Mindcraft: A Dynamical Systems Theory of Cognition

by

David Leo Barack

Department of Philosophy
Duke University

Date:_______________________

Approved:

___________________________
Alexander Rosenberg, Supervisor

___________________________
Michael Platt

___________________________
Karen Neander

___________________________
Walter Sinnott-Armstrong

___________________________
Felipe De Brigard

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Philosophy in the Graduate School
of Duke University

2014
ABSTRACT

Mindcraft: A Dynamical Systems Theory of Cognition

by

David Leo Barack

Department of Philosophy
Duke University

Date:_____________________
Approved:

___________________________
Alexander Rosenberg, Supervisor

___________________________
Michael Platt

___________________________
Walter Sinnott-Armstrong

___________________________
Karen Neander

___________________________
Felipe De Brigard

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Philosophy in the Graduate School of Duke University

2014
Abstract

This dissertation develops a theory of cognition, driven by recent developments in the electrophysiological investigation of the neuronal mechanisms that support adaptive behavior. In the first chapter, I situate the theory in the conceptual landscape of the philosophy of mind, distinguishing componential from systemic dynamical theories of cognition. In the second chapter, I analyze two case studies from electrophysiological cognitive neuroscience, arguing that cognitive neuroscientists are beginning to uncover the dynamical components of cognition. I next define model execution and mechanism implementation, showing how these concepts connect physical mechanisms, dynamical mechanisms, and the formal models of processing that characterize cognitive functions. Drawing on the recent literature on mechanisms and scientific explanation, I propose a revised definition of a mechanism that accommodates these dynamical mechanisms, as well as making room for their implementation by physical mechanisms. These dynamical mechanisms possess a set of organized components and activities that execute the formal models of processing, and are implemented by the physical machinery of the cognitive system, such as the brain. In the third chapter, I argue that the component dynamical mechanisms of cognitive systems are distinct from though implemented by physical mechanisms. I argue that these multiple interacting dynamical mechanisms are the components of cognition, defending this componentiality claim
against several objections. I next discuss the grounds for inferring the existence of dynamical mechanisms that are type distinct from physical mechanisms, their implementing substrate. In the fourth chapter, I argue that these dynamical mechanisms are reused: they can execute different formal models and be implemented by different physical substrates. I define this concept of reuse, situating it in the debate on theories of reuse, and illustrate how dynamical mechanisms are reused in cognitive systems.
Dedication

To my friends, for their love and support.
# Contents

Abstract..............................................................................................................................................iv

List of Tables ........................................................................................................................................... xi

List of Figures .......................................................................................................................................... xii

Acknowledgements ................................................................................................................................. xiv

1 The Enigmas of Thought ......................................................................................................................... 1

   1.1 Classicism, Connectionism and Dynamicism ............................................................................... 8

   1.2 Two Conceptions of Dynamicism ............................................................................................. 27

   1.3 Conclusion .................................................................................................................................. 41

2 Dynamical Mechanisms in Mind ........................................................................................................... 43

   2.1 Two Case Studies in Cognitive Neuroscience ........................................................................... 44

      2.1.1 Case Study 1: Cognitive Mechanisms of Perceptual Decisions ........................................ 44

      2.1.2 State-Spaces, System Evolution Functions, and Dynamics ............................................. 49

      2.1.3 Integrate-to-bound Systems and Perceptual Decision Mechanisms ............................ 59

      2.1.4 Case Study 2: Cognitive Mechanisms of Foraging Decisions ........................................ 64

      2.1.5 Defining Execution and Implementation ............................................................................. 69

      2.1.6 Formal Models of Processing and Model Execution ....................................................... 69

      2.1.7 Implementation .................................................................................................................... 76

   2.2 Dynamical Mechanisms ................................................................................................................... 83

      2.2.1 Recent Approaches to Mechanisms .................................................................................. 85

      2.2.2 Dynamical Systems as Mechanisms? ............................................................................... 91
List of Tables

Table 1: Cognitive Reuse, Recycling, and Composting......................................................... 242
List of Figures

Figure 1: Newell’s framework for cognitive systems. Adapted from Newell 1992, p. 108. ................................................................. 12

Figure 2: Cognitive systems as the interaction of brain, body, and environment. Adapted from Beer 2000, p. 97 ................................................................. 33

Figure 3: Average recorded firing rates from cells in the parietal cortex of monkeys while they make a perceptual decision in different evidential conditions. Adapted from Roitman and Shadlen 2002, p. 9482 ................................................................. 47

Figure 4: The patch-leaving task. Adapted from Hayden et al. 2011, p. 934 .................. 64

Figure 5: Example neuron from the anterior cingulate cortex, recorded during the patch-leaving task. These cells exhibit a peri-saccadic response that increases over the duration of foraging in a patch. Adapted from Hayden et al. 2011, p. 935 ................................................................. 67

Figure 6: Example genetic network for regulating circadian rhythm. Adapted from Leloup et al. 2002, p. 7052 ................................................................. 103

Figure 7: Concentrations of genetic products and periodicities in the generation of circadian rhythms. Adapted from Leloup et al. 2002, p. 7053 ................................................................. 104

Figure 8: Task and neural data from the multi-target cued saccade task. Adapted from Louie et al. 2011, p. 10631 ................................................................. 153

Figure 9: Inversion and excursion in LIP neurons during a 2-alternative and 4-alternative RDMT. Adapted from Churchland et al. 2008, p. 698 ................................................................. 157

Figure 10: The evolution of LIP activity and the decision field D. A: LIP activity is modulated by the value inside the cell’s RF and total value outside the RF (VIN and VOUT, respectively). For constant VIN and VOUT, LIP activity is suppressed following target onset, but exhibits inversion and greater excursion, both as a function of increasing number of targets. For constant value VIN but variable VOUT, activity exhibits suppression only. B: Decision field D(n,t) for RF containing a target. The decision field strength rises as a function of the number of targets (n), time-locked to evidence onset. Two downward arrows shown inside the plot correspond to the time of activating decision field (D(n,t) > 1) and deactivating decision field following a decision just prior to saccade onset (D(n,t) = 1), respectively. ................................................................. 162
Figure 11: Graphical model illustrating different ways to modulate patch leave times by modulating dynamical components in an integrate-to-bound mechanism. Adapted from Hayden et al. 2011, p. 937.

Figure 12: Panel A: As the barrier value increases along the x-axis, the percentage of correct choices increases as well. Panel B: As the motion strength increases along the x-axis, so too does the probability of choosing correctly. Adapted from Gold and Shadlen 2002, p. 304.

Figure 13: As the percentage of coherently moving dots increases, so too does the probability of choosing correctly. Adapted from Roitman and Shadlen 2002, p. 9479.

Figure 14: Simulated and recorded LIP firing rates during the RDMT. Adapted from Mazurek et al. 2003, p.1261.

Figure 15: Diagram illustrating the synaptic reverberation dynamical mechanism. Adapted from Wong and Wang 2006, p. 1316.

Figure 16: Pre-saccadic increase in activity in an LIP cell. Adapted from Barash et al. 1991, p. 1100.

Figure 17: The random dot motion task. Adapted from Gold and Shadlen 2007, p. 548.

Figure 18: The simulated patch-leaving task. Adapted from Hayden et al. 2011, p. 934.

Figure 19: Behavior on the patch-leaving task. Adapted from Hayden et al. 2011, p. 934.

Figure 20: Example neuron from the anterior cingulate cortex, recorded during the patch-leaving task. Adapted from Hayden et al. 2011, p. 935.

Figure 21: Three different processing sequences for selecting an action. Adapted from Cisek 2012, p. 2.

Figure 22: Cisek’s competing proposal for a dynamical mechanism for action selection. Adapted from Cisek 2006, p. 9763.
Acknowledgements

I would like to thank my advisers, Alex Rosenberg and Walter Sinnott-Armstrong, for their encouragement and criticisms, which have greatly enhanced the quality of this dissertation, as well as the quality of my thinking. I would like to gratefully acknowledge the support and insight of Michael Platt, who, forsaking reason and sound intuition, took a chance on having a philosopher in his lab. I had the incredible fortune to interact extensively with Fred Dretske, who sadly passed away before I completed this dissertation. Fred’s philosophical acumen was and continues to be instrumental in my development as a philosopher. I would also like to thank Karen Neander and Rich Mooney, both of whom were sources of wisdom and guidance through my tenure at Duke; Robert Brandon, Felipe De Brigard, Owen Flanagan, and Iris Einheuser all provided careful philosophical criticism and inspiration over many years; Steve Chang, John Pearson, Karli Watson, Geoff Adams, and Nick DeWind for their scientific expertise, experimental prowess, and fine judgment; and to all my professors across my many years of graduate education and training. Special thanks as well to Brian Keeley at Pitzer College, without whom I would never have begun this neuroscientific and philosophical journey. Finally, I would like to thank Leakey and Ramon, my two boys, for being such hard workers and diligent monkeys, all for a chance to drink some juice.
Spending six years in Durham and at Duke has been a challenging but rewarding experience, not just academically, but socially as well. I would like to thank all of my friends, to whom this dissertation is dedicated, for the warmth, love, and support they have provided in what has been at times a trying and difficult path. My life has been enriched by the laughter, conviviality, and adventures we shared.

Though many different people left indelible mark on my life, I would like to thank a few in particular. To Charles, for your companionship, sound advice, and willingness to hear me out, even when I was being ridiculous. To Sara, for your steadfast friendship, insight, and fundamentally sound guidance. To Mathieu, for your insistence on thinking things through, paying attention to alternate explanations, careful and thoughtful opinion, and resolute companionship. To Marc, for teaching me to not give so much of a damn, your compassion and care, and your excellent sartorial sense. To Jeremy, for listening to me, providing much needed wry commentary on what is obvious, and for being a friend to share my thoughts with, often over a beer.

To Nick, to whom I owe so much, for showing me how to be a good friend and for sticking with me longer than was warranted. To Leonore, for being a stalwart supporter and for the moments of cutting loose. Despite the difficulties, I’ve learned a lot from you both and I thank you.

A special thanks to Sumi. Without your guidance as a Buddhist, I would not have come to understand some of the fundamental challenges in my life. Perhaps
counterintuitively, understanding how to be a person has forced me to let go of my Buddhist detachment and to engage more fully with my desires, fears, cravings, and loves. That has made me a more complete person, and I hope that someday, I will understand more fully some of the lessons you so willingly shared.

Special thanks to Sal. You played a critical role in the development of this dissertation, as well as in my development as a person. Your demands for me to explain myself, to state precisely what I mean, and your insistence that I be clear made me realize that I often have no idea what I mean when I say the things I do. That insight has propagated throughout my work, and I am a better thinker and a better person for it. I am so grateful for your friendship and support.

As I move forward in life, I will always recall the time I spent in Durham as a challenging but ultimately fruitful period in my life. Importantly, I want to thank everyone whom I had the good fortunate to befriend while at Duke. I hope that the experiences I’ve had will make me a wiser and more aware person, and that I will be more honest, more forthright, and more compassionate and loving as I move on in life.
1 The Enigmas of Thought

In World War II, a team of codebreakers at Bletchley Park, with the philosopher and mathematician Alan Turing playing a central role and assisted by stolen tables and Polish intelligence, broke the Enigma code, the German military cipher used for clandestine missives. Each message was encoded using an Enigma machine, a complex device that transformed the communiqué into a coded message. These complex machines turned a natural language input into a coded output according to the internal dynamics of the device, as captured by a series of complex mechanical components and activities.

Imagine that the mind works in a similar fashion, but instead of only one Enigma device, there are many such complex dynamical machines. Physical mechanisms possess dynamical properties, which I define as patterns of change in physical components either over time or with respect to other components. I argue that collections of these dynamical properties, subject to certain constraints on mechanisms outlined below, constitute dynamical mechanisms. These dynamical mechanisms execute formal models of processing, mathematical functions that specify a set of variables and their interrelations, and the dynamics in turn are implemented by these systems’ physical mechanisms. Such complex dynamical machines transform the input received from the environment or from other internal machines into behavioral output or internal coded
messages for further processing, resulting in adaptive behavior. These complex
dynamical engines are the Enigmas of thought.

In many ways, contemporary investigators of the mind are in the same position
as the codebreakers at Bletchley: they are examining a complex system that transforms
inputs into coded outputs according to some complex dynamic internal processes.
However, unlike in the case of the codebreakers, the target system is almost
unfathomably more complex and interconnected. While researchers have access to
(perhaps only some of) the inputs, the signals from the environment or from other parts
of the system,\(^1\) as well as to the outputs, the behavior issued by the system or the signals
sent to other parts of the system,\(^2\) the complexity of the system and the ignorance of
researchers regarding which internal transformations are possible results in a reverse
engineering problem of enormous difficulty and intransigence. Nonetheless, the idea of
locating discrete dynamical components that execute special kinds of signal
transformations according to some mathematical formalism can still guide research into
the device, just as in the case of the Enigma machine.

\(^1\) Even when such inputs are accessible to researchers, they are often ill-defined: the causal complexity of the
environment can conceal what aspects of the environment are driving behavior. Similar concerns apply to
internally determined inputs.

\(^2\) Similar to the case of inputs, while the outputs are accessible, they are also often ill-defined: which aspects
of the movements of the system are aspects of its behavior is confounded by the complexity of the
movements of the system, and similar remarks apply to the signals broadcast by internal components.
Here, I provisionally identify cognitive systems as those that behave adaptively, flexibly modulating their behavior on the basis of the environment and goals of the system. Not every system that exhibits such behavior will be a cognitive system, of course, and some cognitive systems may not result in behavior, adaptive or otherwise, hence the provisional status. Imagine a cognitive system that cannot act, perhaps because it has been detached from any effector systems. Since it is detached from any mechanism that permits behavioral output, this system is unable to adapt its behavior to its goals and environment. It may be objected that we shouldn’t permit such systems to be classified as cognitive systems; perhaps behavioral adaptation—or behavior at all—is a necessary capacity for cognitive systems. I won’t be arguing for this behavioristic criterion. All the same, these subtle conceptual issues should be settled by reasoned debate, not by fiat, and identifying the types of systems my theory targets as those systems that are able to modulate their behavior in a way reflecting their goals and environment should be seen as only an initial, provisional, defeasible criterion for a cognitive system. I believe that, ultimately, cognitive systems will be those systems that are explained using the tools and concepts of cognitive science and related disciplines. These tools may be analyses of the representational properties of systems, the mechanistic nature of systems, or some other set of criteria. Conceivably, that set of systems will only partially overlap with systems that modulate their behavior to reflect
their goals and environments. This latter characterization of a cognitive system is just a useful operationalization of the concept of a cognitive system for the current discussion.

There are at least two types of theories of cognition: empirical and philosophical. I am offering a bit of both. A philosophical theory of cognition offers the conceptual grounds for empirical theories. I will be outlining a set of principles that are meant to provide these conceptual grounds, arguing that cognitive systems are partly made of dynamical mechanisms. An empirical theory of cognition offers a detailed scientific account of how cognitive systems work, how they develop, how they evolved, and situates cognitive systems in the natural world. As applied to my approach, these empirical questions include which specific dynamical mechanisms appear in cognitive systems, how those mechanisms interact and combine to produce cognitive phenomena, and how they develop and evolve. It is not my intent in this dissertation to answer these questions, though certainly I illustrate the philosophical theory with examples drawn from the science. Many extant theories of cognition present specific functional architectures, descriptions of the operations, components, and their organization, for cognitive systems. I won’t be offering such a comprehensive proposal, though the case

---

3 There are many examples of such philosophical theories, including Fodor (Fodor 1975, 2008), Pylyshyn (1984), Smolensky (1988), van Gelder (1998), and many others. Some of these I briefly review below.

4 Examples abound in the literature, ranging from classical approaches (e.g. Newell 1992) to more recent proposals (e.g. King and Gallistel 2009; Eliasmith 2013).
studies I discuss are chosen not only for their illustrative purposes, but also because I think they are right, uncovering some fundamental aspects of the way actual cognitive systems are organized. A theory of cognition, of course, can serve both philosophical and empirical roles.5

In this dissertation, I develop a philosophical theory of cognition, grounded in the notion that minds are the effect of the concerted activity of many dynamical mechanisms. This theory of cognition accommodates intuitions about the possibility of minds made out of diverse physical substrates, as well as recent research into the neural basis of much animal cognition.6 In particular, the theory endorses at least the following theses:

1. Cognitive systems’ dynamical mechanisms execute formal models of processing.
2. Cognitive systems’ dynamical mechanisms are distinct from but implemented by physical mechanisms.
3. Cognitive systems repeatedly use the same dynamical mechanism to execute different formal models of processing.

5 Newell’s theory is a good example of this: he provides a conceptual background for a theory of cognition, and then proposes a specific cognitive architecture (SOAR) (Newell 1992).

6 What is the value in supporting the intuition behind diverse physical bases for cognitive systems? There is a sense in which we wish to be cognitive liberalists, and permit the attribution of cognitive states to systems that are very different from humans, such as aliens or computers or other animals. Characterizing dynamical cognitive mechanisms permits us this sort of wiggle room in our attributions of cognition. Of course, we don’t want to be cognitive chauvinists either; just because we identify one dynamical mechanism for a particular cognitive function does not imply that there aren’t other abstract mechanisms that can explain that cognitive capacity in a ‘how possibly’ fashion (Brandon 1990; Craver 2007), and which we may ultimately discover are in fact implemented in certain systems. See (Block 1980). The distinction I make between dynamical and physical (or biophysical) mechanisms is meant to retain some of this flexibility in our psychological attributions.
I defend these theses over the course of this dissertation. The dynamical mechanisms at
the heart of this theory of cognition are not mysterious, immaterial or ethereal in any
way. They are token identical to subsets of properties of physical mechanisms, and they
are implemented by those physical mechanisms, as defined and discussed below. However, they can be implemented by many different kinds of physical mechanism,
preserving the intuitions about the possibility of diverse sorts of minds.

The plan of the dissertation is simple. In the remainder of this chapter, I will lay
out the conceptual space in which this dynamical mechanism theory of cognition is
located. I will also contrast this theory with other approaches to cognition, discussing its
approximate location in the conceptual landscape. In the following chapter, I present
several case studies from cognitive neuroscience, illustrating the discovery of repeated
patterns of dynamical properties. I briefly discuss the relationship between these
component dynamical systems and cognitive functions. A formal model of processing
mathematically characterizes cognitive functions, and this formal model of processing is
executed by those dynamical systems. These models are executed in virtue of a weak
equivalence, or input-output equivalence, between the model and systems that execute
it. I also define the implementation relation, and discuss how physical mechanisms
implement dynamical systems. Following a discussion of what makes a mechanism, I
argue that these dynamical systems qualify as mechanisms and hence are dynamical
mechanisms. In the third chapter, I discuss at length how dynamical mechanisms play a central role in cognitive systems and how they are related to physical mechanisms. I respond to several objections to the decomposability thesis entailed by a componential dynamicist approach, illustrating how to obtain complex dynamics by the simultaneous implementation of multiple simpler dynamical mechanisms. I define what it means for a physical mechanism to implement a dynamical mechanism. Finally, I present an argument that dynamical mechanisms and physical mechanisms are distinct, as well as respond to a number of objections. In the last chapter, I discuss cognitive mechanism reuse, the reuse of the same cognitive mechanisms for different cognitive functions. Dynamical mechanisms can execute many different cognitive functions, and can be implemented by many different physical mechanisms. I provide a formal characterization of use and reuse, in terms of implementation and model execution, and illustrate dynamical mechanisms reuse with examples drawn from cognitive and behavioral neuroscience.

In the remainder of this introduction, I will lay out some of the conceptual terrain, drawing from a subset of the approaches to cognition, focusing in particular on philosophical theories of cognition. The purpose of this survey is not to militate for any
particular position. The purpose of the survey is also not to resolve philosophical debates or contention regarding the particulars of the conceptual framework. Rather, the purpose of the survey is only to peremptorily delineate a set of conceptual distinctions, noting the possible sources of contention, in order to situate the area in conceptual space in which my componential dynamicist philosophical theory of cognition is located. Furthermore, I won’t be assessing the viability of various species of the different types of philosophical theories, though undoubtedly some of the assumptions that distinguish some of the theories result in the closure of various parts of the conceptual space.

1.1 Classicism, Connectionism and Dynamicism

Locating the philosophical theory of cognition developed in this dissertation on the conceptual landscape of philosophical approaches will help characterize and develop it, as well as aid in understanding it. Somewhat arbitrarily, and with awkward fits and shaping where necessary, I will break this landscape into three families of theories: classical computational approaches, connectionist approaches, and dynamicist approaches.

In fact, there are only implicit arguments for my componential dynamicist philosophical theory of cognition contained within this dissertation: in virtue of the systematic treatment of a broad swatch of recent empirical evidence for the nature of cognitive processes; the suggestive treatment of regularities in the functional organization of cognitive systems; and the suggestive consolidation or unification of systems biology under a single conceptual framework. A goal of this dissertation is to outline one particular componential dynamicist theory of cognition, respond to some objections to it, and provide evidence for it, not to detail the philosophical arguments for the theory. I undertake that more controversial forensic process elsewhere.
On the classic computational theory of mind (classicism), mental states are representational states and mental processes are computations, transformations operating over representations. Mental states are representational states, physical states that possess semantic and syntactic properties. Mental processes are sequences of such states, determined by computations, physical processes sensitive to the syntactic properties of the representations. A central exemplar of classicism is the language of thought (LOT). The LOT views our mental states as a species of linguistic state, characterized by both syntactic and semantic properties (Fodor 1975, 2008). As Fodor and Pylyshyn characterize it, the LOT is committed to two principles:

“(1) Combinatorial syntax and semantics for mental representations. Classical theories... take mental representations to have a combinatorial syntax and semantics, in which (a) there is a distinction between structurally atomic and structurally molecular representations; (b) structurally molecular representations have syntactic constituents that are themselves either structurally molecular or structurally atomic; and (c) the semantic content of a (molecular) representation is a function of the semantic contents of its syntactic parts, together with its constituent structure....

(2) Structure sensitivity of processes. In Classical models, the principles by which mental states are transformed, or by which an input selects the corresponding output, are defined over structural properties of mental representations. Because Classical mental representations have combinatorial structure, it is possible for Classical mental operations to apply to them by reference to their form. The result is that a paradigmatic Classical mental process operates upon any mental representation that satisfies a given structural description, and transforms it into a mental representation that satisfies another structural description” (Fodor and Pylyshyn 1988, p. 12-13).

In the LOT, every mental state is composed in part of representations consisting of token symbols that are “spatiotemporal particular[s] [having] syntactic and semantic

See also Pylyshyn 1984, Field 1978, or Schneider 2011.
properties and a causal role…” and stand in some relation to an agent on the one hand and the world on the other (Fodor 1990, p. 167). For some belief content, a representation has that content in virtue of its components contributing to the representation’s semantic properties. The representation’s syntactic properties are also composed from the syntactic properties of its components. These syntactic properties correspond to some subset of the physical properties of the symbol. Furthermore, symbols can contribute their semantic and syntactic properties to molecular constructions, combining in a rule-governed fashion according to some grammar. The rules that govern this composition operate over the syntactic properties of the symbols. In short, the LOT has two of the hallmark features of language: an alphabet, a set of atomic symbols with syntactic and semantic properties, and a syntax, a set of rules governing the composition of molecular sentences out of those atomic symbols.

According to Fodor, and as implied by the quote above, classical approaches assume a number of theses, many associated with the use of representational symbols and operations that are executed over them. As another example, compare the LOT position with a similar though distinct classical theory from Newell (Newell 1980, 1992). First, consider Newell’s presentation of a physical symbol system (Newell 1980). After discussing a number of constraints on an analysis of mind, Newell focuses on two in

---

* Fodor and others use the term ‘vehicle’ for the physical entity that is composed of symbols and over which computations are executed. Here I will use the term ‘representation’ or ‘symbol’ to denote these vehicles.
particular: universality, where “a universal machine is one that can produce an arbitrary input-output function; that is, that can produce any dependence of output on input” (Newell 1980, p. 147), and symbolic behavior, behavior characteristic of physical symbol systems, or universal machines (Newell 1980, p. 154). Physical symbols systems are grounded in the following concepts:

“Symbols as abstract types that express the identity of multiple tokens. Expressions as structures containing symbol tokens. Designation as a relation between a symbol and the entities it symbolizes. Interpretation as realizing the designations of expressions. Operations of assigning symbols, and copying, reading, and writing expressions” (Newell 1980, p. 168).

Newell characterizes the physical symbol system hypothesis:

“Physical Symbol System Hypothesis: The necessary and sufficient condition for a physical system to exhibit general intelligent action is that it be a physical symbol system. Necessary means that any physical system that exhibits general intelligence will be an instance of a physical symbol system. Sufficient means that any physical symbol system can be organized further to exhibit general intelligent action” (Newell 1980, p. 170).

Physical symbol systems possess symbols, abstract types that occur in expressions, possessing representational properties like designation and interpretation, and over which operations are executed. Being a physical symbol system, in turn, is necessary and sufficient for general intelligent action, that is, for the property of universality.10

---

10 Do these physical symbol system considerations, which clearly share central conceptual assumptions with the LOT, entail the postulation of a language of thought, that is, the postulation of an internal alphabet and internal grammar? This depends on whether the representational assumptions shared across classicists entail the existence of such a language.
Figure 1: Newell’s framework for cognitive systems. Adapted from Newell 1992, p. 108.

Next, consider Newell’s presentation of the conceptual foundations of cognitive science (Newell 1992). This presentation is summarized in figure 1 (adapted from
Newell 1992, p. 108). The target of explanation for these systems are behaving systems, and Newell takes the mind to be “the control system that guides the behaving organism in its complex interactions with the dynamic real world” (Newell 1992, p. 43). In particular, Newell is interested in explaining systems that exhibit wide-ranging behavior, systems described “in terms of their having knowledge and behaving in light of it” (Newell 1992, p. 45, italics in original). Emphasizing that describing a system as having knowledge is just one of many ways of describing a system, describing a system as a knowledge system “abstracts completely from the internal processing and the internal representation…. [A]ll that is left is the content of the representations and the goals toward which that content will be used” (Newell 1992, p. 48-49). Newell proceeds to discuss a complex series of claims based on this idea of analyzing intelligence as a particular way of describing certain systems, whose central conceptual components include knowledge, representation, combinatorial transformations, symbols, and the like.\footnote{Newell’s definition of intelligence I find particularly insightful: “Intelligence is the ability to bring to bear all the knowledge that one has in the service of one’s goals” (Newell 1992, p. 90).}

This brief discussion of classicism indicates that it is committed to the following set of processing theses:

1. Mental states are representational states with semantic and syntactic properties.
2. Mental states have a combinatorial semantics and a combinatorial syntax.
(3) Mental states are (in part) composed of local, narrow, discrete, digital, invariant, and atomistic symbols.
(4) The semantic properties and syntactic properties of mental states are at the same level of description.
(5) Mental processes are transformations defined over mental states.\(^{12}\)
(6) Mental processes are sensitive to the syntactic properties of mental states.

The first thesis commits classicism to representationalism, the thesis that cognitive states are representational, with alethic, extensional, and possibly other semantic properties, such as intensional properties. The second thesis commits classicism to combinatorialism, the thesis that there are molecular representations composed out of atomic (or other, simpler molecular) representations, and these molecular representations are both semantically combinatorial and syntactically combinatorial.

Semantic combinatorialism says that the semantic properties of a molecular representation are composed only out of the semantic properties of the constituents such that the meanings are the sum of the parts.\(^{13}\) Syntactic combinatorialism says that the syntactic properties of a molecular representation are composed only out of the syntactic properties of the constituents such that the structure of the molecular representations is the sum of the structure of the parts.\(^{14}\) While the properties of the representational

\(^{12}\) I know of no theory that views mental processes as not composed of sequences, in some sense, of mental states. Furthermore, what is meant by transformations will be discussed below.

\(^{13}\) This vague formulation is clearly underspecified, but will do for this introductory discussion.

\(^{14}\) Note that for this discussion, I am drawing an equivalence between structural and syntactic properties. Devitt has perceptively argued that these should be distinguished (Devitt 1991).
symbols present in classical cognitive systems is debated, they are often attributed the properties in thesis three. These symbols are local: they are not distributed across many of the entities over which mental transformations are defined, but rather they are directly the entities that stand in the relations definitive of mental transformations. These symbols are narrow as opposed to extended: their semantic properties stand independent of the environment.\textsuperscript{15} These symbols are discrete: they are independent of other symbols, clearly differentiated from them.\textsuperscript{16} These symbols are often digital: they are numbers or concatenations of numbers.\textsuperscript{17} These symbols are invariant: they do not change their semantic or syntactic properties while retaining the same causal role in the system. Finally, these symbols are often cast as atomistic, possessing their semantic properties independently of other representations present in the system. The fourth thesis commits classicism to analyzing the semantic and syntactic properties of mental states at the same level of description; connectionism, we will see in a moment, rejects

\begin{flushright}
\footnotesize
\begin{enumerate}
\item This is a notorious assumption in the philosophy of mind, as Putnam, Burge, and many others have eloquently argued that the semantic properties of representations must reach beyond the confines of the head (Putnam 1973, Burge 1979). Fodor has attempted to craft a more narrow notion of content (Fodor 1991). Since I am just painting a broad strokes picture, I’m ignoring these important debates. Furthermore, there are classical approaches that sanction such extended semantic notions (e.g. Wilson 1994).
\item Here I conform to Maley’s usage (Maley 2011). Maley defines discreteness as relative to a representational scheme wherein “representations are distinct from other representations in the same representational scheme, there are gaps between the possible representations, and representations only come in wholes” (Maley 2009, p. 125).
\item This is especially the case for approaches that assimilate the classical view to digital computers.
\end{enumerate}
\end{flushright}
this thesis. The fifth thesis commits classicism to mental processes as made of sequences of mental states and the transformations between the elements of those sequences. Finally, the sixth thesis commits classicism to mental processes as transformations that are sensitive to the syntactic properties of mental states, in that the transformations are functions of those properties. On classical theories, syntax recapitulates semantics, mental processes are indirectly sensitive to the semantic properties of mental states as well. Commitment to these theses picks out the class of classical computational philosophical theories of cognition. In the recent history of the philosophy of mind, this classical approach is often contrasted with connectionism.\(^\text{18}\)

Connectionist theories are a class of theories of cognition, both empirical and philosophical, all of which compose cognitive systems out of networks of units. These units are quasi-neuronal but have a number of properties that are distinct from those of networks of neurons (e.g. see Smolensky 1988, p. 9 for a discussion of the differences). Each unit has an activation level and an activation function, describing how active the unit is as a function of its inputs from other units and from the environment.

\(^{18}\) Note that there are quasi-classical approaches that reject some number of these postulates. Fodor discusses some alternatives, for example railing against the possibility of a theory that rejects syntactic combinatorialism while accepting (1) in the form of atomic representational mental states, which he calls the “fusion” view (Fodor 1978). A view that rejects the semantic components of these theses but retains the syntactic components is Stich’s syntactic computational theory of mind (Stich 1983).
Connectionism comes in a variety of forms, but the different types seemingly take issue
with a number of the theses central to classicism.

For example, consider Smolensky’s discussion of connectionism (Smolensky
1988). The view he advocates, the proper treatment of connectionism view, is a
trippartite hypothesis, composed of a syntactic principle, a semantic principle, and a
levels hypothesis, focused on describing the machine that runs the “kinds of programs
[that] are responsible for behavior that is not conscious rule application”, which he calls
the “intuitive processor” (Smolensky 1988, p. 5). The syntactic principle, about the
formal properties of the states over which operations that correspond to mental
processes are defined, states that “[t]he state of the intuitive processor at any moment is
precisely defined by a vector of numerical values (one for each unit). The dynamics of

19 For a similar view, the parallel distributed processing view, see (Rumelhart et al. 1986). An interesting, less
extreme version of a connectionist approach is Mesulam’s selectively distributed processing (1994, 1998;
Seeck et al. 1995):

“The anatomical areas that play crucial roles in the identification of colour, movement, faces,
words, objects and spatial targets display relative rather than absolute specializations. For example,
V4 is specialized for colour perception but also participates in spatial attention, the identification of
salience and the encoding of form…. In turn, the processing of colour information may involve not
only V4 but also a part of the lateral peristriate cortex…. Furthermore, neuronal ensembles
selectively tuned to canonical features of faces participate, although to a lesser extent, in encoding
other visual entities…. It is quite likely that several ensembles, each composed of neurons
optimally tuned to a different category, form an interdigitating or partially overlapping mosaic and
that the predominant type of ensemble varies from one location to another. This organization has
been designated ‘selectively distributed processing’… it apart from other models based on
equipotentially distributed processing,... parallel distributed processing... and modular
processing...” (Mesulam 1998, p. 1022).

While this discussion does not engage with the more empirical considerations behind different types of
cognitive or neural architecture, Mesulam’s proposal for a partially distributed yet still componential
organization for cognitive systems is very much in the spirit of my proposal.
the intuitive processor are governed by a differential equation” (Smolensky 1988, p. 6).

The semantic principle, about what the objects are that possess semantic conceptual properties, states that “[t]he entities in the intuitive processor with the semantics of conscious concepts of the task domain are complex patterns of activity over many units. Each unit participates in many such patterns” (Smolensky 1988, p. 6). Smolensky calls this the subconceptual unit hypothesis, because the level at which the system components are described is below the level at which semantic conceptual properties are attributed to the system. Finally, the subconceptual level hypothesis, about the appropriate level of description for formulating principles of cognition, is that “[c]omplete, formal, and precise descriptions of the intuitive processor are generally tractable not at the conceptual level, but only at the subconceptual level” (Smolensky 1988, p. 6-7). He summarizes these three principles in the subsymbolic hypothesis: “The intuitive processor is a subconceptual connectionist dynamical system that does not admit a complete, formal, and precise conceptual-level description” (Smolensky 1988, p. 7). Fundamentally, cognition is to be explained via a descriptive dimensional shift from the semantics of the original behavior to the semantics of the subconceptual level and the subsymbolic systems therein. Cognitive systems, in short, are reduced to subconceptual systems, a “kind of reduction familiar in natural science” and exemplified by the reduction of the laws of thermodynamics to mechanics (Smolensky 1988, p.11).
Symbolic descriptions are then merely approximate descriptions of the subsymbolic process that are actually constitutive of cognitive systems: “[s]ubsymbolic models accurately describe the microstructure of cognition, whereas symbolic models provide an approximate description of the macrostructure” (Smolenksy 1988, p. 12). On Smolensky’s connectionism, mental transformations are defined at the level of the vectors precisely describing the level of activation of each unit, the conceptual entities are patterns of activity over many units, and precise descriptions of cognitive processes are only available at the level of the vectors describing the activation of each unit in the network.

Several distinctions can be drawn between classicism and connectionism. Both views accept (1), that mental states have representational properties. There is a difference in the nature of the symbols present in each theory, however. As discussed above, classical approaches see the mind as divvied up into symbol structures. The semantic properties and the structural properties are at the same level, attaching to the entities over which transformations are executed. These symbols are local as opposed to

---

20 Recall that symbols are entities with both syntactic and semantic properties. I’m ignoring classical approaches that eschew semantic properties (Stich 1983), and I’m ignoring connectionist approaches that eschew either semantic properties—of which I know no examples—or syntactic ones—again, of which I know no examples. Fodor has argued that the only non-implementing-classical-approaches versions of connectionism are the former sort, eschewing combinatorial semantic properties altogether (Fodor and Pylyshyn 1988). By Fodor’s lights, connectionists cannot recover the semantics of the components of molecular representations. Eliaashmith reviews the many different possibilities for recovering the semantic structure of representations from connectionist vectorial representation, implicitly proving the lie to Fodor’s claim (Eliaashmith 2013).
distributed, narrow as opposed to extended, discrete, often digital, and finally atomistic instead of holistic. In contrast, connectionist approaches see symbols as distributed, non-discrete, and often holistic, with their semantic properties partly determined by the other symbols present in the system (Churchland 1988, 2012). These representations are often digital, as they are typically vectors, and they are often narrow, though they may also be extended. Thus, connectionists will often reject (3), attributing different properties to symbols than classicists.

For connectionists, the semantic properties may be at a different level of description than the syntactic properties of mental states, thus rejecting (4) above. On Smolensky’s theory, there is a distinction between the conceptual and subconceptual levels. Intuitively, the semantic properties or the syntactic properties of mental states may exist at either level. On Smolensky’s view, the semantic properties are at the conceptual level, corresponding to patterns of activation across many units of the network. The syntactic properties, in contrast, correspond to the activation function that describes the evolution of each unit’s activation over time, and thus reside at the level of the individual unit, the subconceptual level, not at the level of the patterns of activation that correspond to the semantic properties of the system. (Note, though, insofar as the approximate descriptions at the conceptual level are perforce approximations of precise

\[21\] The concept of subconceptual semantic properties has been challenged; see the critiques following (Smolensky 1988).
semantic characterizations at the subconceptual level, there will be semantic properties at both levels; note furthermore that, on a connectionist approach there may also be ‘real’ structure at the conceptual level too.) In particular, connectionism sees the transformations that are central to mental processes as the traversal of the state-space of the system along a trajectory, described by transitions between vectors corresponding to different mental states. These vectors have structure, but the level of description at which they have structure is mismatched with the level of description at which these mental states have semantic properties. This stands in contrast to classicism, in which the very same elements of the system that possess the semantic properties are the elements whose structure mental transformations are sensitive to. Thus, connectionists will reject (4). Mental processes are still transformations defined over mental states, so both approaches sanction (5), but due to the rejection of (4), the transformations may not be sensitive to the structural properties at the same level as the semantic properties of mental states. Classicism, in contrast, accepts (4), and thus that the representations are sensitive to properties that are located at the same level.22

22 The issue of the existence of a grammar, along the lines of a language of thought, for connectionist approaches is a historically fraught one (see Macdonald and Macdonald 1995 for the collected discussion). Fodor has maintained that, insofar as connectionism is a viable position in the philosophy of mind, it is only because it implements a classical approach such as the language of thought (Fodor and Pylyshyn 1988). If connectionism is merely an implementation of the LOT, then (4) must be true at some level of description of connectionist systems, perhaps the subconceptual level. If connectionism is not an implementation, then (4) may not be true of the states of the system, at whichever level they are located. The arguments cited by
As far as mental processes are concerned, both connectionism and classicism accept (5), that mental processes are transformation over mental states. The nature of these transformations, however, is distinct between the two approaches. For classicists, these transformations consist in the production of certain output representations or behavior from input representations or sensory input. The nature of this transformation is typically described using discretized operations executed over those representations and/or inputs. For connectionists, a vector describes the entire network state, with each element in the vector a number corresponding to the level of activation of each unit in the network. Mental states are identified either with network states or with sets of network states, those are that are relevantly similar to each other in some way as specified by a similarity metric. Mental processes are transformations over mental states in the sense of traversing a trajectory in the system’s state-space, described by the (usually differential) equations by which the activation of the individual units change over time.

Both approaches agree that mental processes are sensitive to the syntactic properties of mental states (6). Classical theories hold that mental processes are transformations defined over the syntactic properties of representations, satisfying (6).

Fodor against several quasi-classical approaches could be applied to connectionist theories that reject (4) (Fodor 1978; Fodor and Pylyshyn 1988).
Connections theories hold that mental processes are transformations defined over the units in the network, also satisfying (6). As noted above, because syntax recapitulates semantics, on the classical view mental processes are also sensitive to the semantic properties of mental states. According to the classicists, there is no way for semantic sensitivity to be true unless connectionism is an implementation of classicism; for connectionists who reject this implementation thesis, mental processes could still be sensitive to the semantic properties of mental states because mental processes are indirectly sensitive to the subconceptual semantic properties or, supposing approximate descriptions of subconceptual semantic properties qualify as semantic properties, because mental processes are indirectly sensitive to conceptual semantic properties.\(^{23}\)

We can see that there are a number of theses agreed upon by both classicists and connectionists. Both approaches sanction representational mental states with semantic and syntactic properties. Both see mental processes as transformations over these representational states, and these transformations are sensitive to the semantic and

\(^{23}\) Here we see the need to distinguish certain aspects of (5): mental processes as transformations defined over mental states requires structural properties of the involved representation. If we grant purely semantic transformation, then we may have other versions of these positions, but the only way that I can see transformations being defined over purely semantic properties is to have the semantic properties disconnected from the structural, and by inference physical, properties of mental states. Relinquishing physicalism is not being considered herein. Note that a connectionism that countenances conceptual level semantic and structural relevance for mental states and processes does not entail classicism; insofar as classicism denies the existence of subconceptual semantic and/or syntactic properties, and connectionism essentially requires such subconceptual properties, no amount of conceptual level semantic and syntactic relevance granted by a connectionist theory for mental states and processes will entail classicism. Furthermore, differences in the nature of symbols between the two approaches would remain.
syntactic properties of the processes’ constituent states. There are a number of
differences as well. While classicists see the semantic and syntactic properties of mental
states as occurring at the same level, connectionists do not. Furthermore, they disagree
over the nature of symbols, the properties of the entities over which mental processes
operate and that determine the nature of mental states.

In addition to the processing theses (1)-(6), however, there is a pair of system
theses. These regard how cognitive systems are situated in their environments and
whether cognitive systems can be decomposed into parts. First, theories of cognition
conceptualize cognitive systems as related in certain ways to their environment. In
particular, cognitive systems may be said to extend beyond the physical boundaries of
the organism or artifact we typically take to be the physical foundations of such systems.

(7) Cognitive systems are extended.
Extended theories of cognition endorse (7), that cognitive systems do not end at the
boundaries of the classically construed physical device, such as the boundaries of the
brain or body.24 Environmental properties can act on cognitive systems in the form of
inputs or background constraints on the mechanisms that compose cognitive systems,
but they can also be part of cognitive systems themselves. Most versions of classicism
reject (7) but there are some that accept extending cognitive systems to include aspects of

24 For one locus classicus, see (Clark and Chalmers 1998).
the environment (e.g. Wilson 1994). Though I know of no extended connectionist theory, there is no a priori reason to reject the position as incoherent; for example, some of the units in connectionist systems may fall outside of the boundaries of the brain or body.

Second, there is a difference in the mechanistic assumptions behind each theory, and in particular, differences in componentiality between the two types of cognitive theory. Both views see cognitive systems as composed of mechanisms, varying in nature in different respects to different degrees.

(8) Cognitive systems are made up of components. Classical approaches distinguish processors from the representations over which those processors execute transformations. Classical approaches often see cognitive systems as composed of a central processor or executive that controls and manages the transformation executed over representations elsewhere in the system.25 Furthermore, many aspects of the system’s processing are modular, being executed by systems that are in some sense dedicated to processing certain types of information (e.g., the sensory processing areas in the brain).26 Thus, the classical approach views cognitive systems as composed of many separate processors operating over distinct representations, with all

25 Though this need not be the case. On one approach to cognitive architectures, production systems, there may not be a central executive as traditionally construed (Newell 1973). Furthermore, the central executive must itself be decomposed into some set of simpler component operations, on pain of both infinite regress and homunculus-type objections (Lycan 1981; Dennett 1992).

26 Fodor has presented a stringent definition of a module (Fodor 1983). The notion I’m deploying need not be so restricted.
of this activity possibly coordinated by a central executive. In contrast, connectionist approaches see cognitive systems as composed of many identical elements. Differences in functionality result from differences in inputs and the way these elements are connected to each other. Processors are collocated with representations, and aspects of a coordinating central execution, such as control, planning, and the regulation of other processes, are seen as emerging from the dynamic interaction of the otherwise identical processors. Thus, while many classicists may accept (8), many connectionists would reject it.

Both of these views, classicism and connectionism, have been contrasted with dynamicism. Dynamicist approaches see cognitive systems as a special class of dynamical system, a set of entities that evolve over time and with respect to each other. Furthermore, dynamicist approaches utilize dynamical systems theory, a branch of mathematics, for analyzing cognitive systems. Like connectionist approaches, dynamicist approaches view mental transformations as tracing out trajectories in a system’s state space as described by a set of differential equations, are often antagonistic or agnostic toward representation, and frequently endorse extended and embodied theses about cognitive systems. However, I think the space of dynamicist approaches is

---

27 Classical views positing that cognitive systems are computers may construct such systems out of many components that are largely identical, such as AND gates or XOR gates. However, unlike connectionism, the level at which transformations are executed in such systems is not at this level where the components are largely identical.
larger than generally acknowledged. I will now turn to a basic distinction that underlies the theory of cognition that I am developing herein: the distinction between systemic and componential dynamical systems.

1.2 Two Conceptions of Dynamicism

There are two types of dynamical system, componential and systemic. Componential dynamical systems are composed of dynamical parts; smaller systems that in concert result in the behavior of the larger containing system. Systemic dynamical systems, in contrast, are not composed of dynamical parts; they are complete systems unto themselves. Another way to put this, as composition is relative to the goals of analysis and to how we define componency, is that componential dynamical systems are those that are analyzed into constituent dynamical systems, and systemic ones are those that are not so analyzed. There is a rather straightforward relation between these two, besides one being decomposable and the other not: the decomposable sort, componential dynamical systems, are ultimately decomposed into the non-decomposable sort, systemic dynamical systems. This distinction I contend paves the way for a novel dynamicist approach to analyzing the mind, one that respects the actual development of cognitive neuroscience and opens up new possibilities for understanding and analyzing cognitive phenomena.
To illustrate, consider a car. A car can be analyzed as a dynamical system along either of the senses above. Consider first the systemic sense. A car travels from home to office, from office to store, from store to gas station, and all combinations thereof and many more beside. The course is driver dictated, but setting aside that, we can analyze this movement as a sort of dynamical system. In analyzing any sort of dynamical system, we provide a description of the system in terms of variables and the relations between these variables, and specify how those variables change over time or with respect to each other. The variables in this description include properties of the car, such as the car’s speed, position, fuel levels and possibly other properties. But there are variables external to the car that can describe aspects of the car’s behavior, such as variables relating to the weather, distance to the nearest gas station, (most frustratingly) how many other cars are out on the road, and possibly other properties of the environment.

A car can also be analyzed as a componential dynamical system. A car is a complex physical mechanism, a collection of parts and activities that are organized in a certain fashion so as to be operated by a driver. A dynamical description of this mechanism in terms of its parts would have variables that stand for the various components, such as the engine, the steering mechanism, the wheels and other parts, and the way these items are physically related would be captured by the functional form
of the relationship between these variables. The components that constitute the car appear in the componential dynamical description and the way these components interact in the car is reflected in the equations describing the evolution of those components over time.

Each component, of course, can itself be decomposed into its component parts and activities, and the behavior of that component can be explained by this decomposition. Such a decomposition can be cast in dynamical terms as well, and so each component can be described as a componential dynamical system. This ability to break the component down into its components and provide a componential dynamical description is mirrored by the possibility of providing a systemic dynamical description for each component as well. The component can be considered as a system, with a particular environment populated by other objects, such as the other components of its containing system, and various environmental conditions and properties that can be captured by the variables and parameters in a systemic dynamical description. Any system, then, can receive a componential or systemic dynamical analysis.²⁸

²⁸ There may be some restrictions on this claim; the world may have a ‘top’ and a ‘bottom’, so to speak. The top-most analysis perhaps has as its target the entire universe, that is, everything that exists. There may be little to no sense in trying to provide a systemic description of the universe from this top-most perspective, as there does not seem to be a ‘containing’ environment; ex hypothesi, anything that would be available as part of the containing environment is part of the system targeted for analysis, namely, everything that exists. The bottom-most analysis perhaps has as its target the smallest, non-divisible existent. But, in virtue of its being non-divisible, there are no components to describe dynamically; the system is univocal and exhausted by any variables or parameters that denote it in its entirety. Thus there may be upper and lower bounds on
This division of types of dynamical systems is not meant to be a stark contrast. Mixed systems exist, where the two types of system are co-instantiated. In that case, there will be mixed dynamical descriptions. A car can be described systemically, or componentially, or in a mixed fashion, with some terms for its environment and systemic disposition and other terms for its components, the entities and activities that comprise the car. In a mixed dynamical description of the car, a variable may appear for the weather and another for how much traffic there is, some parameter might account for the weight of the car (which can take on different values, depending on who’s in the car, for example), and other variables might capture aspects of the engine, wheels, or other parts of the car. Most descriptions of systems are mixed in this fashion. Thus, the two types of dynamical description stand as anchors on a scale of dynamical descriptions, with most such systems lying somewhere in the midst of the spectrum.

Furthermore, every systemic description could be turned into a componential one, merely by replacing the various occurrences of the systemic variables with descriptions of the components of the system underlying it. This redescription brings up a related issue in distinguishing systemic and componential dynamicists. Why couldn’t such systems. As the target of description for my purposes and the purposes of this discussion are cognitive systems, I won’t concern myself with these boundaries, at the risk of overlooking cognitive systems that exist in the Lilliputian realm of the very small or the Brobdingnagian world of the very large, that is, everything.
the systemic dynamicist simply note that the approach is componential: it consists in one component, the system, interacting with the environment? I agree that the systemic
dynamicist and the componential dynamicist are on the same mathematical grounds.
The difference lies in the fact that the componential dynamicist advocates for the truth of both a systemic and a componential approach, such that cognitive systems can be
decomposed in just such a fashion, by taking the system variable and replacing it with a
(perhaps complicated) set of variables that stand for the components and their
interactions. In effect, the componential dynamicist views cognitive systems as a type of
mechanism. The nature of this claim will become clearer as the discussion proceeds over
the course of this dissertation.

Many of the extant dynamical approaches analyze cognitive systems as systemic
dynamical systems. For example, Beer argues that dynamicism offers a genuine
alternative to the classical and connectionist approaches. He states that

“…a typical dynamical model is expressed as a set of differential or difference equations
that describe how the system’s state changes over time. Here, the explanatory focus is on
the structure of the space of possible trajectories and the internal and external forces that
shape the particular trajectory that unfolds over time, rather than on the physical nature
of the underlying mechanisms that instantiate this dynamics. On this view, inputs do not
uniquely specify an internal state that describes some external state of affairs. Rather,
they serve as a source of perturbations to the system’s intrinsic dynamics…. a system’s
internal state does not necessarily have any straightforward interpretation as a
representation of an external state of affairs. Rather, at each instant in time, the internal
state specifies the effects that a given perturbation can have on the unfolding trajectory…. an agent’s past experiences influence its future interactions through its internal state, on
multiple time scales” (Beer 2000, p. 96-97).
This quote characterizes one type of dynamicist approach. Dynamicism on this view rejects (1), the representational theory of mind. Since mental states don’t have semantic properties, they don’t have a combinatorial semantics, rejecting (2), and a fortiori no level at which they exist, rejecting (4), though this leaves the question of these states’ structure unanswered. Without representations, there are no symbols carrying semantic and syntactic properties, rejecting (3). Thesis (6) is rejected as well, as there are no semantic properties to which mental states are sensitive. Mental processes may still be defined over mental states, however, potentially leaving (5) intact. Though Beer admits that “a dynamical approach can certainly stand alone”, he contends that “it is most powerful and distinctive when coupled with a situated, embodied perspective on cognition”—that is, when a dynamical approach accepts (7) while rejecting (8) (Beer 2000, p. 97). In the situated embodied perspective, “an agent’s nervous system, its body and its environment are viewed as coupled dynamical systems”, and cognitive systems are to analyzed at this level (Beer 2000, p. 97). In the dynamical approach advocated by Beer and others, cognitive systems are situated and embedded in the world (Figure 2). Beer characterizes this approach as focusing on “continuously engaging an environment with a body so as to stabilize appropriate coordinated patterns of behavior, rather than the sequential sense-think-act processing cycle that is typical of computational approaches” (Beer 2000, p. 97). Insofar as this continuous engagement requires
remaining at the system level, this approach is a systemic dynamicist position.

Figure 2: Cognitive systems as the interaction of brain, body, and environment. Adapted from Beer 2000, p. 97.
As a second example, consider van Gelder’s characterization of the dynamicist position. van Gelder defines dynamical systems as “state-dependent systems whose states are numerical (in the abstract case, these will be numbers, vectors, etc.; in the concrete case, numerically measurable quantities) and whose rule of evolution specifies sequences of such numerical states” (van Gelder 1995, p. 368). On van Gelder’s view, “cognitive systems such as people are dynamical systems..., and cognition is state-space evolution in such systems” (van Gelder 1995, p. 369, italics removed). For van Gelder,

“[t]he core dynamical hypothesis—that the best models of any given cognitive process will specify sequences, not of configurations of symbol types, but rather of numerical states—goes hand in hand with a conception of cognitive systems not as devices that transform symbolic inputs into symbolic outputs but rather as complexes of continuous, simultaneous, and mutually determining change, for which the tools of dynamical modeling and dynamical systems theory are most appropriate. In this vision, the cognitive system is not just the encapsulated brain; rather, since the nervous system, body, and environment are all constantly changing and simultaneously influencing each other, the true cognitive system is a single unified system embracing all three” (van Gelder 1995, p. 373).

van Gelder clearly endorses a systemic approach to dynamical systems. Unlike Beer, van Gelder is open to representational properties and states, arguing that “there is nothing preventing dynamical systems from incorporating some form of representation; indeed, an exciting feature of the dynamical approach is that it offers opportunities for dramatically reconceiving the nature of representation in cognitive systems, even within a broadly noncomputational framework” (van Gelder 1995, p. 376). van Gelder allows that (1) may be true, and depending how representations are incorporated into the dynamicist view, (2), (3), and (4) may be true as well. He explicitly endorses (7), like
Beer, and rejects the classicist interpretation of (5), that mental processes are symbol transformations. van Gelder argues that cognitive systems are composed out of the coupled dynamical interaction of the brain, body, and world. Insofar as each of these is a distinct component of minds, then the mind is componential in this gross-grained sense. However, the central physical cognitive systems such as brains are not to be decomposed into their components, as required by the componential dynamicist, instead remaining as un-decomposed wholes that couple with the environment and body in particular ways, and so the approach is a systemic dynamicist one, rejecting (8).

In this dissertation, I aim to construct a theory of cognition that firmly endorses some of the theses of systemic dynamicists, but also some of the theses endorsed by connectionists and classicists. This theory will accept (8), the componentiality thesis, arguing that cognitive systems are built out of implemented dynamical mechanisms. Thus, my theory is a componential dynamicist theory of cognition. The theory rejects the extended thesis about cognition (7), placing it somewhat at odds with much of the current dynamicist approach, as exemplified in Beer’s or van Gelder’s approach. The theory is also neutral on the status of representation (1), and, due to the avowed neutrality regarding (1), is also neutral on whether mental states have a combinatorial semantics and syntax (2), whether the semantic properties and syntactic properties of mental states occur at the same level (4), and whether mental transformations are
sensitive to the semantic and syntactic properties of mental states (6). The nature of
dynamical theories of cognition more or less forces a rejection of (3), that symbols must
be narrow, discrete, digital, invariant, and atomistic, as dynamical systems are typically
variable, typically non-digital, typically non-discrete, and so forth. However, since the
theory is agnostic about (1), the status of representation, I won’t be discussing what
symbols must look like on such a view. On the view I will defend, mental processes are
transformations over mental states (5), but the types of transformations are more like
connectionists state-space trajectories described by state evolution equations for the
components of the system than the input-state, state-output representational
transformations endorsed by the classicist.

To better situate my assessment of the conceptual terrain, compare my
characterization of different approaches to cognition to Horgan and Tienson’s analysis
(Horgan and Tienson 1994). Horgan and Tienson define classical cognitive science as
adherence to a set of five theses. Their alternative framework rejects their thesis (5),
“[h]uman cognitive transitions conform to a tractably computable cognitive transition
function” (Horgan and Tienson 1994, p. 309) and their thesis (3), “[c]ognitive processing
conforms to precise, exceptionless rules, statable over the representations themselves
and articulable in the format of a computer program” (Horgan and Tienson 1994, p.
308). Their framework retains their thesis (1), “[i]ntelligent cognition employs
structurally complex mental representations” and their thesis (2), “[c]ognitive processing is sensitive to the structure of these representations (and thereby is sensitive to their content)” (Horgan and Tienson 1994, p. 308), as well as their thesis (4), “[m]any mental representations have syntactic structure” (Horgan and Tienson 1994, p. 309). Their motivation for searching for a new framework is that the algorithmic nature implied by their theses (3) and (5) prevents the classical paradigm from handling global properties of cognition, like relevance, revision, simplicity and other attributes of the belief system. Their strategy is to appeal to dynamical systems theory as an alternative mathematical framework for connecting the cognitive level to the physical implementation level.

The new framework they propose has three levels: the Cognitive State-Transitions, Mathematical State-Transitions, and Physical Implementation (Horgan and Tienson 1994, p. 320). Cognitive State-Transitions are at the level of “mental qua mental, [where] the cognitive system has general dispositions to evolve from one total cognitive state (TCS) to another in content-appropriate ways” (Horgan and Tienson 1994, p. 320). The Mathematical State-Transitions refer to how “[c]ognition is mathematically subserved by a dynamical system, under a realization relation linking TCSs to points in the dynamical system’s state space…. The mathematical states and state-transitions subserving cognition need not… constitute an algorithm that computes cognitive transitions” (Horgan and Tienson 1994, p. 321). Finally, the level of Physical
Implementation, where “the dynamical system is subserved by a neural network of some sort…. Points in the state space of the dynamical system are realized by total activation patterns in the associated network…. The dynamical system is thus a high-dimensional activation landscape” (Horgan and Tienson 1994, p. 321). Two points are made about the physical implementation level: first, that physical systems in nature can realize enormously complex dynamical systems, and second, connectionist networks can subserve this complexity.

Horgan and Tienson focus on the systemic dynamical system analysis. Fundamentally, “cognitive processing is… construed as the system’s evolution along its activation landscape from one point in activation space to another—where at least the beginning and end points are interpreted as realizing intentional states...” (Horgan and Tienson 1994, p. 319). From the dynamical systems perspective, the system starts at some point in an attractor basin and arrives at the point attractor. This trajectory corresponds to the cognitive processing. Though they focus on the total cognitive system, referring to total cognitive states (TCSs) and cognitive transition functions (CTFs) defined over the TCSs, “…a total connectionist network can consist of distinct, though interacting, subnetworks; and hence can be mathematically analyzed as subserving separate, though coupled, dynamical systems” though the component dynamical systems “are embedded
in a larger, higher-dimensional, total dynamical system…” (Horgan and Tienson 1994, p. 319). Thus, they do acknowledge the possibility of distinct subnetworks.

Horgan and Tienson’s approach differs from my view in three distinct ways. First, they contrast classicism with a hybrid, dynamicist-connectionist approach, whereas on my view, dynamicism and connectionism need not go together. On their view, dynamical systems are at a higher level, the Mathematical State-Transitions level. They assimilate dynamical systems to dynamical systems theory, a branch of mathematical analysis that utilizes differential equations to describe the evolution of a system through time or with respect to changes in a system variable (cf. Strogatz 1994). But these equations must be interpreted, assigning referents to the mathematical variables, and that interpretation could reflect physical entities or quantities, like neurons or logic gates, as well as functional entities or quantities, like network nodes or Turing machine states. When given a physical interpretation, dynamical systems are no longer merely mathematical descriptions, but interpreted so as to refer to physical entities. So, dynamical systems analyses can be given of physical devices, including classical computational devices, as well as of functionally defined entities. Thus, dynamical systems, be they componential or systemic, need not be connectionist systems.
Second, their view of dynamical systems is fundamentally different from mine. They view dynamical systems as identical to the branch of mathematics called dynamical systems theory. I see dynamical systems as distinct from the mathematical tools used to analyze them. Dynamical systems are described using dynamical systems theoretic concepts and mathematics, but that group of concepts and formal tools can be applied to many kinds of systems, including classical computational systems. Dynamical systems are better contrasted with discrete systems like Turing machines or various computational automata, which can also be described using dynamical systems theory.

Third, their approach glosses over the distinction between componential and systemic dynamical analyses. They acknowledge the possibility of subnetworks, but these are not yet components. The componential analysis of connectionist systems finds a uniform set of components, with the more complicated dynamics arising from the way those components are interconnected and interact. But componential dynamical accounts of the sort I develop herein contain a diversity of components. In particular, the componential approach attempts to find a number of distinct types of dynamical systems that are components of cognitive systems and that occur across such systems. The subnetwork decomposition discussed by Horgan and Tienson is neutral about the scope of the decomposition: on their view, different subnetworks may be found for each token distinct connectionist network. The componential approach implies a stronger
claim than this sort of decomposition into uniform parts. The componential approach is not merely an avowal of the possibility of a decomposition of a system. Rather, the componential approach is a prediction about how the science of cognition will proceed: cognitive science will identify a finite set of diverse components that are present in a wide range of cognitive systems.

1.3 Conclusion

Here, I have painted a picture of the conceptual landscape with a very broad brush. I discussed classical approaches to cognition, which view cognitive systems as representational, componential, narrow, and often utilizing input-state-output transformations. I contrasted classical approaches with connectionist ones, which also view cognitive systems as representational and narrow, but often not componential nor utilizing input-state-output transformations. Finally, I discussed dynamical approaches, which often view cognitive systems as non-representational or a-representational, extended, embodied, non-componential and not utilizing input-state-output transformations.

Next, I distinguished two types of dynamicist approach, systemic and componential. Most extant dynamicist approaches are systemic, and are characterizing by the theoretical commitments just related. In contrast, componential dynamicist approaches view cognitive processes as being either representational or non-
representational, narrow (not extended), either non-computational or computational, either embodied or not, not utilizing input-state-output transformations and, of course, componential.

In the remainder of this dissertation, I will argue that recent research into the physical bases of cognitive processes is consistent with such a componential dynamicist view. Cognitive systems are composed of many tokens of different types of dynamical mechanisms. These dynamical mechanisms are the Enigmas of thought: they execute the formal models of processing, the cognitive functions, and they are implemented in turn by physical mechanisms. These dynamical mechanisms are reused for different cognitive functions, and they can be implemented by different physical mechanisms, even radically different ones. These dynamical mechanisms are truly the mechanisms of cognition.
2 Dynamical Mechanisms in Mind

In this chapter, I will be laying the foundation for making good on the dynamical ideas presented above. I begin by discussing several active research programs in electrophysiological cognitive neuroscience. These case studies are conceptual fodder for developing a componential dynamicist theory of cognition, serving as empirical benchmarks and providing theoretical motivation for developing the theory. In the course of discussing these case studies, I will discuss how these research programs invoke dynamical systems, defined below, in explaining the various cognitive phenomena under investigation. I then define execution and implementation, two central notions in my theory of cognition.

Next, I argue that these dynamical systems are in fact a species of mechanism. I present what I take to be the current consensus on mechanisms, the organized interaction of a collection of spatiotemporally characterized entities and their activities. This definition excludes the dynamical systems at the heart of cognition. I turn to an analysis of what makes a mechanism different from mere aggregates or from other emergents, beginning with Wimsatt’s analysis of aggregativity and considering other constraints on organized systems. I conclude with a revised definition of a mechanism, one that is less conservative than the contemporary consensus and that classifies the dynamical systems that constitute cognitive systems as dynamical mechanisms.
2.1 Two Case Studies in Cognitive Neuroscience

Two case studies in electrophysiological cognitive neuroscience will serve as the main empirical illustrations of the dynamicist theory constructed herein. The first case study is of perceptual decisions under noisy sensory conditions, and the second is of strategic decisions about allocating time between exploiting resources and exploring for them. The same dynamical system, called an integrate-to-bound mechanism, is implemented when making these perceptual or strategic decisions. I also briefly discuss the relationship between these dynamical mechanisms and the formal models of processing that those mechanisms execute. The discussion of these case studies will provide two central empirical examples for understanding how neuronal dynamics execute the processes relevant to decision making. I contend that the lessons learned from these case studies extend to other cognitive processes. Though I call this integrate-to-bound system a mechanism, making the case for its mechanism status will require lengthy discussion after presentation of the case studies.

2.1.1 Case Study 1: Cognitive Mechanisms of Perceptual Decisions

The first case study concerns the perceptual and decision mechanisms for decision-making under noisy conditions, when animals make decisions with uncertain

---

1 Appendix A contains a discussion of these two case studies at much greater length.
sensory evidence. This case study will illustrate the distinction between the formal model of the behavior, the dynamical system that executes the formal model, and the physical mechanism that implements the dynamics. Rhesus monkeys (*Macaca mulatta*) are presented with a visual display of moving dots, only some fraction of which move in a particular direction, e.g. either left or right (“motion strength”). The noisiness in the signal is due to the fraction of dots that are not moving in the same direction (moving coherently), and different fractions of dots move coherently on different trials. To make a decision, the animal looks at a target that is in the direction of the perceived motion of the dots, receiving a juice reward for a correct decision. Monkeys modulate their behavior on the basis of the motion strength, showing greater accuracy and faster response times for stronger evidence (Roitman and Shadlen 2002).

A formal model of processing known as the drift diffusion model (DDM) describes an evidence sampling process for determining the direction of motion (and, in fact, does so optimally; Bogacz et al. 2006) for making a decision during the random dot motion task (RDMT). The DDM starts with a prior log-odds ratio, the ratio of the prior probability of the dots moving left or right, and then computes the likelihood fraction for the observed evidence over some period of time, adding this weight of evidence to the priors to arrive at a posterior log-odds ratio. The evidence here is the motion signal

---

2 See (Gold and Shadlen 2001, 2002, 2007) for extensive discussion of this research.
over some small period of time (Shadlen and Newsome 2001). If a threshold for the posterior ratio has not yet been reached, the process iterates, until a threshold is reached, the system runs out of evidence (e.g., if the dot display disappears), or some other stopping criterion is met. Once the process halts, a decision is made according to some decision rule, such as on the basis of which threshold was reached, or on the basis of which alternative direction of motion has the greater amount of evidence.

Investigation of the neuronal mechanisms of this noisy perceptual decision have revealed the presence of a particular set of dynamical properties, the integrate-to-bound system. The lateral intraparietal area (LIP), an eye movement (saccade) control region in the parietal cortex (Platt and Glimcher 1999), exhibits stereotyped dynamics during this decision process (Roitman and Shadlen 2002; Gold and Shadlen 2007). LIP cells receive projections from area V5/MT (MT) (Britten et al. 1992), and area MT cells encode the strength and direction of motion in a part of the visual field (Britten et al. 1993).

---

3 The priors and the threshold values are variable, though in analyzing the data from the RDMT, the DDM is typically modeled with a fixed prior and threshold.
Figure 3: Average recorded firing rates from cells in the parietal cortex of monkeys while they make a perceptual decision in different evidential conditions. Adapted from Roitman and Shadlen 2002, p. 9482.

The recorded electrical activity from LIP shows three relevant dynamical properties (figure 3). First, these cells have a ‘reset’ point or baseline firing rate just after motion onset (time 0 in the plot on the left in figure 3). Second, the activity of cells in LIP varies with the motion strength of the stimulus, with steeper increases in firing rate for stronger evidence (the different colored lines in both plots). Third, the activity of these cells converges on a common firing rate across different motion conditions just prior to
the animal initiating a saccade (time 0 in the plot on the right). This pattern of activation partly constitutes an integrate-to-bound dynamical system: a baseline or starting point, a period of integration, and a threshold or boundary.

Finally, note that this dynamical pattern maps on to the formal model of processing, the DDM. Recall that the DDM consists in an evidence sampling and summation process that starts with a prior log-odds ratio, samples the evidence to form a likelihood ratio, and then adds this to the prior to obtain a posterior log-odds ratio. This summation proceeds until some stopping criterion, such as a threshold, is reached. First, the DDM starts with a prior log-odds ratio that is the same across different strengths of evidence, just as LIP cells exhibit a baseline at the beginning of presentation of the motion stimulus. Second, the sequential evidence sampling process at the heart of the DDM is mapped on to the differential increases in activity in LIP cells that encodes the strength of the motion stimulus. Finally, third, the DDM evidence sampling process halts when a threshold is reached, also the same across different strengths of evidence, similar to how LIP cells rise to a common value across different motion conditions.  

This case study illustrates the three components that are central to explanations of decision processes, as well as the relations between them. First, there is a formal model of the decision process, which in the case of the RDMT is the DDM, and this

---

4 A variable threshold is possible.
formal model is executed by a dynamical system. Second, there is a dynamical system, which in the case of the RDMT is the integrate-to-bound system, and this dynamical system is partly implemented by the neurons in LIP. Shortly, I will argue that this system is a mechanism. Third, there is the physical mechanism, the neurophysiological components and activities in neurons in LIP. In this dissertation, I describe the nature of these three components, provide an account of the execution relation between the formal model and the dynamical mechanism, and provide an account of the implementation relation between the dynamical mechanism and the physical mechanism.

The dynamics of this neuronal mechanism are key to the execution of the formal model. The dynamics of LIP neurons’ integration of perceptual evidence executes the DDM. The dynamical properties of the LIP neurons are a dynamical system. To understand what is meant by a dynamical system, a brief interlude is required.

2.1.2 State-Spaces, System Evolution Functions, and Dynamics

We can define a system as a collection of entities that evolve over time, according to their system evolution function. A system evolution function specifies the way the system properties and the properties of the system components change over time or

---

5 This sense of ‘evolution’ is not Darwinian evolution, though such systems may be subject to selective processes.
with respect to each other. Systems can be discussed and reasoned about by considering the mathematical description of the evolution of the system over time, as succinctly presented in the state evolution equations of the system. Often these equations are differential equations, operating over continuous time, but they can also be difference equations, operating over discrete time. (For a mathematical presentation of a view similar to this, see Strogatz 1994.) The state evolution equations specify how the variables that stand for the components of the system evolve as a function of time, various system parameters, and other components of the system, including environmental variables such as input into the system. The state evolution equations are mathematical descriptions, specifically formulae; the system evolution function is the interpreted counterpart to those equations taken altogether.

As an initial example, consider a simple switch. Suppose the switch can be in one of two states, 0 and 1 (or, off and on), and that the switch flips from 0 to 1 or 1 to 0 depending on the value of some input variable x, such as the amount of light in a room, as determined by a photodetector. If the photodetector determines the amount of light is equal to or below some constant c, the switch S goes into state 0 (‘off’), and if it is above

---

6 There may be some concern about shifting between talk of states and talk of evolving systems, especially for systems evolving over continuous intervals. In Appendix B, I show how talk of states translates into talk of continuous intervals.
that value, S goes into state 1 ('on'). This change in the state of the switch corresponds to its system evolution function, as described by its state evolution equation

\[
f(x) = \begin{cases} 
1 & \text{if } x > c \\
0 & \text{if } x \leq c
\end{cases}
\]

where \( f(x) \) describes the state of the system. S is very simple: there are two states, 1 and 0; S must be in one or the other; and in this case S’s state, \( f(x) \), is a function of the input \( x \).

But we could complicate S by adding a second switch, also able to be in state 1 or 0.

Then we would have doubled the number of states the system can occupy: 1 1, 0 0, 1 0, 0 1. The state space of S contains 4 states. The second switch may have a different cut off from the first switch. Let’s suppose that the cut off is greater than the first switch. Each switch will have as its state either 0 or 1, as determined by its own system evolution function, described by the state evolution equation. Letting \( f_1(x) \) be the state of switch 1 and \( f_2(x) \) be the state of switch 2, the state evolution equations are

\[
f_1(x) = \begin{cases} 
0 & \text{if } x \leq c \\
1 & \text{if } x > c
\end{cases}
\]

\[
f_2(x) = \begin{cases} 
0 & \text{if } x \leq d \\
1 & \text{if } x > d
\end{cases}
\]

for \( d > c \). For example, consider how this simple system evolves. First, for zero light level, they both start off {0 0}; as the light level rises, first one turns on {1 0}, and then the

\[7\] S’s state space, the set \{0 1\}, is a complete partition over S’s input \( x \), the amount of light detected by the photodetector. There is no value for \( x \) that does not determine the state for S. S’s state space might have executed a partial partition if there were values for \( x \) for which S’s state was undefined.
other \{1 1\}. Since one cut off is higher than the other, one of the states, state \{0 1\}, can never be occupied. This system is composed of two components, the two switches, each of which can occupy one of two states 0 or 1, and these two switches’ behavior is described by the state evolution equations $f_1(x)$ and $f_2(x)$. We can concatenate these state evolution equations together to form a total state evolution equation for the system composed of both switches.

With the notions of a state space and system evolution function in hand, we can now return to our initial example of an integrate-to-bound system.\(^8\) This system is called a one-dimensional attractor with a single stable fixed point in dynamical systems theory: the system is drawn toward one particular point in its state space (Strogatz 1994). The integrate-to-bound dynamical system responds to input by changing its state until a particular boundary is reached, at which time the system resets. This change in state is an integration: a trajectory through the system’s state-space toward an attractor point. Once that value is reached, the system resets to its initial state. From this initial state, the system could begin the integrative activity anew. The components (at least) include a baseline, the starting point for the system, and a threshold, the point in the system’s state-space where it returns to the baseline. These activities (at least) include resetting,

\(^8\) For an interesting contrast, this discussion can be compared to Eliasmith’s recent treatment of neural integrators (Eliasmith 2013, p. 43ff). Eliasmith is concerned with understanding cognition from an emphatically biological, and specifically neural point of view, and his discussion of neural integrators is correspondingly focused on the biological details.
when it the system returns to the baseline state, and thresholding, as the system approaches and cross the threshold. The system also exhibits complex behavior, including integrating to a threshold, beginning to integrate, resetting to baseline, and so forth. All of these components and activities acting and interacting over time taken together define the system, and the way they change over time or with respect to each other define the system evolution function.

There are many different ways to build such integrate-to-bound systems from simpler components, even switches, and examining the different possible ways of building these systems provides insight into the criteria for being such a system. Imagine we have a bank of switches. They are all wired to some input device, such as a photodetector. Each of these switches has a different flipping value, the value at which they flip; above this value, they flip on, and below it, they flip off. At first, none of the switches are flipped. As the input rises, first one switch flips, then another, and so on until they have all flipped. In aggregate the system exhibits a continuous trajectory as the switches are flipped. The system is also encoding the input, as the level of the input correlates with the number of switches that have flipped. So far, the system has no bound. The switches will flip as the input increases, and they will flip back once the input drops. However, we could specify a more complex switch that, once flipped, stays flipped until a different signal is sent to it. This is one way a threshold is built.
In our more complicated system, we can imagine the same bank of switches. But now they have a more complex state space. Instead of flipping back and forth between states in response to changes in the input, the switches, once flipped, stay flipped until a separate signal is received to reset the switch. Once a switch is flipped, its flipped state encodes that the input crossed a particular value (the activation value for that switch), regardless of the current input level, until the switch is reset by an outside signal. Now imagine that the switches, which stay flipped until reset, don’t get reset unless they are all flipped. This is the threshold, and the presence of such a threshold establishes an attractor point and the trajectory can now be classified as integration. In order to describe the thresholding function, additional terms in our state evolution equations are necessary. For example, suppose an additional mechanism sends a reset signal once a certain time has passed, a periodic reset, which resets the switches, resulting in a new system evolution function. We can also add a mechanism that flips all the switches off once the last switch is flipped: this sets the threshold at the level of activation for the last switch.9

--------------------

9 The system as described is overly complicated, however, since if the level of activation for reset is just the last switch’s level, then why have all the superfluous switches at the lower levels of activation?
The array of switches can be arranged such that the threshold is adaptive.\textsuperscript{10} Adaptive thresholds are akin to our simple system’s reset mechanism, but the particular value at which the reset signal gets sent can change. For example, perhaps we want our reset signal to occur at a lower threshold in one environment (e.g. a dark room) than another (a light room). Or, perhaps some internal aspect of the system modulates the threshold setting. Either way, the threshold is adjusted, which changes when the reset signal is sent to the array of switches.

With only ten switches, our ‘integration’ is more aptly described by a Heaviside step function than a smooth ramp up. However, there is nothing essential about that particular type of system evolution function, and note that in the system’s state-space, the trajectory is continuous, with each transition between states passing through any intermediate state as determined by some ordering of the states. The flip transition function could be binary, or it could be smooth. As the number of switches increases, with each switch mapping on to a different flip value from the others, the ramp up will look more and more smooth. A smooth evolution function can also be an integration: the threshold could be set at an arbitrary value. Likewise, there’s nothing special about the aggregate number of flipped switches, as we could interpret a set of serial switchers as a

\textsuperscript{10} An adaptive threshold mechanism is hypothesized as the underlying mechanism for what is known as the speed-accuracy tradeoff in decision-making (e.g., Bogacz et al. 2008) and such an adaptive threshold has been observed in monkeys on the RDMT (Roitman, personal communication). By dynamically adjusting the threshold for a response, a system can reply quicker with more errors, or slower but more accurately.
ramp up. Imagine that, instead of our row of 10 switches flipping on in succession as the
input grows, that the switches flip and stay on only if the input is in a certain range,
flipping off if the input lies outside that range, with non-overlapping ranges for each
switch. Then the value of the input can be read off directly from which switch is
currently flipped.

The integrator as described integrates with respect to the input, but integration
can also occur with respect to time. Imagine the same sort of system as before: a row of
switches. However, instead of tracking the value of the input, the switches track how
long the input has been on (or, if you prefer, how long the input has had a particular
value or range of values). The longer the input has been on, the larger the number of
switches that have been flipped. The system is now temporally integrating the input.
Suppose we want our system to execute a further function once the light has been on for
a certain amount of time (such as, turn the light off). Then we can design our system
such that the switches all flip off if the light goes off, and if the last switch (or n switches)
have flipped, an output signal is sent (which turns off the light). Now our integrator
integrates to a temporal threshold in order to change the world around it. This simple
machine augments the world around it once the world attains a particular state; in particular, our simple machine will turn off the light if it has been left on too long.\textsuperscript{11}

Clearly, there are many different combinations of system evolution functions, that is, different combinations in changes of the properties of the system or its individual components, that together compose an integrate-to-bound system: tying the reset to the identity of the switches (‘wait for the last’ or ‘wait for the n\textsuperscript{th}’), looking at the overall activity in the switches (‘reset once n have flipped’) and so forth. These are all different ways of building an integrate-to-bound system. Above, I defined the system evolution function as the description of the way the components and properties of the system

\textsuperscript{11} The integrator system is relatively simple, meant to illustrate the dynamics of a simple system. By coupling together integrators (or other simple systems), however, very complex dynamics can emerge. Imagine, for example, that on top of our light integrator, which switches off the light, we add a second integrator, which switches on the light if it has been off for a period of time. That second integrator integrates the absence of a signal over time, and there is a coupling between the two integrators, as they switch the light on and off. Instead of a second integrator for how long the light has been off, imagine instead we couple a second integrator to the first. Every time the first integrator thresholds, it sends a signal to turn off the light. In addition, suppose we have that signal feed into our second integrator. This second integrator now keeps track of how many times the first integrator has ‘integrated’, that is, how many times the ramp up has reached the barrier or threshold. This second integrator could also have a threshold, and upon reaching it, an output is sent to another mechanism that (let’s say) turns down the multiplicative gain on the input to the first integrator. Turning down the gain reduces the efficacy of the input to the first integrator by lowering its value by a multiplicative constant. The effect of this reduction in signal is to have the first integrator take a longer amount of time to ramp up, because the slope of the ramp up is lower for a multiplicative gain. We could have an additive gain change as well, where the time to ramp to threshold might be the same, but the beginning of the ramp up would be different, starting later if the additive gain change is negative (i.e., subtractive). The function of our little system (function being a critical topic which will reappear later) is to shift the system’s light controller into a new control regime. Suppose the light is often flipped on: as the first integrator continually ramps up, it pushes the second integrator to threshold, which modulates the time to reach threshold or the time of integration initiation of the first. The effect is to reduce the number of integration events that the first goes through, thus resulting in a longer delay period before the light is turned off. Again, these integrators are coupled, exhibiting an interaction where the activity of one integrator affects the activity of the second. Complex dynamics can thus arise from concatenating or combining many such simple systems.
evolve over time or with respect to each other. What do all integrate-to-bound systems share? They all share a number of functionally defined components like a baseline, an activation point specifying where in the state-space the system lies, a threshold. They also share a number of functionally defined activities like resetting, thresholding, and integrating. Generally, on my account, two dynamical systems are type-identical if they share system evolution functions, a specification of how the properties of the system and system components evolve over time. In the case of integrate-to-bound systems, the system evolution function does not specify what plays those functional roles. Two dynamical systems are both integrate-to-bound systems then if they have both have baselines, integration paths towards an attractor point, a reset following the arrival at the attractor, and so forth.

To summarize, the dynamical systems that are components of cognitive systems are like simple mechanisms consisting of a set of components, the activities of those components, and the behavior they exhibit, all implemented in one of a variety of possible ways by physical mechanisms of some sort. The behavior of the system, how it wends its way through the state-space of the system, is described by the state evolution equations, the mathematical description of the system’s components and activities. These state evolution equations are a mathematical description of the system evolution

---

12 I define implementation later in this chapter. Of course, my use here of ‘mechanism’ to describe these component dynamical systems is loaded: I make the case for these systems being mechanisms below.
function, which describes how the system’s properties and the properties of its components change over time or change with respect to each other. The whole system’s state evolution equation mathematically results from all of the state evolution equations of all the different components (the componential state evolution equations) combined. There are many different ways to combine simpler components into more complex machines that implement a given state-space structure, as illustrated by the different ways to combine simple switches to implement an integrate-to-bound circuit. Furthermore, implementing such a system can be done with different combinations of components and only one basic componential system evolution function or with a mix of different components and different componential system evolution functions.

2.1.3 Integrate-to-bound Systems and Perceptual Decision Mechanisms

Returning to our empirical example, neurons in LIP implement an integrate-to-bound system, which in turn executes the DDM. (Below, I argue that this dynamical system is in fact a dynamical mechanism.) The formal model of processing during the RDMT, the DDM, is executed by the integrate-to-bound dynamics of the neuronal mechanism in LIP, a gradual change in the state of the neuron through a stereotyped sequence until some threshold state is reached, at which time the neuron resets to a

---

13 The conceptual apparatus of state-spaces and system evolution functions can be used to describe brain functioning more generally than at the lowly level of individual neurons. Large-scale brain dynamics appear to exhibit transitions between states, as described by system evolution functions.
baseline state. As the system receives input, it traverses its state space in an input-dependent trajectory that settles into the point attractor at different times correlating with the quality of evidence in the motion stimulus. An integrate-to-bound dynamical system minimally consists in a starting point or baseline, a period of integration, a threshold, and a reset, as discussed above. The motion signal encoded in MT is an input onto the LIP neurons, and LIP neurons integrate this motion signal over time until some threshold is reached and a decision is executed. The observed increase in activity appears to correspond to the mathematical function of integration, though a pure integration function is biologically unrealistic and a mixed integration and exponentiation function may be more plausible when specifying the state evolution equation.\textsuperscript{14} While integration describes the changes in the neuron’s firing rate, it captures

\textsuperscript{14} The mathematical description of the dynamical system, present in what I call the state evolution equations, has been framed as an integration, an exponentiation or a mixed computation (Mazurek et al. 2003; Usher and McClelland 2001; Wang 2002; Carandini and Heeger 2012). Carandini and Wang both discuss the neurons as computing an exponentiation function (Carandini and Heeger 2012; Wang 2002). In a case of pure exponential growth, for a given strength of evidence, the relative change in the firing rate, or the change in firing rate divided by the current rate, is constant (Stewart 2011). However, upon removal of the stimulus, LIP neurons will not maintain their firing rates indefinitely, and adding state-dependent leak and recurrent excitation exponential decay terms captures this decay (Usher and McClelland 2001; Wang 2002). Furthermore, the integration is noisy, requiring the addition of a noise term. Thus, neurons such as those in area LIP are often seen as ‘noisy leaky integrators’; their firing rate is a function not only of the change in firing but also contains a noise term and a loss (or ‘leak’) term (e.g. see Usher and McClelland 2001 or Wang 2002; for a philosophical discussion, see Eliasmith 2013, p. 43ff). The leak is dependent upon the state of the system, and so the internal system dynamics contribute to the state of the system. Wang’s model is even more complex and is discussed below, but in short contains an additional recurrent excitation term that is driven by the state of the system. These noise and leak terms effectively curtail the integrative growth function as well as enforcing a biologically plausible decay in the firing rate at some time constant.
neither the initial starting point component nor the thresholding activity that results in a decision. These additional components and activities require their own neural mechanisms. Nonetheless, the dynamical integrate-to-bound system describes the relevant parts of the neuronal mechanism observed in LIP during the RDMT for executing the DDM.

The previous discussion suggests that there are different ways for LIP neurons to implement such a system, and indeed there is currently investigation into the precise nature of the implementation. Do LIP neurons proceed through a cascade of discrete jumps in state, akin to a multi-step Heaviside function? Or is there a large jump in state occurring at different time points, as though the LIP cells are executing a binary operation, and averaging over the cell responses results in the integration activity? Or is absent any driving force, while the differences in the strength of evidence correspond to differences in the relative growth rate constant driving the neuron’s activity.

Lo and Wang (2006) model the thresholding operation as resulting from a ramp up in the firing rate of cortical cells that causes caudate neurons to fire, inhibiting the substantia nigra pars reticulata and hence disinhibiting superior colliculus cells to burst, resulting in a saccade. This ‘inhibition of inhibition’ model is similar to models of the cortico-basal ganglia-thalamo-cortical loop, widely deployed by neuroscientists to understand how cortical activity results in action, and situates the integrative activity in LIP neurons in a broader theory of the neuronal mechanisms of behavior.
there in fact a slow ramp-up in activity? All of these are possibilities for implementing the integrate-to-bound system.\textsuperscript{16}

The case study of perceptual decision suggests several important principles. The remainder of the dissertation discusses these principles at length. First, the cognitive systems and the formal models of the behavior are distinct. Despite the apparent similarity between the dynamical system and neuronal mechanism in LIP and the DDM, it is important to conceptually separate them from the DDM. The cognitive systems faced with these kinds of decision problems might execute alternative formal models, such as by deploying simple choice heuristics (Simon 1957; Simon 1990; Gigerenzer et al. 1999).\textsuperscript{17} Since the integrate-to-bound system refers to a specific dynamical pattern found in the physical system, they could be utilized to execute models other than the DDM, such as the formal models characterizing those heuristics. Second, the dynamics of the physical mechanism constitute a complete, intermediating execution of the formal model, and these dynamics are in turn implemented by the physical mechanism. No

\textsuperscript{16} Recent evidence uncovers a diversity of response profiles at the level of individual LIP neurons, suggesting that the population responses, such as those depicted in Figure 3, may be the result of the aggregation of the response dynamics of a large neuronal pool (Premereur et al. 2011, Meister et al. 2013). Other data indicate that single neurons exhibit a slow increase in firing within individual trials, mirroring the population response and suggesting that LIP neurons do not act like binary switches (Bollimunta et al. 2012).

\textsuperscript{17} In particular, there may be reasons to question whether optimal models are the models executed. For example, there are motion discrimination conditions in versions of the RDMT in which monkeys perform suboptimally per the DDM (Kiani et al. 2006). Simple choice heuristics such as satisficing, first past the gate, and so forth are all simple but suboptimal formal models that may better describe the system’s behavior.
other properties are required for the execution of the formal model but those present in
the dynamical system; hence, the execution of the formal model is complete. And since
the dynamical properties are all that is needed, and no other properties of the physical
mechanism are required, the dynamical system stands as an intermediating system,
between the formal model and the physical mechanism. Third, the physical mechanism
implements the dynamical system. These three components are central to my theory of
cognition.

Neuroscientists and cognitive scientists have uncovered the integrate-to-bound
dynamical system executing or helping to execute a wide range of functions, like
perceptual decision-making across sensory modalities, strategic decision-making (as
discussed momentarily), and motor control. Additionally, many different types of
physical mechanisms, even potentially reaching non-neural mechanisms, implement the
integrate-to-bound dynamical system.\textsuperscript{18} To further illustrate the difference between the
formal models of the behavior, the dynamical systems that execute them, and the
physical mechanisms that implement the dynamics, consider another case study, from

\begin{footnotesize}
\begin{enumerate}
\item Integration of evidence and a concomitant integration activation function in neurons has been taken to
underlie a variety of perceptual decisions, including disparity discrimination (Uka and DeAngelis 2003,
2006), olfactory discrimination (Uchida and Mainen 2003; Kepecs et al. 2006; Uchida et al. 2006), face/house
discrimination (Heekeren et al. 2004), and vibrotactile perceptual decisions (Romo et al. 2002, 2004;
Hernandez et al. 2002). Integration has been most famously used to explain how the brain keeps the eyes
still (Robinson 1989; Seung 1996), and integration has been hypothesized to serve as a substrate for how bee
swarms select nest sites (Seeley et al. 2012), as will be discussed later in more detail.
\end{enumerate}
\end{footnotesize}
the neural mechanisms of strategic decisions. Research into these strategic decision mechanisms serves as an additional illustration of the points made above regarding the case study of perceptual decision mechanisms, as well as empirically demonstrating that diverse physical mechanisms can implement type-identical dynamical systems.

2.1.4 Case Study 2: Cognitive Mechanisms of Foraging Decisions

The second case study concerns a class of strategic decisions, decisions related to exploration or exploitation, allocation of behavior or resources across time, and the like. When foraging in an environment where rewards are clustered in patches, animals must determine when to leave the current depleting patch to travel to a new one: the patch-leaving problem. In order to investigate the neural mechanisms of patch-leaving decisions, Hayden and colleagues (Hayden et al. 2011) devised a simulacrum of the patch-leaving problem suitable for neural recordings, the patch-leaving task.

![Figure 4: The patch-leaving task. Adapted from Hayden et al. 2011, p. 934.](image-url)
In the patch-leaving task, rhesus macaques are presented with two choices: a small blue rectangle (‘stay in patch’ choice) and a large gray rectangle (‘leave patch’ choice) (Figure 4, adapted from Hayden et al. 2011, p. 934). Upon choosing to stay in the patch, the monkey receives a squirt of juice. On the next trial, if the monkey choose to continue to stay in the patch, a smaller squirt of juice will be delivered, and as the animal successively chooses to stay in the patch, smaller and smaller juice rewards are delivered, mimicking patch depletion. At some point, the rewards will decrease to the point where the monkey chooses to leave that patch. Upon choosing to leave, the monkey waits while the gray bar shrinks to nothing, without reward, mimicking the travel time between patches. Once the next trial starts, the locations of the two targets have switched sides, the reward associated with the stay in patch option has reset to the full value, and a new travel time is selected, as encoded by the height of the gray bar.

Like the case of perceptual decisions, a formal model describes the processing necessary for optimal patch-leaving decisions: the marginal value theorem (MVT) (Charnov 1976). The MVT determines the energy intake rate as a function of the value associated with the food item, the handling time for consuming the item, the average travel time between patches, and other environmental variables. Maximizing this rate results in a simple decision rule for leaving a patch: leave a patch when the
instantaneous energy intake rate in the current patch falls below the average intake rate for all patches in that environment.

In the patch-leaving task, monkeys show a significant if slight effect of travel time on patch residence time. On average, animals tended to stay longer in a given patch for longer travel times to a new patch. If the analysis of optimal leave times is calculated based on the whole environment, using the average travel time to a new patch, then this influence of travel time on patch residence time is suboptimal. However, monkeys appear to treat each patch as though it has been drawn from its own environment. If optimal leave times are calculated separately for each travel time, then monkeys obtained almost the optimal amount of reward (Hayden et al. 2011).19

19 This suboptimal behavior may be due to a noisy implementation of the MVT, or it may suggest that some other model would better capture the behavior. Wang (2012) interestingly suggests that deviations from optimality might result from the neural mechanisms used to implement optimal models.
Figure 5: Example neuron from the anterior cingulate cortex, recorded during the patch-leaving task. These cells exhibit a peri-saccadic response that increases over the duration of foraging in a patch. Adapted from Hayden at al. 2011, p. 935.

Given that patch residence time varies with travel time to a new patch, to investigate the neural mechanisms of foraging decisions Hayden et al. searched for a neural signal that encoded patch residence time and that varied with the travel time. Neuronal recordings from the anterior cingulate cortex sulcus (ACCs), a medial prefrontal cortical structure, revealed a different neurophysiological implementation of an integrate-to-bound system. The increase in patch residence time as the monkeys foraged in a patch are encoded by an increase in the peri-saccadic peak response in ACC neurons (see example neuron in Figure 6). The firing rate just prior to a decision rises over the course of the patch, akin to an integration. For similar travel times to a new patch, the firing rates in those neurons also rose to a common threshold for different patch leave times. Furthermore, for similar travel times, the initial firing rates at the beginning of the patch were the same. All three elements, of baseline, integration, and
threshold, present in the LIP implementation of an integrate-to-bound system are also present in the ACC data collected during the foraging task, suggesting the same system is implemented in both regions.\(^\text{20}\)

These two case studies illustrate the repeated implementation of the same type of dynamical system to execute cognitive functions, in these cases, decisions. Though different formal models characterize the behavior in the foraging and perceptual tasks, the brain executes these models by implementing two tokens of the same dynamical system, an integrate-to-bound system, albeit with different neurophysiological mechanisms. In the case of LIP, the system executes an integration of evidence function. The specific function executed by the dynamical system in ACC remains unclear, though assessments of opportunity cost or a comparison of instantaneous reward intake rates to average reward intake rates are among the possibilities. The total dynamics of the cognitive system appear to execute the DDM in the case of noisy perceptual decision-making and the MVT in the case of patch-leaving partly with the use of the same type of dynamical system. This dynamical system is implemented by distinct physical

\(^{20}\) Despite implementing the same dynamical system, the physiological mechanisms are different in LIP and ACC. In the case of LIP and the RDMT, integrative activity occurring on the timescale of hundreds of milliseconds during a single trial is implemented by an increase in firing by individual neurons. In the case of ACC and the patch foraging task, integrative activity occurring on the timescale of tens of seconds over many trials is implemented by an increase in the peak activity of individual neurons during a trial. These two different timecourses indicate distinct physiological mechanisms implement the integrate-to-bound system.
mechanism, LIP neurons during the RDMT and ACC neurons during the patch-leaving task.\(^\text{21}\)

### 2.1.5 Defining Execution and Implementation

Two important concepts in my theory of cognition are execution and implementation. Formal models of processing, such as the drift diffusion model (DDM) in the random dot motion task (RDMT) and the marginal value theorem (MVT) in the patch-leaving task, are executed by some subset of the set of dynamical properties of physical mechanisms present as components of cognitive systems. Below, I argue that the dynamical properties of these physical mechanisms satisfy the definition of a mechanism just elucidated, and thus dynamical mechanisms execute the formal models of processing. These dynamical mechanisms are implemented by physical mechanisms, the physical structures and properties of cognitive systems.

### 2.1.6 Formal Models of Processing and Model Execution

I will briefly discuss formal models and their execution by the dynamical properties of cognitive systems.\(^\text{22}\) This discussion, like that of component dynamical

---

\(^{21}\) I contend that there are many such cases of tokens of the same type of dynamical system being used in the execution of a formal model (divisive normalization, linear filtering, synaptic reverberation and center-surround inhibition are other examples). I discuss more examples later in the dissertation, as well as providing a definition of the use, and reuse, of a mechanism in a later chapter.

\(^{22}\) A much more in depth discussion of formal models, cognitive functions, and the justification of the ascription of cognitive functions to systems occurs in Appendix D.
systems, is motivated by current research programs in cognitive science, and though the focus in this dissertation is on the component dynamical systems of cognition, talk of formal model execution will occur repeatedly in the discussion below. As noted at the outset of this discussion, the first principle of my theory of cognition states that dynamical mechanisms execute the formal models of processing. Formal models of processing mathematically describe the entities and relations \textit{inter alia} that organisms must track for adaptive action. These formal models are formalizations of the cognitive functions of systems as they act and navigate their world.\footnote{I won’t be arguing for any particular view of cognitive functions, and these remarks are meant to be neutral about assimilating cognitive functions as causal role functions, selected effects functions, or some other type of function.} I will say that formal models of processing are executed by cognitive systems, as defined below.

Cognitive systems are faced with a range of behavioral challenges that reflect the particular exigencies they face in navigating and manipulating the world. Different challenges face different organisms, and the particular challenges faced by an organism reflect the ecological context within which the organism is acting. An animal in the forest will face the need to detect predators occluded by foliage, whereas an animal in the desert will face different predatory dangers. Humans have perhaps the most diverse environments to navigate. Most generally, perception, decision and other sorts of cognitive processing require the organism to keep track of objects in the environment,
the various rewards and uncertainties associated with actions, and so forth. Different objects will require different sorts of processing, different rewarding contexts will place different evaluative demands on the organism, different perceptual contexts will require different attentional commitments, and so forth.

Formal mathematical models describe the processing necessary to overcome these challenges. These formal models specify the variables that the system must encode, the mathematical relationships between these variables, and the way that these encoded variables must be transformed in order to accomplish the goal of the organism, to respond to the behavioral challenge (providing such an account Marr calls a “theory” of the relevant cognitive domain; Marr 1982; Shagrir 2010). In the case of the RDMT, animals integrate evidence from the environment in order to make a perceptual decision, and in so doing they implement the DDM. In the case of patch foraging, animals keep track of average and instantaneous reward rates, travel times between patches, and other variables, and in so doing they implement the MVT. In either case, there is a mathematical formalism that describes the variables in the environment the animal must keep track of and the relationships between those variables. Importantly, simply providing a mathematical function with uninterpreted variables and parameters
is insufficient to specify the formal model; the formal model must be interpreted, that is, there must be a referential assignment for the variables and parameters in the model.24

These formal models are distinct from though executed by the components of cognitive systems, specifically the dynamical systems that make up cognitive systems. The dynamical properties of the system result from physical processes in mechanisms that transform signals and execute the formal models of processing. The accumulation of evidence present in the DDM is executed in the integrate-to-bound dynamical system whose properties were revealed by research in LIP.25 Similarly, a mathematical function of comparing instantaneous and average reward intake rates or tracking opportunity

24 Such assignment has been discussed extensively by Pylyshyn amongst others (Pylyshyn 1984). Due to limitations on length, I won’t be able to delve into the interesting question of how to make such an assignment, nor on whether there can be intensional as well as extensional interpretations to variables.

25 I have intentionally chosen these elements corresponding to the integrate-to-bound mechanism. A prima facie objection is that while these dynamical processes work well e.g. for the transformations in the mathematical model, they do not play any role in the encoding of the variables present in the formal model, such as the travel time or average reward rates in the MVT or the prior probabilities or likelihood ratios in the DDM. This, however, is not true. Variables are encoded in neural systems using a number of different dynamical mechanisms, including monotonic rate codes, where the firing rate of neurons encodes the value of the relevant variables, place codes, where the firing rate of a neuron relative to its neighbors encodes the relevant variable, or population codes, where the firing pattern of a neuronal population encodes the relevant variables. In all of these cases, however, the dynamical activities of the neuronal processes are the relevant aspect for encoding. Though I won’t argue for this view herein, this implicitly assimilates neuronal encoding mechanisms into the class of dynamical mechanisms that compose cognitive systems on my theory. Incidentally, these encoding mechanisms, I would contend, are prima facie candidates for representational vehicles.
cost is partly executed in the integrate-to-bound dynamics of ACC.\textsuperscript{26} The first principle states that formal models are executed by dynamical mechanisms; arguing that the dynamical systems that are components of cognitive systems are mechanisms will require an analysis of the concept of a mechanism. Arguing that the dynamical mechanisms execute the formal model is an empirical induction from the various case studies of cognitive systems, such as those presented earlier.

How do the dynamics execute the formal model? Processing models are executed by dynamical mechanisms iff there is a mapping between the elements of the model and the elements of the mechanism. If there is such a mapping, I will say that the formal model is equivalent to the dynamical mechanism. There are two types of equivalence: strong and weak. Weak equivalence occurs when the input/output description of the formal model maps on to the dynamical mechanism, such that there are distinct states of the dynamical mechanism for distinct inputs to the formal model, distinct states of the dynamical mechanism for distinct outputs of the formal model, and a counterfactual mapping such that the sequence of dynamical mechanism states are arranged to preserve the input-output mapping that characterizes the formal model. Strong equivalence occurs when the input, output, and mathematical relations between the

\textsuperscript{26} Note that this is a prospective claim, as the causal role of the ACC in this process has not yet been demonstrated. Nonetheless, for the purposes of illustrating the features of the theory, the case study will serve.
inputs and outputs are mapped on to the dynamical mechanism.\textsuperscript{27} In cases of strong equivalence, not only are there distinct dynamical mechanism states for the inputs and outputs, but in addition the way that the dynamical mechanism transforms the input into the output can be described using the very same mathematical relations that occur in the formal model.

I will now argue that, assuming the dynamical properties of cognitive systems do execute formal models of processing, strong equivalence is too strong a requirement for executing a formal model. The mathematical functions that describe the mechanisms need not correspond to the functions that are present in the formal model. Integrative activity such as is seen in LIP during the RDMT can be described using different mathematical functions, such as integration, exponentiation (a simple first-order differential equation), or even using a prototypical ‘squashing’ function (Mazurek et al. 2003; Usher and McClelland 2001; Wang 2002, 2008; Churchland 2012). None of these functions are identical to the mathematical function of the sequential probability ratios calculated in the DDM. And yet, these mathematical descriptions of neural activity in LIP don’t prevent a mapping from the formal model described by the DDM to the

\textsuperscript{27} Insofar as we think of a mathematical relation as simply a set of ordered n-tuples F such that F \subseteq \mathbb{R}^n where \mathbb{R}^n is the n\textsuperscript{th} Cartesian product of the set of real numbers on itself, then there is no distinction between the two. However, this is only the case if our domain is the entire set of reals. While this is perhaps true for a formal model executed by the system (and even then, only for some formal models), this is simply not the case for mechanism states, unless there are uncountably many such states to map on to \mathbb{R}.
neural activity. There is still a weak equivalence between the two mathematical
descriptions sufficient for model execution. Thus, strong equivalence is not necessary for
a dynamical mechanism to execute a formal model of processing.

Though the dynamical mechanisms execute the formal model, there is no
necessary connection between any particular formal model and a dynamical mechanism.
A formal model will be executed by some set of dynamical mechanisms, but which
dynamical mechanisms execute that model may vary from system to system or within a
system over time. Likewise, as amply illustrated by the two case studies discussed
above, the same type of dynamical mechanism can execute different formal models. The
integrate-to-bound mechanism in LIP executes the DDM, whereas the integrate-to-
bound mechanism in ACC executes some aspect of the MVT. Thus, there is a
many:many mapping between formal models and dynamical mechanisms.

Though weak equivalence is all that is required for the execution of a cognitive
function by a set of mechanisms, there is a counterfactual constraint on cognitive
function execution as well. Not only must it be the case that the mechanism is weakly
equivalent to the cognitive function, but it must also be the case that if the inputs to the
system were different, such that the formal model dictates a different set of values for
the variables should obtain, then the mechanism would also be in a different set of states
corresponding to those different values. The mechanisms of cognitive systems will be
weakly equivalent to any number of different formal models; enforcing the counterfactual constraint restricts the range of alternate models to which the mechanisms are weakly equivalent.

2.1.7 Implementation

Before moving on to the discussion of the mechanistic status of the dynamical systems that execute the formal models of processing in cognitive systems, the implementation relation must be defined. I won’t herein argue at length for this definition, except to note that it may handle some well-known problems with other extant definitions. The definition I will propose ties the implementation of a dynamical system by a physical mechanism to a mapping between the physical mechanism and a formal model of processing, and to a process of selection for the dynamics of the physical mechanism.

In the context of discussion of the implementation of computational functions, implementation has been defined in the past as a type of mapping; in particular, a system implements a computational function if there is an isomorphism between the abstract formalization of the function and the causal structure of the system. However, this simple approach faces some now-classic problems. First, requiring a mere mapping results in something like pan-computationalism. Causal complexity is cheap in the world; the worry is that any rock or wall will execute any computation (Putnam 1988;
The problem of pan-computationalism stands as a challenge for any theory of implementation. Second, requiring a mere mapping results in something like a failure to individuate mental states. Granted a physical system that has a mind, and that having a mind is to implement some (presumably complicated) computational function, the complexity of such a system will result in its implementing any number of different minds (Shagrir 2001, 2012). The problem of multiple simultaneous implementations is equally a challenge for any theory of implementation.

Chalmers’ view is a good example of the mapping approach (Chalmers 1995, 1996). For Chalmers, a computational function is defined by a series of transitions from inputs and states to states and outputs; the formal structure of the function is mirrored by the causal structure of the system; and the causal structure of the system can be described by inputs to states, and states to outputs. Chalmers presents his definition of implementation in terms of combinatorial state automata (CSA). First, a finite state automaton (FSA) is:

“An FSA is specified by giving a set of input states I_1, …, I_k, a set of internal states S_1, …, S_m, and a set of output states O_1, …, O_n along with a set of state-transition relations of the form (S, I) → (S', O'), for each pair (S, I) of internal states and input states, where S' and O' are an internal state and an output state respectively” (Chalmers 1995, p. 392-3).

CSAs are like FSAs but

“differ from FSAs only in that an internal state is specified not by a monadic label S, but by a vector [S^1, S^2, S^3,…]…. There are a finite number of possible values S^i for each element S^i, where S^i is the jth possible value for the ith element…. Inputs and outputs can have a similar sort of complex structure…. State-transition rules are determined by specifying, for each element of the state-vector, a function by which its new state depends

77
on the old overall state-vector and input-vector, and the same for each element of the
output-vector” (Chalmers 1995, p. 394).
A CSA is a more complicated version of an FSA; the states, inputs, and outputs of the
automata can have complex internal structure, as reflected in the vectors for the states,
inputs, and outputs possibly possessing multiple elements. A CSA is an abstract
formulation of a computational structure, much like the Turing machine concept
presented earlier. In fact, as Chalmers notes (Chalmers 1995, p. 395), it is possible to
describe a Turing machine as a (finite or infinite) CSA: the total Turing machine state can
be described as a vector consisting of a numerical encoding of the state of the read/write
head together with the encoded state of the tape, each square on the tape described by
an ordered pair consisting of a number, indicating the square’s value, and a binary flag,
indicating whether or not the tape head is on that square. CSAs possesses an internal
state, described by a state vector, and the system transitions between states based on its
current state and the input-vector, resulting in a new state and output-vector. On the
basis of these definitions, Chalmers defines the implementation of a CSA as

“A physical system $P$ implements a CSA $M$ if there is a decomposition of internal states
of $P$ into components $[s^1, s^2, \ldots]$, and a mapping $f$ from the substates $s^i$ into corresponding
substates $S^i$ of $M$, along with similar decompositions and mappings for inputs and
outputs, such that for every state-transition rule $([I^1, \ldots, I^k], [S^1, S^2, \ldots]) \rightarrow ([S'^1, S'^2, \ldots], [O^1, \ldots, O^l])$ of $M$: if $P$ is in internal state $[s^1, s^2, \ldots]$ and receiving input $[I^1, \ldots, I^k]$ which
map to formal state and input $[S^1, S^2, \ldots]$ and $[P, \ldots, P]$ respectively, this reliably causes
it to enter an internal state and produce an output that map to $[S'^1, S'^2, \ldots]$ and $[O^1, \ldots,
O^l]$ respectively” (Chalmers 1995, p. 394).
This definition applies to any computational function that can be described as a CSA,
covering any Turing computable function, since any Turing machine can be described
by a CSA as just described. This definition of implementation also applies to the physical implementation of a computational function; however, there are no spatiotemporal limitations present in the definition, so the class of physical systems capable of implementing computational functions is correspondingly broad.

Intuitively, this definition is capable of handling the problem of widespread implementation, but falters on the problem of multiple simultaneous implementations, as Shagrir has persuasively argued (Shagrir 2012). Chalmers’ definition handles the problem of pan-computationalism, because every state-transition rule must be reflected in the casual structure of the physical system that implements the CSA. For very complex CSAs, presumably including those sufficient for possessing cognitive states, very few physical systems will have the physical structure required to implement the CSA. On the other hand, the problem of multiple simultaneous implementations appears to threaten. On Chalmers’ definition of implementation, implementation occurs wherever there is a mapping between the CSA and the causal structure of the system. Applied to the case of cognition, there will be some complex system that implements a CSA that corresponds to a cognitive function. But this complex system will then possess

---

28 Or so the reasoning runs; I’m unconvinced, as there is plainly a large amount of causal structure even in relatively simple objects, such as rocks.
a great amount of causal complexity, and could implement many other CSAs, especially those that correspond to relatively simpler cognitive functions.29

Ideally, a definition of implementation will handle both types of problem. I propose the following definition of implementation for this discussion:

A physical mechanism implements a dynamical system iff
i. the physical mechanism possesses as a subset of its properties, a set of dynamical properties composing a system that is token identical to the dynamical system;
ii. that subset of the physical mechanism’s dynamics are weakly equivalent to a (part of a) formal model of processing;
iii. the physical mechanism was selected for because of its possession of these dynamics;30 and
iv. the physical mechanism was selected for because its dynamics were weakly equivalent to a (part of a) formal model of processing.

Implementation involves token-identity of subsets of the properties of physical mechanisms and the properties of the dynamical system. Furthermore, some mapping is required; in this case, weak equivalence corresponds to input-output equivalence between the dynamical properties of a physical mechanism and the formal processing model describing the cognitive function. But there is also a selection component to implementation, thus preventing the problem of pan-computationalism (or, in this case, pan-cognitivism). Not every physical mechanism with the right dynamical properties

29 Shagrir’s argument is a good deal more complicated; the details aren’t required to have the requisite intuition, however.

30 Selection for is required, not just selection, to avoid piggy-back counterexamples where one mechanism gets selected in virtue of another mechanism being selected for.
will implement the corresponding dynamical system because that physical mechanism was not selected for in virtue of possessing those dynamics. This selective process can occur as a result of natural selection, but there are other selective processes that would qualify, such as learning or self-engineering selection, that would also qualify a physical mechanism for implementation of a corresponding dynamical system (see Garson 2011). In particular, the selection component reduces the severity of the problem of multiple simultaneous implementations of dynamical systems by physical mechanisms. There may be widespread weak equivalence in a causally complex system with many different formal models of processing. However, not all of the properties of such systems will have been selected for. Furthermore, though there may be many properties that are selected for, in order to count as an implementation, the properties must have been selected for in virtue of that weak equivalence. This last

---

31 Garson defines three selection processes besides natural selection that can endow traits with functions: neural selection, immune selection, and learning. What I’m calling self-engineering selection is a general term meant to cover effect such as synaptic selection, as just as there are diverse sorts of selection that might occur in the brain (covering synapse selection, neuronal group selection, and possibly others), so too might there be other physical systems that exhibit selective processes of diverse nature.

32 I would like to respond to one problem with this definition of implementation before moving, the problem of one-off implementations, which really results from a confusion. Organisms in the struggle to survival may face specific, novel processing problems, and those organisms do not suddenly freeze in inaction. And yet, my definition of implementation appears to prevent such organisms from implementing any processing model, as there has not been a process of selection. But the definition of implementation applies to dynamical mechanisms and their physical implementing substrates, not to the formal models of processing discussed earlier. Processing models, as defined above, require only a mapping for their execution. What they get mapped on to makes all the difference: on my view, cognitive systems result from the execution of processing models by dynamical mechanisms implemented, in the sense just defined, by physical mechanisms.
condition connects the physical mechanism, dynamical system, and formal model into a complex that has been selected for in virtue of the relationship between the dynamics and the formal model. While there may be many instances of weak equivalence between the dynamics of a physical mechanism and a formal model, there will be fewer such instances where the physical mechanism is selected for, and fewer yet where the physical mechanism is selected for because of that weak equivalence, thus reducing the severity of the problem of multiple simultaneous implementations.33

Though this approach to implementation still requires a fuller discussion and more robust defense, I hope I’ve illustrated some prima facie advantages the definition offers. Note that, as I argue next, the dynamical systems that are components of cognitive systems are dynamical mechanisms, and so this implementation relation will hold between mechanisms: physical mechanisms implement dynamical mechanisms. Similarly, the comments above will apply equally to the implementation of dynamical mechanisms as they do to the implementation of dynamical systems. Together with the

33 There may still be cases of multiple simultaneous implementation if the physical mechanism’s dynamics are weakly equivalent to more than one formal model that is plausibly what was selected for in the organism or in the environment. Such cases may occur where there are competing formal models offered to describe the behavior of an organism, and either those competing formal models contain some of the same mathematical functions or the outputs of the formal model are identical over some range. The former describes cases where the dynamical mechanism executes only part of a formal model and two (or more) formal models are strongly equivalent in that part, and the latter describes cases where the dynamical mechanism is weakly equivalent to two (or more) formal models in whole or part. In cases where there is only one plausible formal model, then there will no longer be the possibility of multiple simultaneous implementations.
definition of model execution as a form of weak equivalence, I hope I have now elucidated the connection between the formal model, the dynamical mechanism, and the physical mechanism that is central to my dynamicist theory of cognition.

2.2 Dynamical Mechanisms

I will now argue that these dynamical systems that play a crucial role in cognitive phenomena are dynamical mechanisms. What about the integrate-to-bound system, or other dynamical systems in cognitive systems, makes it a mechanism? Dynamical systems are not identical to their physical implementations, and aren’t characterized by their concrete organizations. However, in the recent literature on mechanisms, mechanisms are taken to be concrete physical organizations whose parts causally interact to produce some phenomenon (Machamer et al. 2000; Craver 2007). So what about dynamical systems makes them mechanisms? This is a general question that has implications outside of the mechanisms of cognition; for example, abstractions like Turing machines (Otto and Rusanen 2011) or network motifs in structural biology (Alon 2007; Levy and Bechtel 2013) are also plausibly mechanisms, though lacking the typical physical properties invoked in definitions of mechanisms (Dupré 2013; Woodward 2013; Levy 2014; Levy and Bechtel 2013).

---

34 My argument is limited to the dynamical systems that play a role in cognitive phenomena. Persuasive parallel arguments can be offered for seeing dynamical systems in computer science, biology, geology and other sciences as dynamical mechanisms. From now on, unless explicitly noted, when I speak of dynamical systems, I am picking out the dynamical systems that are components of cognitive systems.
First, I examine the recent discussion of mechanisms in the philosophy of science, constructing a general definition that reflects this discussion. These recent definitions are unduly restrictive in what counts as a mechanism, preventing dynamical systems from being classified as mechanisms. My argument that dynamical systems are indeed mechanisms consists in two parts. First, I will argue that mechanisms are emergents of a certain sort and not mere aggregates, by appealing to Wimsatt’s contrast of emergents with mere aggregates (cf. Craver 2001, 2007). Second, I argue that the dynamical systems appealed to in the case studies above satisfy the requirements for being a mechanism in virtue of certain aspects of their organization. I present a revised definition that will accommodate dynamical systems, abstract machines like Turing machines, and possible other abstract organizations like network motifs in structural biology as mechanisms.

Before continuing I would like to forestall the complaint that all of this is so much furious discussion over what is ultimately an issue of terms. I am sympathetic to this terminological objection, welcoming a recasting of my theory in non-mechanistic terms for the dynamical systems at the heart of cognition. The concepts denoted by our terms and the entities in the world to which our concepts refer are the focus of the discussion, not whether we call these systems ‘mechanisms’ or ‘machines’ or ‘abstract causal organizations’ or what-have-you (cf. Woodward 2013, p. 41). In the case of

---

35 These recent discussions are motivated by considerations arising from scientific explanation.
mechanisms, consideration of how mechanisms are differentiated from other sorts of systems results in a more inclusive definition, on which the dynamical systems present in cognitive systems are to count as mechanisms. There are interesting differences between dynamical mechanisms and physical mechanisms. Approaches that only stress physical mechanisms will miss these differences. Insofar as dynamical mechanisms lie at the heart of cognitive phenomena, these approaches will also miss the heart of cognitive systems—regardless of what terms we use to label them.

2.2.1 Recent Approaches to Mechanisms

The recent turn to mechanism gathers together a group of ideas about entities and activities acting to produce some output. In an influential 2000 paper, Machamer, Darden and Craver define mechanisms as “entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions,” where activities are “producers of change” and entities are “the things that engage in activities” (Machamer et al. 2000, p. 3). Bechtel and Abrahamsen define a mechanism as “a structure performing a function in virtue of its component parts, component operations, and their organization” (Bechtel and Abrahamsen 2005, p. 423), while Wright and Bechtel, in their review on mechanisms and explanations of cognition, note “…mechanisms are composed of component parts and their properties. Each component part performs some operation and interacts with other parts of the
mechanism..., such that coordinated operations of parts is what constitutes or comprises the systemic activity of the mechanism” (Wright and Bechtel 2006, p. 45). Finally, Piccinini defines a “mechanism M with capacities C [as] a set of spatiotemporal components A1, ..., An, their functions F, and F’s relevant causal and spatiotemporal relations R, such that M possesses C because (i) M contains A1, ..., An, (ii) A1, ..., An have functions F organized in way R, and (iii) F, when organized in way R, constitute C” (Piccinini 2010, p. 285). All of these accounts bring together a group of objects or entities, imputes some organization to them, and attributes some activity to the organized collection of entities, which produces the explananda phenomena.

What are the features of the objects or entities that appear in mechanisms?

According to Machamer et al., the entities “often must be appropriately located,

36 Piccinini defines a mechanistic explanation of a system X as “a description of X in terms of spatiotemporal components of X, their functions, and their organization, to the effect that X possesses its capacities because of how X’s component and their functions are organized” (Piccinini 2007, p. 506). Elsewhere, Piccinini applies the mechanistic approach to explanations of computers and computer science, noting that labeling the brain a computational mechanism is an empirical hypothesis needing investigation to confirm. Piccinini defines mechanistic functionalism about a system S as “the thesis that S is the functional organization of the mechanism that exhibits S’s capacities”, (Piccinini 2010, p. 288-289) where “… a mechanism’s functional organization includes the states and activities of components, the spatial relations between components, the temporal relations between the components’ activities, and the specific ways the components’ activities affect one another” (Piccinini 2010, p. 286). In discussing the computational theory of mind, Piccinini argues that constructing a mechanistic explanation requires characterizing the system’s functional organization, which may or may not be the functional organization of digital computers. Piccinini talks simply of ‘computers’ and ‘computation’, but in more recent work has differentiated between digital and other sorts of computation (e.g. Piccinini and Bahar 2012). Piccinini’s definition of a mechanism and mechanistic functionalism about a system clearly preclude a purely dynamical approach. At a minimum, the dynamical mechanisms that constitute a cognitive system do not include the spatiotemporal characteristics required by Piccinini. Note, though, that the definition of mechanistic functionalism that Piccinini provides is fairly general, even though he requires more stringent details in specifying the functional organization of a mechanism.
structured, and oriented...” (Machamer et al. 2000, p. 3) and “[t]raditionally one identifies and individuates entities in terms of their properties and spatiotemporal location” (Machamer et al. 2000, p. 5). Bechtel and Wright aver that “mechanisms are composed of component parts and their properties...” though they acknowledge that “not just any sequence of causal interactions will suffice” to constitute a mechanism (Bechtel and Wright 2006, p. 45). Bechtel and Abrahamsen hold that “[t]he component parts of the mechanism are those that figure in producing a phenomenon of interest” and “may involve multiple levels of organization” (Bechtel and Abrahamsen 2005, p. 424). The parts of a mechanism are individuated on the basis of their properties (spatiotemporal, structural or otherwise) and their role in producing the explananda phenomena.

The activities of a mechanism, resulting from the parts and their various interactions, produce or bring about the mechanism’s product. Machamer et al. characterize these activities as “producers of change,” possessing a “temporal order, rate, and duration” and “constitutive of the transformations that yield new states of affairs or new products” (Machamer et al. 2000, p. 3-4). Fundamentally, mechanisms “do things. They are active and so ought to be described in terms of the activities of their entities, not merely in terms of changes in their properties” (Machamer et al. 2000, p. 4-5). Like entities, “[a]ctivities... may be identified and individuated by their
spatiotemporal location…” as well as “by their rate, duration, types of entities and types properties that engage in them” (Machamer et al. 2000, p. 5). For Bechtel and Abrahamsen, “[e]ach component operation involves at least one component part” of the mechanism, and the mechanism’s “[o]perations can be organized simply by temporal sequence, but biological mechanisms tend to exhibit more complex forms of organization” (Bechtel and Abrahamsen 2005, p. 424). For Wright and Bechtel, “[t]he relevant organizational and architectural properties…enable the parts to work together effectively and perform the phenomenon Φ targeted for explanation…” (Wright and Bechtel 2006 p. 46). Activities are classified and characterized by the types of entities that engage in them, their rate or duration, or their spatiotemporal properties. The activities result from the concordant action of the components of the mechanism operating together to produce some result.37

Two brief examples will suffice to illustrate what I take to be the received definition of a mechanism. Consider the heart, an organ that circulates blood throughout the body.38 The heart is comprised of an aorta, valves, ventricles, cellular tissue and

37 Bechtel et al. talk of ‘operations’ whereas Machamer et al. talk of ‘activities’. Use of ‘activities’ perhaps suggests an active process, which may be an intentional invocation of actual activity as opposed to a mere disposition (cf. Craver 2001 p. 58 on Cummins, capacities and activities). If we suppose that operations are more like dispositions, then activities may be the result of the execution of an operational disposition. At any rate, I will not examine if there are differences between activities and operations here.

38 Bechtel and Abrahamsen 2005, p. 424ff discuss the heart as mechanism.
other structures, all of which act in concert to pump blood received from the body via the vena cava through the pulmonary artery to the lungs and from the lungs via the pulmonary veins through the aorta to the body. This mechanism is composed of a set of parts, such as the valves or the ventricles, with associated operations or activities, such as contracting rhythmically or preventing backflow. When organized in the appropriate fashion, these parts and activities comprise the mechanism, which produces the explanandum phenomenon, in this case pumping blood.

As another example, consider the biophysical mechanism of a chemical synapse, such as a voltage-sensitive Na⁺ channel. The mechanism consists of a set of entities, including the cell membrane, vesicles, microtubules, ions, etc. and a set of operations, including biosynthesis, coupling, diffusion, transport, depolarization, etc. The process of Na⁺ channel opening proceeds as follows. The intracellular matrix rests at -70 mV relative to the extracellular fluid. Spreading depolarization from elsewhere along the cell causes alpha helix conformational changes in the Na⁺ channel, a result of the positive charges in the corkscrew proteins comprising the channel being repulsed, opening the voltage gates and permitting an influx of sodium ions. The influx of positive charge peaks at approximately 50 mV relative to the extracellular fluid. This depolarizing

____________________

30 Machamer et al. 2000, p. 8ff and especially Craver 2007, p. 114ff discuss the chemical synapse example.
process is explained by the set of entities acting together in an organized fashion, causing the alpha helix conformational changes that result in depolarization.

On the basis of the foregoing discussion of entities, activities and these two examples, the following definition of a mechanism can be constructed:

(M) A mechanism M is a set E of entities that (i) is characterized by their physical properties (such as spatiotemporal location, shape, electrochemistry, etc.), which (ii) are arranged in an organization O such that (iii) when O-organized, E exhibits a set A of activities that (iv) are characterized by their physical properties (such as spatiotemporal location, rate, duration, etc.) or their constituents such as the subset e of E that initiate, participate in, or result from them or the changes in the properties of some subset e of E they produce, and that (v) causally produce some phenomena.40

A mechanism’s parts can exist at multiple sizes, such as in the heart, where the parts range from pacemaker cells to ventricles, and the organization of the parts exhibit the activities that allow for the production of the explanandum. The parts of a mechanism are individuated on the basis of their properties (spatiotemporal, electrical, etc.) and their role in producing the phenomena by being oriented, structured, etc. in the right fashion (Machamer et al 2000). In the case of the heart, these parts include the cells, the aorta, the valves, and so forth, while in the chemical synapse, these include the cells, the ion channels and proteins forming them, the ions, and so forth. The parts and their various interactions result in the mechanism’s activities, which produce or bring about

---

40 Cf. Machamer et al. 2000, p. 3ff; Bechtel and Abrahamsen 2005, p. 423ff; Wright and Bechtel 2006, p. 45ff; Piccinini 2010, p. 285ff; Illari and Williamson 2013, p.119ff. Recently, Kaplan and Craver have offered a revision to this approach that moves the definition much closer to mine; see the discussion below in section 2.3.
the phenomena. These activities are defined by their temporal order, rate, and duration, and individuated by the entities that engage in them, their spatiotemporal properties, or their organizational properties (Machamer et al. 2000; Wright and Bechtel, 2006). The actions of the parts of a mechanism, typified by the types of entities that engage in them and their properties, constitute the mechanism’s activities.

### 2.2.2 Dynamical Systems as Mechanisms?

The case studies from cognitive neuroscience detailed above challenge this account of mechanisms, specifically clauses (i), (iv), and (v) in (M) above. Recall that the dynamical properties of neural mechanisms are dynamical systems, that those physical mechanisms implement the dynamical systems, and the dynamical systems execute the formal models that mathematically describe the processing in the system. I will now make the case for considering these dynamical systems as dynamical mechanisms.

I argue that there are two kinds of mechanism in cognitive neuroscience, dynamical and physical, with physical mechanisms implementing their dynamical counterparts. By contrasting dynamical mechanisms with physical mechanisms, I do

---

41 Temporal sequence in biological mechanism may be less important, as they often exhibit more complex organization (Bechtel and Abrahamsen 2005; Bechtel 2012, 2013).

42 Distinguishing senses of mechanism distinct from (M) is common (see, e.g., Anderson 2014a, b). Many other philosophers have drawn a similar philosophical distinction, including Matthewson and Calcott (2011), Kuorikoski (2009), Levy and Bechtel (2013), and approaching the issue from a dynamical systems perspective, van Gelder and Port (1995), Giunti (1995) and van Gelder (1996). The idea of repeated patterns
not mean to imply that the dynamical mechanisms are metaphysically non-physical. Recall that at the outset, I defined dynamical mechanisms as composed of the dynamical properties of physical mechanisms. This firmly places them on physical grounds. Physical mechanisms correspond to the mechanisms that implement the dynamical mechanisms, and have the properties adverted to in (M). The individual neurons in LIP that exhibit a ramp-up in their firing rates over the course of individual trials are good examples of physical mechanisms: they have a set of entities, in particular the neuron with its cellular parts and the ions that constitute the charge carriers, organized in a particular fashion with ion pumps embedded in cell walls that maintain the electrolytic balance, and these parts exhibit a set of activities, such as the ion pumps opening and closing in response to changes in membrane potential, ions flowing into and out of the cell, intracellular changes resulting in generation and propagation of $\text{Ca}^{2+}$ flows inside the cell and the consequent maintenance and transmission of an action potential along the length of the axon. The individual neuron constitutes a physical mechanism if anything does.

_____________________

of organization has a long history in biology and philosophy, though it has recently been codified in a series of articles on network motifs; see Alon (2003, 2007a and especially 2007b, a textbook treatment), Milo et al. (2002), Milo et al. (2004), and in application to neural data specifically, see Qian et al. (2011), Rubinov and Sporns (2010), and Whalen et al. (2012).
Dynamical mechanisms, on the other hand, possess different entities and different properties from the physical particulars of their implementation. Consider the integrate-to-bound system again. It has components as well—a threshold, an activation point (i.e., where in the integration period the mechanism lies), a baseline, perhaps we might include the state space of the system as one of its ‘components’—and these components appear to be organized, such as the threshold being at a higher level of activity than the baseline, the activation point lying between the threshold and baseline, etc. Furthermore, these parts seem to exhibit activities, such as thresholding, resetting, and the crucial integrating. All of these components and activities are implemented by a physical mechanism that stands in an implementation relation with the dynamical system. These dynamical systems violate conditions (i), (iv), and (v) in (M). But are they still mechanisms, in some more general sense?

### 2.2.3 Wimsatt on Emergents and Aggregates

Dynamical systems seem to satisfy some of the constraints laid out in (M), such as the possession of component entities that engage in certain activities. Furthermore, the component entities are organized in a certain fashion such that their organization and interaction results in the activities characteristic of the dynamical system. Dynamical systems, then, don’t appear to be any old collection of entities with arbitrary relations and properties. How should we characterize these systems, and if they are mechanisms,
what separates them from non-mechanisms? I will argue that mechanisms are a subset of what Wimsatt calls emergents, where emergents are contrasted with mere aggregates.

Wimsatt has discussed the difference between emergents and mere aggregates (Wimsatt 1986, 1997). Wimsatt defines aggregativity as “the non-emergence of a system property relative to properties of its parts” (Wimsatt 1997, p. S375). Wimsatt provides four criteria for aggregativity, which he believes are jointly sufficient, though not necessary. Determining aggregativity first requires decomposition, breaking a system down into its components and their properties. The decomposition step provides a composition function $P(S_i) = F([p_1, p_2, \ldots, p_n(s_i)], [p_1, p_2, \ldots, p_n(s_2)], \ldots, [p_1, p_2, \ldots, p_n(s_m)])$ “for system property $P(S_i)$ in terms of parts’ properties $p_1, p_2, \ldots, p_n$, of parts $s_1, s_2, \ldots, s_m$.

The composition function is an equation—an inter-level synthetic identity, with the lower level specification a realization or instantiation of the system property” (Wimsatt 1997, p. S376). Then, if some system property satisfies the following four conditions, it is an aggregate with respect to such decomposition:

“(1) IS (InterSubstitution) Invariance of the system property under operations rearranging the parts in the system or interchanging any number of parts with a corresponding number of parts from a relevant equivalence class of parts…. (2) QS (Size Scaling) Qualitative Similarity of the system property… under addition and subtraction of parts…. (3) RA (Decomposition and ReAggregation) Invariance of the system property under operations involving decomposition and reaggregation of parts… (4) CI (Linearity) There are no Cooperative or Inhibitory interactions among the parts of the system for this property” (Wimsatt 1997, p. S376).
On Wimsatt’s view, satisfying these conditions is jointly sufficient for aggregativity, so violating them does not mean the system is not an aggregate. Suppose, as an illustration, that we have a pile of sand with some weight. The property of having weight is the relevant system property (the weight property). The weight property satisfies IS: rearranging the grains, such as by shifting the spatial relations between them, does not change the pile’s property of weighing something or other, and neither does interchanging some number of grains with other grains. The pile property satisfies QS: adding and subtracting grains won’t change the pile’s property of having weight, though it will quantitatively change that weight. The weight property satisfies RA: taking apart the pile, such as by removing some number of subsets of the whole pile, and then reaggregating them by putting them back into the pile does not change the weight property. Finally, the pile property satisfies CI: the grains of sand in the pile do not have cooperative or inhibitory interactions.

Dynamical mechanisms are clearly not aggregates. In order to demonstrate this, some method of decomposing these systems into their parts and their parts’ properties is necessary. Consider our running example of the integrate-to-bound mechanism. Translating into Wimsatt’s terminology, consider the system property of being an integrate-to-bound mechanism. The property can be decomposed into its components along the lines suggested above. There is a baseline, a point in the mechanism’s state-
space, a threshold, among other possible parts. These parts have properties; for example, the threshold will have a particular level, may be more or less flexible, and so forth. Similar points hold for the other parts. This decomposition of the system into parts and parts’ properties is the composition function for the system. Now, take each Wimsatt criterion in turn. The integrate-to-bound property fails to satisfy IS: rearranging the parts, such as swapping the baseline and threshold, radically alters the nature of the mechanism. Swapping out the components can also alter the mechanism, for example, swapping out a threshold for a different threshold can radically shift the state-space of the mechanism. The integrate-to-bound property fails to satisfy QS: adding thresholds or points of activation in the state-space of the mechanism drastically alters the mechanism. The integrate-to-bound property fails RA: moving the threshold around, or the activation point, will change the system property. Finally, the integrate-to-bound mechanism fails to satisfy CI: the parts cooperate in order to execute the system property. Failures to satisfy CI are especially important for other dynamical mechanisms, ones that involve inhibitory actions between components, such as in the synaptic reverberation mechanism discussed later.

The four Wimsatt criteria are sufficient to establish that a system property is an aggregative property. Violation of these criteria, however, do not entail that the property is not aggregative, though it may suggest it. Consider, for example, a lump of graphite.
A lump of graphite satisfies QS, as adding or subtracting carbon atoms in the lump does not change its lumpiness. However, not every rearrangement, decomposition and reaggregation will result in a lump of graphite, violating IS and RA, as graphite has a characteristic chemical structure, a honeycomb lattice of carbon atoms. Finally, the graphite lump violates CI as well, since the cooperative interactions of the atoms result in graphite’s organization. Specifically, there are two such organizations, hexagonal and rhombohedral, resulting from different ways of stacking graphene layers to form the lump. Is a lump of graphite not an aggregate? This is somewhat a terminological question, but the important point is that there is a range of organization and different systems will fall at different points on this range, resulting in greater or less degrees of emergence. Wimsatt contrasts aggregative properties with emergent ones, but there are many types of emergent properties. If we fail to distinguish different types of emergents, for example, classifying all emergents as mechanisms, then we will fail to capture differences in the class of emergents, in particular failing to make potentially empirically relevant distinctions, and trivializing the designation of a system as a mechanism. In such a case, intuitively distinct examples of emergents, such as organisms, statues, diffusing gases, lumps of graphite and dynamical mechanisms, all of which are emergents in virtue of violating some of the Wimsatt criteria, would be co-classified, while glossing over the differences between these emergents.
2.2.4 Mechanisms and Emergents: Organization

I just argued that a definition of mechanism that fails to distinguish mechanisms from other non-aggregates trivializes the concept of a mechanism and glosses over empirical and conceptual distinctions between types of emergents. Avoiding the charge of triviality requires a more refined definition of a mechanism. Identifying features of some system that violate the Wimsatt criteria are insufficient to pick out a mechanism, as I demonstrated above. What more is required to identify a mechanism? Both IS and RA conceptually involve the notion of organization. Recall (M) above: that approach appeals to the organization of a system as a criterion for mechanism-hood. Perhaps, then, a closer examination of how systems are organized can provide further differentiating criteria.

Craver argues that distinctively mechanistic forms of organization can be brought out in a contrast with Wimsatt’s aggregates. He lists modified Wimsatt conditions:

“Suppose that a property or activity (ψ) of the whole (S) is explained… by the properties or activities {ϕ₁, ϕ₂, …, ϕₙ} or its parts {X₁, X₂, …, Xₘ}. The ψ-property of S is an aggregate of the ϕ-properties of X’s when: (W1) ψ is invariant under rearrangement and intersubstitution of X’s; (W2) ψ remains qualitatively similar (if quantitative, differing only in value) with the addition or subtraction of Xs; (W3) ψ remains invariant under the disaggregation and reaggregation of Xs; and
(W4) There are no cooperative or inhibitory interactions between Xs that are relevant to \( \psi \) ” (Craver 2007, p. 135).43 Craver utilizes Wimsatt’s criteria for aggregativity to help distinguish mechanisms from mere aggregates. Stating that the criteria are “for diagnosing the importance of organization in a system and are also a set of strategies for discovering a mechanism’s organization”, a system’s organization serves as one of the central criteria (Craver 2007, p. 135). He argues that

“components of mechanisms, in contrast to those of mere aggregates, have an active organization; they act and interact with one another in such a way that the \( \psi \)-ing of S is more than just the sum of \( \varphi \) properties. In fact, the \( \varphi \) properties of mechanisms... are the activities of and among the entities of the mechanism.... [M]echanisms are composed of different kinds of entities... engaging in different kinds of activities... and acting in cooperation or competition with specific other entities in the mechanism.... It matters which Xs \( \varphi \) with which others, and it matters where, when, how much, and how often” (Craver 2001, p. 59-60, italics in original).

There are three varieties of such organization: active, spatial, and temporal, with different degrees of prevalence of each type in different mechanisms. Active organization is necessary, on Craver’s view, for mechanisms. Mechanisms, such as the action potential mechanism, are actively organized, and their components “act and interact with one another in such a way that the \( \psi \)-ing of S is more than just a sum of \( \varphi \)-properties.... [T]he \( \varphi \)-properties of a working mechanism... are the activities of and interactions among the entities in [it].... It matters which Xs \( \varphi \) with which others, and it

43 The main changes appear to be a recasting of the criteria as for explanation and simplification of the language. I am ignoring both differences.
matters how they interact” (Craver 2007, p. 136). This active organization violates W1-W3. Craver goes on to stress the spatiotemporal organization of such mechanisms, noting that active organization “is sustained by the spatial and temporal organization of the component parts” (Craver 2007, p. 137, italics removed). Spatial organization includes “the sizes, shapes, structures, locations, orientations, directions, connections, and compartments of” the mechanism’s components (Craver 2007, p. 137). Temporal organization includes the “order, rate, and duration of successive component activities” and the “sequence of stages from beginning to end, and it is not possible to change their order without interfering with how the mechanism works…” (Craver 2007, p. 138). In sum:

“Mechanistic explanations… are anchored in components, and those components occupy space and take time to act. A description of a mechanism… is a description of how they work together. That description involves—in addition to a list of component entities \{X_1, X_2, ..., X_m\} and activities \{ϕ_1, ϕ_2, ..., ϕ_n\}—an account of how they are organized together actively, spatially, and temporally in S’s ψ-ing” (Craver 2007, p. 138).

In sum, on Craver’s account, in addition to violating the Wimsatt criteria, specifying the active, spatial and temporal organization helps differentiate mechanisms from non-mechanisms. It is unclear on the basis of this whether on Craver’s theory there is something other than spatiotemporal organization that contributes to a mechanism. On my theory there is: the dynamics are actively organized but not necessarily in a physical
spatiotemporal fashion, though certain temporal relations, such as the order of activities, is present.

On Craver’s view, specifying the active organization can occur via a specification of the spatial and temporal features of the system components that result in the active organization. I will argue that specifying the spatial and temporal properties of the system components is neither necessary nor sufficient for specifying the active organization of a system. First, the specification of the spatial and temporal properties of the components of the system (the entities and activities), i.e. the specification of the spatiotemporal organization of the system, need not result in a specification of the active organization. In short, specification of the spatiotemporal organization is not sufficient for specification of the active organization. Second, specification of the active organization can occur without specification of the spatiotemporal organization. As illustration, there are examples drawn from computer science where specifying the active organization occurs without specification of the spatial and temporal properties of

44 I take a ‘specification’ to be a description of the target of the specification. A note on description before continuing. Descriptions for this discussion are sets of sentences, and in the case of descriptions about something, they are sets of sentences that denote the (parts of the) target of the description. This denotation may fail, that is, the system being described may not exist. Nonetheless, often we think of nonexistent systems as mechanisms. By convention, the sentences that denote nonexistent system would be trivially true. Insofar as the reference to a system succeeds and the description of the system is accurate, the description will be true of the system, and the system will have the properties denoted by the description’s predicates.
the components. In short, specification of the spatiotemporal organization is not necessary for specification of the active organization.

Specification of spatiotemporal organization is insufficient for specification of the active organization of a system. Since the active organization is sustained by the spatial and temporal properties of the components, specifying the spatial and temporal properties of the components results in the specification of the basis of the active organization. However, this specification need not result in the specification of the active organization, especially if the spatial and temporal properties are very complex, the relations between the components nonlinear, such as cyclical behavior in many biological mechanisms, and the behavior of the system dictated by extra-componential constraints (Bechtel 2011, 2012). For example, consider the following model of oscillatory activity generating circadian rhythms in mammals (figure 6; Leloup et al. 2002, p. 7052):
Figure 6: Example genetic network for regulating circadian rhythm. Adapted from Leloup et al. 2002, p. 7052.

The model specifies the components of the mechanism, the transcription factors and genes and their interactions, generating the oscillatory activity regulating circadian rhythms. Importantly, this schematic describes a mechanism that results in specific periodicities and quantities of mRNA (figure 7; Leloup et al. 2002, p. 7053):
Figure 7: Concentrations of genetic products and periodicities in the generation of circadian rhythms. Adapted from Leloup et al. 2002, p. 7053.

Here, the amount of genetic product oscillates over time in a fashion that sustains the circadian rhythm. As Craver claims, specifying the spatial and temporal properties of the components sustains the active organization of the mechanism, in the sense that the active organization causally results from those spatial and temporal properties. This organization need not be linear or simple, though, and as a result the active organization is not specified. The spatial and temporal properties of the components can result in a complex set of interactions, whose resultant organization cannot be simply ‘read off’ from the list of spatial and temporal features of the system. The specification of the
active organization requires an analysis of the spatial and temporal properties of the components, often with the aid of computational and other types of models (Bechtel and Abrahamsen 2010). Thus, the specification of the spatial and temporal properties of the component entities and activities sustains but may not eo ipso specify the active organization. That is, the specification of the spatiotemporal organization of a mechanism sustains but may not eo ipso specify the active organization.

A simple objection applies here: while we may not understand or know what active organization results from the spatiotemporal organization, the spatiotemporal organization does in fact specify the active organization. The description of the spatiotemporal organization may be very complex, and so the active organization thereby specified highly non-obvious, but as a matter of fact the spatiotemporal organizational specification specifies the active organization.

Simply put, this is false. The description of the spatiotemporal properties of the mechanism components is insufficient for describing what the mechanism does. In addition to those properties, a description of the way that they are related to each other and the way these relations are constrained and limited is required. This functional organization captures the interrelations between the parts and is described by the equations governing the concentrations of the various mRNA and other genetic products (Leloup et al. 2003, supplementary p. 3ff). For example, the equation for the
concentration of *per* mRNA, $M_v$, the product of *per* (short for *period*) and that regulates production of the protein PERIOD, which exhibits circadian oscillations in concentration, is

$$\frac{dM_P}{dt} = v_s P \frac{B_n^P}{K_{Ap} + B_n^P} - v_m P \frac{M_P}{K_{mP} + M_P} - k_{dmp} M_P.$$  

This equation describes how the concentration of *per* changes as a result of various parameters, such as the maximum rate of accumulation of *per* ($v_s$), threshold constants ($K_{Ap}$), and other parameters and variables. These equations, which describe the active organization of the mechanism, are not merely the order, rate and duration of successive activities, such as the rate of *per* production. They also specify how the various other components of the mechanism, as well as parameters that capture intrinsic limitations and other aspects of the organization, affect the rate. These rate-limiting parameters, for example, are not entailed by the spatiotemporal description of the mechanism.

On the one hand, I don’t think Craver would be opposed to expanding what’s required to account for the active organization of the mechanism. After all, rate-limiting parameters and similar influences on the active organization of the mechanism are ultimately dictated by the physical properties of the entities and activities that compose the mechanism. To obtain these parameter values and other influences, a complete spatiotemporal description, down to the subatomic components if necessary, would result in a specification of these values. However, this requirement for full specification...
is unduly burdensome. Every mechanism, including physical ones like hearts and
synapses, are constrained by various physical parameters and aspects of their physical
organization that come from the fundamental material constituents of the mechanism.
And yet, we clearly can describe these mechanisms without providing all of that detail.
Furthermore, including those details would weaken explanations that rely on those
descriptions of mechanisms, a point appreciated by Craver himself (Craver 2007). I
conclude that the specification of the spatial and temporal organization of mechanisms
is not sufficient for specifying the active organization.

On the other hand, these constraints on active organization appear to require the
spatial and temporal organization. While specifying the spatiotemporal properties of the
components may be insufficient—because of extra-componential constraints on the
system’s behavior—the active organization still appears to be necessarily constrained by
the specification of the spatiotemporal organization.

However, specification of components without providing the spatial and
temporal details can sustain the active organization of a mechanism; that is, specification
of the spatiotemporal organization is not necessary for specification of the active
organization. For example, consider a basic finite Turing machine that executes the
addition function (the ADD function). A Turing machine consists of a set of basic
components, including a read/write head, an (infinite) tape, a mechanism for moving the
tape to the left or right, and a look-up table for all the transitions, as well as potentially a ‘state register’ for encoding the state of the machine with accompanying read/write machinery for the state register. The basic functions here include reading a square on the tape, erasing what is in the square, writing to the square, moving the tape left once, moving the tape right once, and potential similar operations for the state register, a mechanism that encodes the state of the system (and can be as simple as an individual square of tape).

The ADD function takes as input two numbers and provides as output their sum. The true ADD function can take as input any two numbers and output their sum; this would require something along the lines of uncountably many entries in the machine table, due to the ability to add real numbers. Even if we restrict ourselves to the natural numbers, we will require an infinite tape. Is there a more parsimonious way to implement the ADD function, at least for natural numbers? Yes, and here’s how: suppose we structure our input in certain ways so our machine knows when we want it to add. We can do this, for example, by starting our input to the machine, our tape, with two ‘0’ s. Then, we input a series of ‘1’s corresponding to the cardinality of our first summand. We end this string with a ‘0’, and then enter a second series of ‘1’s corresponding to our second summand. To let the machine know we’ve finished, we can enter another ‘0’.
From the machine’s perspective, we can structure our machine table to reflect this simple input code. The machine starts by reading the first square. If there is a ‘0’, then the machine moves one square to the left. If there is a second ‘0’, then the machine moves another square to the left. If there is a ‘1’, then the machine proceeds to move another square left, until a ‘0’ is found. Then, it erases the ‘0’, writes a ‘1’, and moves one square to the left. The machine continues moving left as it reads ‘1’s until it reads a ‘0’. Then, it moves one square to the right, prints a ‘0’, and outputs the whole string, say by moving all the way back to the first zero. The machine adds numbers simply by concatenating the two strings of ‘1’s into one long string of ‘1’s.

The active organization of this little adding machine is specified in the rules for transforming the states of the machine, such as writing ‘1’s and ‘0’s or moving the read/write head. Nonetheless, the spatiotemporal organization of the machine is left unspecified. Prima facie, describing the spatiotemporal organization is unnecessary for specifying the active organization of a system like the finite adding Turing machine described above.

In reply, Craver and allies could contend that what is specified is the abstract functional organization of the system, not its active organization. Active organization of a mechanism refers to how the mechanism components act and interact (interaction for short). But Craver and allies do not specify precisely in what this interaction consists.
The simple adding machine described would seem to have interactions, as different inputs result in different outputs. Despite not having a spatiotemporal specification, describing these interactions appears to describe the active organization of the system. Absent an argument that such descriptions of interactions don’t describe the active organization, we should admit that a specifying the functional organization is one way of specifying the active organization.

I conclude that specification of the spatial and temporal organization of a mechanism is neither necessary nor sufficient for specifying the active organization. In the case of Turing machines, the specification of components and activities results in an active organization, though the spatial and temporal organization has not been specified. A similar situation holds for the dynamical systems of cognition. To distinguish this variety of active organization from the sort that attends the specification of the spatial and temporal organization Craver requires, I will call this functional organization.

Violation of Wimsatt’s criteria does not license the inference that Turing machines or other abstract systems are mechanisms. As noted above, a lump of graphite violates the criteria; so do organisms, statues and diffusing gases. Satisfying additional criteria is necessary to further distinguish mechanisms from other non-aggregates. Many (all?) non-aggregates possess some organization as well. One criterion, then, separating
emergents from mere aggregates is the specification of the organization of the system. I will now argue that mechanisms do seem to be differentiated from other emergents by their active organization, be it functionally, spatially, or temporally specified.

Does active organization obtain for all non-aggregates? Consider the lump of graphite. The lump of graphite is composed of carbon atoms arranged in a certain fashion. The atoms interact in a certain fashion, forming covalent bonds that result in graphite’s structure. Furthermore, the different layers stack differently for different types of graphite. These are all interactions between the components that make up graphite, namely the carbon atoms or sheets thereof. A lump of graphite does seem to be actively organized.

However, not all non-aggregates possess active organization. Some non-aggregates possess an organization, but one that is not an active organization. Active organization, as Craver argues, requires that there be differentiation in the interactive properties of the components; uniformity in the interaction of the components is insufficient for active organization, though perhaps sufficient for organization simpliciter. This uniformity can be due to the substitutability of the components—the identity of what is interacting with what makes no difference to the product of the mechanism. It can also be due to the substitutability of the nature of the interactions—the particular sort of interaction makes no difference to the produce of the mechanism.
Consider a diffusing gas made up of a set of uniform molecules. The gas may have an organization, a description of the spatial and temporal properties of the molecules, such as their relative locations, their kinetic energy, and so forth. These molecules may have different speeds or relative locations, thus interacting with each other in different ways, yet still be classified as a diffusing gas, that is, as a set of molecules that exhibit decreasing density over time. Similarly, swapping out one molecule in an interaction for another won’t affect the property of diffusing. I conclude that active organization does differentiate some non-aggregates from others.

In sum, I have argued that organization is important for non-aggregates. In order to differentiate mechanisms from other sorts of non-aggregates, I have argued that mechanisms have an active organization. Specifying the functional, spatial, or temporal organization of a system can specify active organizations. Finally, some non-aggregates, though having an organization, fail to have an active organization, one due to a differentiation either in the parts or ways of interacting between the parts of a system. But, as I illustrated with the case of a lump of graphite, some non-aggregates do have an active organization. In order to further restrict the class of mechanisms, further constraints are needed.
2.2.5 Mechanisms and Emergents: Mechanistic Role Functions

I have argued that mechanisms are at least non-aggregative, actively organized systems. As argued above, the dynamical systems invoked in cognitive neuroscience are non-aggregative, actively organized systems, just like other mechanisms. The initial investigation into what further properties distinguish mechanisms from other non-aggregative systems resulted in a particular definition of a mechanism that distinguishes component entities and activities that are characterized by spatiotemporal properties, active organization, and causal powers. However, prima facie, this prevents considering the kinds of dynamic mechanisms *au courant* in contemporary cognitive neuroscience from being classified as mechanisms. And the relaxation of these constraints to only an active organization of parts was too general; lumps of graphite and some other non-aggregates satisfied these criteria. We need a more fine-grained approach.

Craver argues that in addition to active organization, identification of the function of the mechanism’s components plays a key role in differentiating mechanisms from aggregates. Part of analyzing mechanisms involves the attribution of role functions to the mechanism’s components. Specification of role functions involves two descriptions. First, the attribution of a role function involves describing a system component in terms of its components, activities and how they are organized. Specifying mechanic role functions, those activities that are played by the components of the
mechanism, requires either the spatial, temporal, or functional characteristics of the components also be provided.\textsuperscript{45} Second, the attribution of a role function involves describing the component’s role in terms of the contribution it makes to the larger containing system, the properties and activities of that component. As Craver succinctly puts it, “[a]ttributions of mechanistic role functions describe an item in terms of the properties or activities by virtue of which it contributes to the working of a containing mechanism, and in terms of the mechanistic organization by which it makes that contribution” (Craver 2001, p. 61).\textsuperscript{46} Thus, in addition to violating the Wimsatt criteria, specifying active organization and ascribing mechanistic role functions to the components identified in the organization of a system are additional sufficiency criteria for identifying mechanisms.

Levy argues for a similar point in clarifying the conceptual situation by distinguishing and analyzing what he calls ‘machines’ (Levy 2014). The concept of

\textsuperscript{45} These mechanistic role functions just are a species of causal role functions, defined in terms of capacities, after Cummins (Cummins 1975). The activities that Craver cites are explicitly the capacities that Cummins cites (Cummins 1975; the shift in terminology is meant to emphasize “the way that mechanisms work as opposed to the way that a set of parts has the capacity to work or is disposed to work” (Craver 2001, p. 58, footnote 2)). Craver is emphasizing the actual work performed by the components as opposed to the broader counterfactual or dispositional properties of the components.

\textsuperscript{46} Craver claims that ascriptions of mechanistic role functions “are detailed and precise to the extent that they can be explicated in terms of specific details of how an item fits into the active, spatial, and temporal organization of a mechanism that we seek to understand” (Craver 2001, p. 61). In light of the immediately previous discussion on spatial, temporal and functional organization, we can recast these ascriptions in terms of specific details of how the system components are functionally organized, without the need for the spatial and temporal properties of those components.
orderliness in a system is at the crux of Levy’s distinction between machines and other sorts of systems. Suppose there is a system S that exhibits behavior B. Then, “S is orderly to the extent that: (a) Distinct components of S play different roles in bringing about B. (b) Components play their roles in virtue of local relations to other components” (Levy 2014, p. 4). For the first condition, a central component of orderliness is that the different parts of the system play different roles, that is, different parts have different functions.

“The underlying idea is that in a machine-like system, an overall job is allocated to distinct sub-elements, and therefore understanding the system involves identifying these elements and their contribution to the system’s overall behavior. Talk of distinct components brings to mind spatially defined parts. Components are often spatially separate, but need not be. The important point concerns functional distinctness, i.e. distinctness with respect to the difference made by components” (Levy 2014, p. 5). Spatial properties of these components are not required for their individuation. Rather than the spatial properties of the components, the functions of those components serve to individuate them. Function here is understood as differences in how the components make a difference. “[I]mplicit in the definition is a notion of difference-making: I take it that to say, for instance, that components contribute in virtue of the relations among them, is to say that such relations make a difference to the system’s overall behavior”, and these different contributions can serve to individuate those components (Levy 2014, p. 4). As discussed above, the interactions between the components constitute the functional organization of such systems, but as Levy makes clear, this functional organization can be analyzed in terms of the different functions of the components, the differences in how components affect the system. Thus, there is no functional
organization without a specification of the functions of the system components.

However, different functions can result in the same functional organization. There are, after all, different ways to put together a car or a cognitive system.\textsuperscript{47}

Though Levy focuses on machine-like systems, by a separate but parallel line of reasoning, a mechanism’s components possess role functions that determine the mechanism’s organization. Mechanisms have components, and these components have mechanistic role functions, the way these components contribute to the overall activity of the containing system in virtue of the component’s properties and activities. These mechanistic role functions are the differences that the component contributes to the containing system and are described by the way the component interacts with other components of the system. In addition to being non-aggregates, mechanisms are actively organized, and the components of the mechanism have mechanistic role functions.

\section*{2.2.6 Mechanisms and Emergents: Functional Locality}

While I have argued that mechanisms can be differentiated from other non-aggregates on the basis of actively organized, mechanistic role function assigned

\textsuperscript{47} There may not be different ways to put together some mechanisms, depending on how much liberality we permit in analyzing mechanisms. For example, while different groups of mechanistic role functions may result in the same output for an organism, some combinations of such functions may not be found in the natural world. If we limit our classifications of creatures to their actual biological organization, then there may be only one way to put together a mechanism of that type. The other possible combinations of mechanistic role function that produce the same output once instantiated in the actual world would then be ersatz tokens of that type.
components, this definition is still not sufficient. A lump of graphite will have an active organization, and the different parts of the graphite will possess different contributions to maintaining the layering structure of the lump. Levy’s second condition provides further constraint. It constrains how the functions of the components are executed, viz. in virtue of local relations between the components.

“Condition (b)… should be understood in terms of difference making — the idea is that in machine-like, orderly systems, local relational properties of components make a difference to the system’s overall behavior…. local relations are ones a component has with a designated, typically relatively small, subset of the system’s other components — its causal neighborhood, so to speak. Thus, we have order to the extent that a system’s behavior depends on interactions among small subsets of its components….“ (Levy 2014, p. 5-6).

The need for local relations, as opposed to general or non-local relations, is imposed to capture the sense that orderliness in a system arises from the system’s parts being integrated, as “an orderly system exhibits internal integration. It is possible for a system to fulfill condition (a), i.e. for its parts to contribute differentially, without these contributions being integrated” (Levy 2014, p. 6). These local relations are defined over the local causal neighborhood. For Levy, “[a]n orderly system exhibits an internal division of labor, analogous to that present in many manmade machines: each part does something distinct and recognizable, but there is also interdependence among parts, so that the system’s overall behavior is an integrated product of their activities” (Levy 2014, p. 4). Orderly systems are sets of integrated components with particular functions, and the local relations between the components determine those functions.
Local causal relations can have two different interpretations: local functional relations and local physical relations. Local physical relations refer to the types of properties Craver and other mainstream mechanists invoke in their concept of mechanism, embodied in (M) above. These are spatial and temporal properties of the component entities and activities of mechanisms. But local functional relations may also exist that do not specify the spatiotemporal properties of the bearers of those functional properties.

Consider the simple system for executing addition described above. The different entries on the input-output tape, the different states of the read/write head, and the different moves of the read/write head stand in particular functional relations with each other as described by the machine table and that constitute the functional organization of the device. These relations are the local functional relations, in counterpart to the local physical relations seemingly picked out by Levy. What does it mean for relations to be functionally local, if not that the relations are determined by the spatial and temporal properties of parts in the local causal neighborhood?

I propose that functionally local properties are those determined by the current property of a component based only on the properties of the containing mechanism. The nature of these dependencies determines just how functionally local the interactions between the system components are that determine the overall behavior of the system.
Recall that a dynamical system is described by its system evolution function, how the parts and properties of the system evolve over time and with respect to each other, and that the state evolution equations mathematically described the evolution of those components. I propose that if the system evolution functions refer to intrinsic properties of the system, that is, the variables and parameters that appear in the state evolution equations don’t refer to properties or components external to the system, then the system is functionally local. The adding machine’s components satisfy this notion of functionally local properties. The adding machine is an integrated system, with behavior determined by the integrated product of the components of the system according to functionally local interactions. The read/write head, the tape, and the actions of the read/write head on the tape are functionally organized in such a fashion that when operating together, particular inputs (the summands) result in a particular output (the sum). This property also seems to hold for dynamical systems of cognition, due to the differential contribution of the components to the system. Those systems appear integrated, with dynamical parts and activities that result in the overall functional organization of the system. These systems’ state sequences are governed by the component dependencies described by the system evolution functions that characterize the dynamical system.48

48 This is crucially linked to the distinction between systemic and componential dynamical systems
Furthermore, we can define different types of functional locality by examining how these properties determine the functions of the components. For example, the influence of the history of the system on the functioning of the components can vary, and at its most limited, only the immediately previous time-step is relevant: time-restricted functional locality (in probability theory, this is called the Markov property). What counts as the immediately previous time-step will vary between different systems. In the case of the adding machine, time is discretized by the transition between states according to the description encoded in the machine table. The adding machine satisfies time-restricted functional locality because the sequence of transitions prior to the previous transition do not matter to the next transition. The output of the adding machine results from the most recent state of the read/write head and the input to the system. Thus, the system adds because of the integration of the components operating according to time-restricted functionally local interactions. Only the previous transition matters, as the previous transition situated the mechanism in its current context, that is, where it is in its machine table.\textsuperscript{49} Time-restricted functional locality need not hold for the discussed in the introduction; see chapter one, subsection two.

\textsuperscript{49} This need not be the case: we can imagine a ‘memory register’ for Turing machines, a state register whose value is determined not by the most recent state the system has been in, but rather by the history of such states.
dynamical systems of cognitive processes. These processes may exhibit hysteresis, the
dependence of the current system state on the system’s history. This is hypothesized to
hold for certain neural integrators (Koulakov et al. 2002). In general, the dynamics of
neural systems are modulated by the history of the organism and the recent history of
the physical mechanism, what states it’s been in.

Similar dimensions determining functional locality of the interaction of parts can
be imagined, and are widely discussed in the philosophical literature under the general
heading of ‘complexity’. For example, non-linearities, sensitivity to initial conditions,
feedback and other dynamical properties present in the evolution of systems can serve
as different dimensions of functional locality, and this but a partial list (Lewin 1992;
Kauffman 1995; Bak 1996; Lorenz 1996; Goldbeter 1997; Sole and Goodwin 2001;
Camazine et al. 2001; Mitchell 2003; Amaral et al. 2004; Mitchell 2012). I thus propose
that functional locality, which determines the roles played by the components of a
system, is itself a multi-dimensional complex of properties, rejecting the monolithic
interpretation imposed by Levy’s analysis.

Granted that dynamical systems can have local functional relations and that the
behavior of these systems is the result of the integrated activity of the components, a less
restrictive understanding of Levy’s second condition is possible. Recall that the second
condition states that the system’s components play their role in virtue of local relations
to other components, understood as the causal relations the components stand in. These local relations, though, can be functionally local as much as causally so, and the system’s behavior results from the functional or causal interactions among small subsets of its components. We might classify all the systems that satisfy this more liberal condition as machine-like, and the systems that satisfy Levy’s more robust constraints as machines.

The just completed discussion has found a graded assessment of the class of non-aggregates, with machines as well as machine-like systems. The superset of these systems I propose to identify as mechanisms. Mechanisms are sets of components actively organized in a certain fashion, where those components exhibit activities defined by the mechanistic role functions that the mechanism components possess. Insofar as machine-like systems, like machines, exhibit an actively organized set of components, entities and activities that comprise the system, and differential roles for these components that are determined by the functionally local relations between the parts, machine-like systems are mechanisms as well.50

50 As part of an extended discussion on the role of mechanisms in biology, Woodward pushes back on the idea that biological explanations are mechanistic (Woodward 2013). Since Woodward adopts his interventionist approach to causation in order to level his objections, I won’t discuss him in great depth. Woodward identifies three properties of mechanistic explanation that provides a deeper understanding of why an input-output explanation holds. First, in “successful mechanistic explanations, the generalizations describing intermediate links are typically expected to be... more stable than the overall” input-output relationship we are trying to explain (Woodward 2013, p. 49). The intermediate links are more stable that the overall input-output relationship in the sense of having a larger extent of changes in background circumstances yet still preserving satisfaction of Woodward’s interventionist causal condition (M). Second, mechanistic explanations are modular: for a given set of causal relationships G that compose the system S,
2.3 A New View on Mechanisms

In light of the above discussion, conceptual space has been created for a broader notion of mechanism. (M), the initial definition of a mechanism above, is unduly restrictive, excluding Turing machines, the dynamical systems in cognitive neuroscience, and potentially other systems. The entities and activities, previously

such a system is modular “to the extent that each of the individual Gi remain at least somewhat stable under interventions that change the other Gi... the extent to which one can change or disrupt the causal relationships governing one subset of components in the system without changing or disrupting the causal relationship governing other, distinct subsets of components” (Woodward 2013, p. 51). Finally, third, mechanistic explanations exhibited fine-tunedness, the idea that “… the specific details of the causal properties possessed by the components, and the way in which those components are connected to each other, including their spatio-temporal relations, matter a great deal to the outcome produced, in the sense that if these features or organization were different, the outcome would be different…. Not only is the overall behaviour of typical mechanisms sensitive to the details of their spatio-temporal organization, but we also expect that there will be systematic relationships or correspondences between the causal relationships that characterize such devices… and spatio-temporal facts about the relationships among components” (Woodward 2013, p. 56). Woodward’s third condition is similar to Levy’s orderliness, and the role of spatiotemporal organization, causal structure, and the correspondence between the two is reminiscent of Craver as well.

In a similar vein as Levy, Woodward concludes that “[t]his sensitivity of typical machines to details of organization—to spatio-temporal relationships among components, and to just which components are present—is, of course, connected to and motivates the idea that it is important to make such details explicit in constructing mechanical explanations.... To the extent that such details do not matter—to the extent that there is reason to think that some system would behave in the same way (for some behaviour of interest) even if the spatio-temporal organization of the components were changed considerably, or even if some or many of the components were replaced with others with different causal properties— then the motivation for constructing explanations that focus on them is correspondingly diminished, at least on a conception of explanation according to which this involves the exhibition of difference-making factors” (Woodward 2013, p. 57).

While the spatiotemporal particulars would no longer be of importance to the explanation, this does not mean that the system does not exhibit telltale signs of mechanism or machine-likeness in a broader sense, as I’ve been at pains to argue. A full response to Woodward’s concerns would involve investigating the relationship between the view of dynamical mechanisms laid out herein and Woodward’s interventionist approach to causation; alas, an investigation for another time.
characterized by picking out their spatiotemporal properties, need not receive a
spatiotemporal characterization. This flexibility permits different types of active
organization, including functional organization. Furthermore, mechanisms can have the
functions of their parts determined by functionally local interactions, as well as
spatiotemporally local causal interactions. Accommodating these broader constraints
forces specific revisions in the definition of a mechanism.

But how radical a revision is required? In considering the role of mechanisms
in cognition, including the role of dynamical models, Kaplan and Craver recently
reassessed what makes a mechanism (Kaplan and Craver 2011). They contend that for a
model to possess explanatory force, it must not only save the phenomena—that is, be
empirically adequate—but “the model must in addition reveal the causal structure of the
mechanism. This will involve describing the underlying component parts, their relevant
properties and activities, and how they are organized together causally, spatially,
temporally and hierarchically” (Kaplan and Craver 2011, p. 605). This appeal to a
mechanism, as Kaplan and Craver emphasize, has been recently freed from many of the
prejudices historically attached to the concept of a mechanism:

“First, to insist on mechanistic explanations is not to insist on explanation in terms of
simple machines governed by strict deterministic laws or in terms of physical contact,
energy conservation, or any other fundamental or otherwise privileged set of activities.
Second, mechanistic explanation is not only downward looking, peering down into the
mechanisms within mechanisms by which things work, but it is also contextual, situating
mechanisms within higher-level causal structures. The parts need not be spatially
localized within the system. Nor need their activities be sequential, from beginning to
end; they might involve (negative or positive) feedback loops or recurrent connections between components. Frequently, features of the spatial and temporal or dynamic organization of the components and their activities are explanatorily relevant and so are included in the models.... Other times, it matters more who communicates with whom than precisely where the participants are located....” (Kaplan and Craver 2011, p. 605-606).

While this relative loosening of the role of mechanisms in cognition is welcome, Kaplan and Craver do not go far enough. The fundamental elements of mechanisms adverted in (M) above are still present: spatiotemporal characterization of entities and activities, albeit the explanatory importance of the precise locations of the entities is de-emphasized and the activities may be complex and dynamic. Similarly, though describing mechanisms as simple machines characterized in physical terms or in terms of deterministic laws is not required for mechanisms (reminiscent of Levy), mechanistic explanations still impute a particular causal structure specifying how the input to the mechanism is transformed into the output, and situate the mechanism in a broader causal complex. However, these causal and spatiotemporal restrictions are too tight. A fundamental revision to (M) is required; minor revisions just won’t do.

Recall the first part of the above definition of a mechanism (M), regarding the set $E$ of entities, which

(i) is characterized by their physical properties (such as spatiotemporal location, shape, electrochemistry, etc.).

Specifying the type of dynamical mechanism need not include spatiotemporal properties of the mechanism’s parts. In the integrate-to-bound mechanism, the threshold is a part of the mechanism that does not specify any aspect of its spatiotemporal implementation.
Likewise, the integrative activity that precedes thresholding also lacks spatiotemporal specification. The dynamical mechanism can’t be said to spatiotemporally contain those parts, either, since it does not have any spatiotemporal specification. In the RDMT physiological data presented above, the integration of evidence corresponds to an increase in the firing rate of LIP neurons over the course of the trial. Compare that particular implementation of the DDM in an integrate-to-bound mechanism to the implementation of an integrate-to-bound dynamical mechanism adduced in the patch-leaving task. Both physiological responses are taken to implement a part of the dynamical mechanism, the integration-to-bound mechanism, but the physiological responses are fundamentally different. Despite these physiological differences, both physical mechanisms implement the same type of dynamical mechanism in virtue of implementing the same dynamical component entities and activities.\(^{51}\) In general, any material implementation specification may be absent by intent, to permit the various biophysical implementations that might be found. This abstraction away from the spatiotemporal particulars results in a revised first condition, with the set of entities \(E(i^*)\) is characterized by their physical properties (such as spatiotemporal location, shape, electrochemistry, etc.) or their formal properties (such as function, relation to other entities, etc.).

\(^{51}\) This identity criterion for mechanisms presupposes that a set of component entities and activities is already in hand. Identifying these components will depend on the degree of abstraction in our description of the mechanism, another problem altogether; see Craver for discussion (Craver 2009, p. 585ff).
Rejection of the spatiotemporally individuating properties of components forces the turn to the function of components or their relations to other components, that is, the active organization of the mechanism, the role functions attributed to the components and the functionally local nature of their integrated interaction.

Recently, Piccinini and Craver have suggested a similar sort of revision to the notion of a component, in order to accommodate concerns about psychological mechanisms (Piccinini and Craver 2011). Noting that components are “sometimes identified by their structural properties”, “including their location, shape, orientation, and the organization of their sub-components”, they say that “[t]he term “structural” does not imply that the components involved are neatly spatially localizable, have only one function, are stable and unchanging, or lack complex or dynamic feedback relations with other components. Indeed, a structural component might be so distributed and diffuse as to defy tidy structural description, though it no doubt has one if we had the time, knowledge, and patience to formulate it” (Piccinini and Craver 2011, p. 291, italics in original).

Components also have functional properties that contribute to their identification, properties “specified in terms of effects on some medium or component under certain conditions” (Piccinini and Craver 2011, p. 291). Once again, the suggested revision is in the right direction but not strong enough. Certain structural aspects of the components, such as their location, shape, or orientation, may be absent in the specification of the components of dynamical mechanisms, regardless of the time, knowledge or patience we possess. The list of structural properties needs to be curtailed as recommended in (i*) above.
Furthermore, the characterizations of the activities or operations also need to be revised. Recall that in M activities (iv) are characterized by their physical properties (such as spatiotemporal location, rate, duration, etc.) or their constitutive properties (such as the subset e of E that initiate, participate in, or result from them or the changes in the properties of some subset e of E they produce).

However, activities of functional mechanisms are not necessarily individuated by spatiotemporal location, rate, or duration, and nor are their structural or organizational properties specified in such a concrete way. Since integrate-to-bound mechanisms can be implemented by different token mechanisms for different behaviors, as depicted in the contrast between the LIP responses and the ACC responses, the activity of integration can’t have a spatiotemporal, rate, or duration specification. The integration doesn’t even have a uniquely specified function that it executes, as for the DDM, it is taken to literally integrate evidence, but in the patch-leaving task, it is hypothesized to execute some part of the Marginal Value Theorem, the formal model that specifies optimal behavior in such foraging decisions. The specification of these activities revises this fourth conditions, where activities (iv*) are characterized by their physical properties (such as spatiotemporal location, rate, duration, etc.), their functional properties (such as how they transform the input to the mechanism), or their constitutive properties (such as the subset e of E that initiate, participate in, or result from them or the changes in the properties of some subset e of E they produce).

In addition, the mathematical functions that are present in the characterization of the cognitive mechanism may not map on to the physical implementing operations at all or
in any consistent way within organisms, across organisms or even across cognitive systems. The nature of dynamical mechanisms establishes the formal relations that the implementing mechanism’s parts and activities must map on to and whose outputs it must reproduce in their concerted action. In short, the dynamical mechanism constrains the physical implementation, without requiring that those very same activities or functions be implemented.

In reply to these suggested revisions, the proponent of (M) could object that the “mechanisms” that result are just collections of organized activities. For example, the operations that are specified in the dynamical mechanism may not be the operations that are executed by the concrete counterpart because all that is needed is some preservation, up to some degree of precision, of the products of the dynamical mechanism by the results of those operations as are performed by the implementation. For example, a linear operation in the dynamical mechanism may be implemented by a monotonic sigmoidal activation curve, such that within a particular dynamic range, the sigmoid preserves the cardinal relations present in the linear operation. Likewise, the operations specified in the formal model being implemented to allow the system to achieve its processing goals and exhibit the specified behavior may also not be the actual operations executed by the dynamical mechanism that implements the formal model. This important point has consequences for accounts of mechanistic explanation in cognitive neuroscience, such as presented by Kaplan and Craver (Kaplan and Craver 2011; Kaplan 2011). I explore these consequences elsewhere.

This objection is related to the idea of mechanism schemas and mechanism sketches (Machamer et al. 2000; Craver 2007). Mechanism sketches are “[i]ncomplete models—with gaps, question-marks, filler-terms, or hand-waving boxes and arrows…. Mechanism sketches are incomplete because they leave out crucial details about how the mechanism works” (Piccinini and Craver 2011, p. 292). Piccinini and Craver have recently argued that psychological explanations which proceed via functional analysis, the explanation of a system’s capacities in terms of internal states, result in mechanisms sketches: ““internal” states either are not really internal, in which case they constitute a system-level explanandum for a mechanistic explanation, or they are internal in the sense of being states of components. As we have seen, there are two ways to think of components. On one hand [sic] are structural components. In this case, functional analysis by internal states is a promissory note on (a sketch of) a mechanistic explanation. The analysis postulates states of some structural components, to be identified by a full-blown mechanistic explanation. On the other hand, components may also be functionally individuated components or black boxes. (For instance, the read-write head of
properties alluded to in (i*) for defining the entities that are mechanism components, it could be objected, are in fact merely certain sorts of activities. The components of dynamical mechanisms are functionally individuated, but this functional individuation reifies activities into entities (or so the objection goes). Consider the integrate-to-bound mechanism again. One of its components, as argued earlier, was a threshold: a particular point in the state-space of the system that, upon reaching, the system resets to its baseline state. But the threshold is individuated on the basis of its functional properties, and in particular, the point in the system’s state-space that results in resetting. But resetting (and the attendant retroactive identification of the process of thresholding) is an activity: it is defined in terms of the change in the system evolution function, from one area of the system’s state space to a different area.

Turing machines is a paradigmatic example of a black box.) When components are construed as black boxes, functional analysis by internal states becomes boxology….” (Piccinini and Craver 2011, p. 300).

Boxology is the analysis of a system’s capacities in terms of components (so-called black boxes) individuated by their inputs and outputs.

It’s somewhat unclear the strength of the claims of Piccinini and Craver. For example, by reducing the outputs of the functional analysis of a system capacity to a mechanism sketch, does that incomplete model qualify as a non-aggregate? Insofar as such models are representations, and representations are some sort of emergent, they would seem to be non-aggregates. However, insofar as such sketches are merely models, they would not seem to denote any unique system, but rather a class or family of physical mechanisms. This imputation of dynamical models as mechanism sketches merely representing physical mechanisms belies the error in their reasoning: such models do not have the physical mechanisms as their representational targets, but rather the dynamical systems (or, as I’m presently arguing, mechanisms) that constitute cognitive systems. Further discussion of this issue occurs in the next chapter.
This objection begs the question against dynamical mechanisms. In (M), activities are defined in terms of the physical properties or the constitutive properties of the activity. The physical properties are not germane to the current case of dynamical mechanisms’ activities. The constitutive properties of dynamical mechanisms’ activities are the entities that participate in the activities or the changes in the properties of those entities that the activity produces. Thus, to identify thresholding or resetting as an activity implies the existence of some set of entities that constitute the activity or whose properties change in a way that constitutes that activity. The objector could maintain that the entities involved are spatiotemporally specified, as laid out in condition (i) of (M). But determining whether such specification is necessary is precisely the point of the current discussion; insisting that it is necessary begs the question against the dynamical mechanism view because it fails to acknowledge the more liberal notion of mechanism currently presented.

Recently, as part of a broad skeptical commentary on mechanisms in biology, Dupré has opined that “[i]t seems to me that there are good reasons to think that biological systems—organism, cells, pathways, etc.—are in many ways quite

\[\text{\textsuperscript{54}}\] A banal interpretation of this objection notes the difference is merely terminological; I’m ignoring this response.
misleadingly thought of as mechanisms” (Dupré 2013, p. 28). Dupré militates for a processual approach to biology in general (and microbial biology in particular), including recasting the fundamental units of analysis in biology as processes (Dupré 2013; Bapteste and Dupré 2012). The deepest issue according to Dupré is that biological entities are not stable, and constitute explananda as much as explanantia.

---

Woodward’s criticisms are similar in spirit (Woodward 2013; vide supra).

Dupré’s criticisms are fundamental and far-reaching, and I address only one in the main text. In the context of a discussion of mitotic spindle production, Dupré argues that “[a] variety of mechanisms exist that contribute to the serving of a biological function, but none of them is necessary, and in many cases none is sufficient, to bring about the performance of the function. Of course we could at this point… move up to some higher level and say that this was where the relevant mechanism must be found…. Now the trouble is that even if… there must be such a higher-level mechanism, this just isn’t what scientists typically mean when they say they are investigating the mechanism(s) of spindle production: they are referring to the various processes that can, at different times in different contexts, contribute parts of spindles” (Dupré 2013, p. 26-27). Thus processes, and not mechanisms, are the focus of biological investigation. Dupré notes that “[p]aradigmatic machines—cars, dishwashers, computers—consist of a number of parts, typically more or less rigidly connected. The constituent parts gradually wear out, and the machine lasts as long as they are replaced piecemeal. Certainly we can relax the assumption of rigid connection, and extend the concept perhaps to the reactions in a vat in a chemical works. Here very large numbers of parts move around freely and undergo predictable kinds of interactions when they encounter appropriate partners. We have here fairly clear representatives of the entities and activities demanded by new mechanists” (Dupré 2013, p. 28-29).

But biological mechanisms are not like this, for a variety of reasons, in addition to the concern about the processual nature of biological entities mentioned in the text. First, biological systems constantly engage in self-maintenance: “organisms for instance, constantly rebuild and replace their worn parts” (Dupré 2013, p. 29). Furthermore, such systems do not have clear termination conditions. “Contrary to some versions of mechanism, there is no a priori reason why the process should end, and hence no terminal condition. Lineages of organisms have no mechanistically inbuilt tendency to terminate, though in the end no doubt most will do so” (Dupré 2013, p. 29). Third, the entities that comprise biological systems exhibit many causal roles and interactions. “In at least the paradigm cases of machines, we can very precisely identify the causal role of parts and the effects of manipulating or removing them, and their interactions with other entities are generally specific and limited. These characteristics are again inscribed in the details of recent accounts of mechanism: it is the specification of parts and their behaviours (activities) that provides the mechanistic explanation. However, in biological systems parts often have multiple roles and multiple causal interactions” (Dupré 2013, p. 29).
“All these accounts start with an inventory of entities, and though these may not be called ‘things’, they are pretty clearly conceived as a fairly stable inventory of fairly stable things. But the entities that form the hierarchy of biological ontology are not stable. They are, rather, stabilized over a very wide variety of timescales, and the processes of stabilization are a fundamental part of the explanation of the activities of living systems. Living things are the explananda in biological sciences at least as much as they are the explanantia” (Dupré 2013, p. 30).

This leads Dupré to make the propaedeutic point that biology is about processes, not entities. Processes are fundamentally active, and this focus on the active nature of biological phenomena challenges the argument that biology is concerned with uncovering mechanisms. The more liberal definition of activities and entities outlined above, however, should make ample room for such active processes. No longer need entities be conceptualized as a stable set of fairly stable things. The entities that enter into mechanisms can be specified by their formal properties as much as by their physical ones, and these formal properties can be intrinsically unstable. Similarly, the activities in which these entities engage need not be stable either, being specifiable by their functional properties as much as by their physical or constitutive ones.

To conclude this discussion on mechanism, I will present a revision to the definition (M) presented above. The picture resulting from this discussion, in combination with the discussion of the case studies earlier, is of two types of mechanisms, dynamical and physical, with physical mechanisms implementing dynamical mechanisms in virtue of some kind of mapping, in addition to the dynamical mechanism executing the formal model for the behavior in virtue of a second mapping.
Incorporating this new focus on implementation, we thus have the following revised definition of a mechanism:

(M*) A mechanism M* is a set E of entities that (i*) is characterized by their physical properties (such as spatiotemporal location, shape, electrochemistry, etc.) or their formal properties (such as function, relation to other entities, etc.), which (ii) are arranged in an organization O such that (iii) when O-organized, E exhibits a set A of activities that (iv*) are characterized by their physical properties (such as spatiotemporal location, rate, duration, etc.), their functional properties (such as how they transform the input to the mechanism) or their constitutive properties (such as the subset e of E that initiate, participate in, or result from them or the changes in the properties of some subset e of E they produce), and that (v*) either directly causally produces or, once implemented by some mechanism M, causally produces the explanandum phenomenon.

The disjunction in the last condition allows for the implementation of a dynamical mechanism by a physical mechanism to cause the behavior. The physical mechanism M need only satisfy the definition (M) stated at the outset. This implementation condition skirts the issue of whether the organization of the mechanism is causal, and whether the concept of causality operative in mechanistic contexts is necessarily physical. This definition of mechanism expands the conceptual boundaries of mechanisms so as to include the dynamical mechanisms that implement the formal models of cognition.

### 2.4 Conclusion

There is a different, more inclusive sense of mechanism, captured by (M*), that is more general than that extant in the recent literature, as captured by (M). Furthermore, I hope I have motivated viewing the dynamical systems of cognition as mechanisms of
this more generic variety. These mechanisms violate the Wimsatt criteria for aggregativity, and in addition, their parts are described formally or relationally, possess an active organization that exhibits certain activities described functionally or constitutively in terms of those formal parts, and that are implemented by physical mechanisms.

In the remainder of this dissertation, I will develop a theory of cognition, founded on these dynamical mechanisms. As I stated earlier, I will highlight three main principles. The first, that cognitive systems execute formal models of processing, conceptually distinct from the mechanisms that implement them, was discussed earlier. Further discussion can be found in Appendix C. The second, that cognitive systems are composed of dynamical mechanisms implemented by physical mechanisms, will be discussed in the next chapter. Finally, the third, that the same dynamical mechanisms are reused to implement different formal models, thus accomplishing different processing goals, will be discussed in the last chapter. Strap yourself in: the ride is very kinetic!
3 Mental Machines

The second principle states that cognitive systems are composed of dynamical mechanisms implemented by physical mechanisms. This principle has two components. The first component states that cognitive systems are composed of dynamical mechanisms, where these dynamical mechanisms are type distinct from physical mechanisms. As I argue below, the presence of regular reoccurrences of the same set of dynamical properties in many different physical mechanisms justifies the conclusion that these are distinct types of mechanisms. The second component of the second principle states that the physical mechanisms implement the dynamical ones. The logical nature of the implementation relation was discussed above. I won’t be discussing the second component further.

In this chapter, I will illustrate how these dynamical mechanisms are the mechanisms of cognition. First, recalling the case studies illustrated earlier, I discuss the integrate-to-bound mechanism again. The integrate-to-bound mechanism is just one example of the types of simple dynamical mechanisms that compose cognitive systems. The second principle entails that cognitive systems are decomposable into sets of dynamical mechanisms. A strong skeptical thesis, the epistemic objection, holds that physical cognitive systems, such as brains, have such complex dynamics that they are not decomposable. A related skeptical thesis, the systemic objection, holds that physical
cognitive systems are complex in particular ways that prevent such a decomposition into component mechanisms. I will respond to both objections below. I will also demonstrate how implementing multiple dynamical mechanisms in the same neural area results in more complex dynamics.

Second, I present the argument that dynamical mechanisms and physical mechanism are distinct. The grounds for this claim are to be found in the widespread use of models in reasoning about cognitive systems. I present a series of objections to inferring dynamical mechanisms from dynamical models, models of dynamical mechanisms.

The presence of these dynamical mechanisms, which I contend are found in a wide variety of cases (divisive normalization, synaptic reverberation, center-surround inhibition, and others), constitute a fundamental level of functional organization for the system. They also play a prominent role in explaining the capacity that cognitive systems have to behave intelligently and flexibly. In short, they are necessary for cognitive systems, and the second principle reflects this central role for dynamical mechanisms.

### 3.1 Simple Mechanisms for Cognition

Recall again the case studies of LIP neuronal responses during the RDMT and ACC neuronal responses during the patch-leaving task. These two case studies illustrate
the repeated dynamical mechanisms the brain deploys to execute cognitive functions, in these cases, different types of decisions. The same dynamical mechanism is present in both LIP and ACC, but for different formal models of processing. Though different formal models characterize the behavior in the foraging and perceptual tasks, and thus different processing demands are present for the system, the brain executes these models by implementing the same dynamical mechanism, albeit with different neurophysiological mechanisms between the two cases. The physical mechanisms are different between the two areas: LIP neurons exhibit gradual increases in activation, while ACC neurons exhibit a series of increasing transient activations at a much longer timescale. Nonetheless, despite the difference in neuronal activity patterns and formal models implemented, the two areas are both instances of the same type of dynamical mechanism, an integrate-to-bound. What makes both of them instances of that particular kind of dynamical mechanism is the implementation, in both areas, of the entities and activities of the integrate-to-bound mechanism: a baseline, a build-up, and a boundary, and resetting, integrating, and boundary crossing.

The examples I have described are for relatively simple dynamics observed in physical systems during relatively simple behavioral experiments.¹ I maintain that

¹ I don’t mean anything more than an intuitive distinction between simple and complex dynamical mechanisms. What makes a mechanism simple or complex will, I suspect, be determined by factors that are relative to research program, science, and scientist. There are ways of approaching simplicity through
cognitive systems are made up of repeated instances of such mechanisms, just as more complex electrical circuits are composed of basic circuit components such as resistors, capacitors, and solenoids, or more complex simple machines are composed of basic machines such as levers, pulleys, and block-and-tackle. But this entails a decompositionality thesis: that cognitive systems are decomposable into their component mechanisms. And, given the dynamical nature of my position, the decomposition will result in a set of component dynamical mechanisms. These component dynamical mechanisms are token identical to subsets of dynamical properties of the physical systems.

Following Bechtel and Richardson, decomposition and localization are central to decompositionality (Bechtel and Richardson 2010). Decomposition is the analysis of a system into a set of functions, the assumption that “one activity of a whole system is the product of a set of subordinate functions performed in the system” (Bechtel and Richardson 2010, p. 23). Localization is the “identification of the different activities proposed in a task decomposition with the behavior or capacities of specific dynamical systems theoretic analysis. For example, Kelso defines simple behavior as “a stable fixed point or periodic orbit that appears for low values of” a control parameter, where the point or orbit picks out the location of a system in its state-space (Kelso 1995, p. 20). But even this sort of behavior might be taken as not simple, in another sense, one that is relative to the interests of a scientist or discipline.
components” (Bechtel and Richardson 2010, p. 24). Pursuing these research strategies assumes that the system is decomposable,

“modular in character, with each component operating primarily according to its own intrinsically determined principles. Thus, each component is dependent at most upon inputs from other components, influences other components only by its outputs, and has a specific, intrinsic function…. Localization presupposes that we are confronted with a modular organization such that the components of the system can be subjected to separate study and investigation; it requires that the components have discrete intrinsic functions intelligible in isolation, even if such functions do not independently replicate those of the system as a whole” (Bechtel and Richardson 2010, p. 24-25).

Systems that cannot be analyzed in this fashion are minimally decomposable. Minimally decomposable systems are those that are integrated, wherein “systemic organization is significantly involved in determining constituent functions” (Bechtel and Richardson 2010, p. 26). These systems typically feature feedback relations, mutually interacting subsystems, and are subject to order parameters, system-level parameters that control the local dynamics of the subsystems. For integrated systems, “systemic organization… provides primary constraints on constituent functioning, and constituent functioning is no longer intrinsically determined” (Bechtel and Richardson 2010, p. 26). But intrinsically determined constituent functioning is central to component systems, as in those systems “the behavior of the parts is intrinsically determined” (Bechtel and Richardson 2010, p. 26). These complex dynamical relations prevent the use of decomposition and localization strategies in analyzing integrated systems.

On the theory of cognition being offered here, cognitive functions result from the actions and interactions of many simple dynamical mechanisms. But this requires a type
of decomposition; specifically, cognitive systems must be decomposable into component dynamical mechanisms. The demonstration of the implementation of simple dynamical mechanisms by physical ones does not entail how widespread such mechanisms are, nor that more complicated dynamics, or more complicated cognitive activity, is susceptible to such analysis. Assessing the soundness of this decomposability hypothesis requires assessing the dynamical properties of the physical systems that underlie the dynamical mechanisms, and in particular, assessing whether cognitive systems are minimally decomposable in Bechtel and Richardson’s sense.

Two related objections arise from the relatively simple nature of the mechanisms in these case studies. The first objection claims that the presence of simple dynamical mechanisms does not ‘scale up’: while simple dynamical mechanisms might be inferred, more complex dynamics will confound any attempt at such identification of dynamical mechanisms, as these complex dynamics arise from systemic effects. I call this the systemic objection. On the systemic objection, the top-down effects of integrated systems, such as feedback, effects of order parameters, and interaction and other extrinsic componential effects prevent decomposability.

Both Anderson and Silberstein and Chemero seem to endorse the systemic objection. Silberstein and Chemero argue that systems neuroscience is increasingly turning to network theory and global system properties of the brain to explain cognitive
phenomena. For example, these explanations feature scale-free invariant power laws, such as the decrease by a constant factor in the probability of finding a node in a network with twice as many connections as an arbitrary number. They argue that

“[s]imply put, such global organizational principles or features of complex systems are not explicable in principle via localization and decomposition…. The… many-to-one relationship between the structural and the [graph-theoretic] features illustrates that specific structural features are neither necessary nor sufficient for determining global topological features…. [P]ower laws are explanatory and unifying [because] they show why the macroscopic dynamics and topological features obtain across diverse lower-level structural details. And the “why” has nothing to do with similar structural details of the disparate systems” (Silberstein and Chemero 2013, p. 967).

These power laws and other properties of the brain are determined by the system’s organization. Putative local components will have their behavior dictated by these properties. Thus, the effects of these system-level properties in determining the behavior of putative components violate the requirement for intrinsic determination of function by a component.

Anderson seems to endorse the systemic objection as well.2 In discussing starburst amacrine cells in the retina, whose dendrites are individually sensitive to different motion directions, partly as a result of the influence of other cells in the retina, he writes:

“In considering these issues, I have come around to the idea that we need to give up on the notion of componentiality in the brain. But does the functional architecture of the brain justify this move? I think it does…. [W]hether a local network is helping with motion perception or optokinesis—and whatever functional selectivity it achieves in

---

2 At the end of the next subsection, I briefly discuss how Anderson is open to a revised notion of componentiality.
these different contexts—may be... a matter of global network properties. In this case, apparent local selectivity is not a stable intrinsic fact about a particular neural sub-system, but a result of interactions between local functional biases and global network organization. This reverses the relationship typically assumed when one analyzes the functions of componential (component-dominant) systems….. Here the functional properties of the parts in a given situation seem determined by the functional organization of the whole” (Anderson 2014, p. 70-71).

Anderson argues that we must give up on our usual notion of componentiality in the brain. The function of starburst amacrine dendrites in detecting motion direction and similar such phenomena suggests that the functionality of putative components is limited by higher-level or global system-level properties. In line with Bechtel and Richardson’s discussion of integrated systems, the brain is a non-decomposable, integrated system. Insofar as cognitive phenomena are to be identified with aspects of physical systems, Anderson’s and Silberstein and Chemero’s concerns about neural non-decomposability apply to my decomposability hypothesis.

The second objection argues that the simplicity of the examples conceals the complexity of the cognitive systems behind them. The complexity of physical cognitive systems will prevent us from decomposing them into their components, even if such components exist. I call this the epistemic objection. Both Anderson and Chemero and Silberstein endorse this objection. In discussing his theory of neural reuse, Anderson argues:

“[N]eural reuse doesn’t claim that there is occasional overlap [in the implementation of cognitive modules]; it expects that it will typically (overwhelmingly) be the case that a given [area] is used in multiple complexes; that therefore the members of the implementing set for any given complex... will be implicated in many functional
complexes; and that therefore although not every complex fb would be impacted by a modification to some member(s) of the implementing set of complex fa, many others would. Insofar as this is so, then even though specific pairs of functional complexes will be functionally separable, it would appear that separate modifyability will not be a general characteristic of the brain.

Given this analysis, I want to insist that... the brain is not a nearly decomposable system consisting of separately modifiable parts.... In the brain, function depends much more on the interactions between parts than on the actions of parts....” (Anderson 2014, p. 31).

Anderson is here arguing that our experimental techniques for decomposing cognitive systems rely on separate modifyability, and that these techniques are inapplicable to neural systems. Insofar as our experimental techniques are not up to the task of decomposing complex systems like cognitive systems, then we will not have the methodology to decompose those systems. There is no doubt that much evidence points to fast online modulation of neural networks, widespread plasticity, and many top-down effects and interactive modulations that suggest the brain is an integrated system. This does not entail that the brain, or other physical substrates of cognitive systems, are not componential. However, if top-down, feedback, interactive, and other modulations are the rule, then our experimental practices, which execute discrete manipulations on isolated parts of the system, will not be able to uncover these components. In short, our knowledge of the decomposability of the system is constrained for complex integrated systems like the brain.

Chemero and Silberstein are also representative of this objection. They distinguish functional from structural decomposition, noting that “[o]ne can use
different methods to decompose a mechanism functionally into component operations or structurally into component parts ....” (Silberstein and Chemero 2013, p. 961). They then argue that physical cognitive systems are not decomposable, and specifically that

“brain networks... are nondecomposable and nonlocalizable. There is a degree of functional decomposition for these networks but not structural decomposition. There is no question that graphical and dynamical simulations do describe mechanisms, but they are not merely abstract descriptions of structural mechanisms. The key question here is what is really doing the explanatory work, and the answer in this case is not in the structural or lower-level mechanistic details, or even in the functional details a la ‘boxology’.... [I]n such nondecomposable complex systems, the global topological features act as order parameters (collective variables) that greatly constrain the behavior of the structural elements.... The dynamical interactions here are recurrent, recursive, and reentrant. So there is no sense in which the arrow of explanation or determination is in principle exclusively from the 'lower-level' structural to the 'higher-level' graphical dynamical” (Silberstein and Chemero 2013, p. 966-967).

More complex dynamics will serve to confound the attempt to account for these dynamics in terms of simple dynamical mechanisms. As formulated, complex integrated systems do not exhibit explanatory relations between lower-level structural and higher-level dynamical properties. Insofar as their concept of explanation maps on to epistemic properties, Silberstein and Chemero are endorsing the epistemic objection.

How to respond to the epistemic objection, that the complexity of these systems prevents us from executing a decomposition strategy? My reply to this objection is to adopt Bechtel and Richardson’s distinction between analytic and synthetic strategies for decomposing systems. While analytic decomposition is subject to the epistemic constraints, synthetic decomposition is theoretically motivated, providing inferential
grounds for decomposition. I then argue that a dynamical mechanism approach to
cognition provides the constraints required to satisfy the second strategy.

Bechtel and Richardson provide two different strategies for characterizing
component functions. First, the analytic strategy “is to isolate components physically
within the system and then determine what each does” (Bechtel and Richardson 2010, p.
18). This strategy is “confronted by the fact that smoothly operating systems conceal
their component operations” (Bechtel and Richardson 2010, p. 18). Complex systems
such as brains, and presumably other physical substrates of cognitive systems, will
simultaneously possess many dynamical properties. Out of this bewildering array of
properties, how to identify those that correspond to the component operations, those
functions that are relevant to cognitive processing? Furthermore, this strategy requires
us to physically isolate the components. But physical isolation of the parts of complex
integrated physical systems will prevent those parts from functioning in the way they
do when fully integrated; top-down, feedback, interactive, and other sorts of systemic
modulation will be absent when they are isolated in this way. A complex integrated
system is not subject to the analytic strategy, even if in reality, it is in fact composed of
distinct components.

Second, the synthetic strategy “is to conjecture how the behavior of the system
might be performed by a set of component operations, and then to identify components
within the system responsible for the several subtasks” (Bechtel and Richardson 2010, p. 18). As they note, a “synthetic strategy requires some prior hypothesis about the organization and operation of the system. From an initial hypothesis about the underlying mechanisms, one formulates a model of how the system functions” (Bechtel and Richardson 2010, p. 20). The synthetic strategy requires a preexisting theoretical stance from which an analysis of the system can proceed. In particular, a set of component operations or component mechanisms is drawn upon in determining how to decompose a system. So long as this set of components is justified, then we can appeal to this set to analyze more complex dynamics.

Silberstein and Chemero complain that functional decomposition, but not structural decomposition, is possible for complex integrated systems like brain networks. And Anderson complains that we should be willing to acknowledge “irreducible functional complexity where there appears to be irreducible functional complexity” (Anderson 2014, p. 106). I don’t mean to imply that all of the observed neuronal dynamics—or physical dynamics generally—in cognitive systems will be decomposable in the way suggested.³ The second form of the objection, the epistemic objection, may be right; it might be the case that complex cognitive systems simply cannot be decomposed in a way that leads to our understanding of the mechanisms of

³ Just how far the theory can go in accounting for the dynamics will depend on the science.
cognition. However, we should be open to the strategy of finding decompositions of complex dynamics into simpler dynamical mechanisms. Specifically, prior theoretical commitments can provide the epistemic bootstraps necessary for mapping those functional decompositions on to mechanisms. In reply to the epistemic objection, drawing on a set of previously justified simple dynamical mechanisms to perform the operations identified in a functional decomposition can execute just the sort of structural decomposition that Silberstein and Chemero deny is possible. And this set of simple dynamical mechanisms is justified by looking at cases where the dynamics of the physical device, such as the activity of individual neurons or neuronal groups, clearly indicates the implementation of a particular dynamical mechanism, such as the integrate-to-bound mechanism.

In reply to the objection that simple dynamical mechanisms can’t scale-up, the systemic objection, I contend that simultaneous implementation of multiple such mechanisms does provide more complex dynamics. Positing the existence of component mechanisms accounts for these more complex dynamics. As Bechtel notes, “there are some reasons to be optimistic about the prospects of mechanistic analysis. The primary reason is that the strategy of decomposition and localization is compatible with discovering great amounts of interactivity that gives rise to complex dynamics” (Bechtel 2002, p. 239). This note goes against the spirit of Bechtel and Richardson’s comment that,
with regard to integrated systems, “attempting to understand the operation of the entire machine by following the activities in each component in a brute force manner is liable to be futile” (Bechtel and Richardson 2010, p. 18). The best way to argue for this reply is to work through an example of more complex dynamics arising from simpler dynamics.

3.2 Complex Dynamics from Simple Dynamical Mechanisms

To briefly illustrate how more complicated dynamics arise from the implementation of multiple simple dynamical mechanisms, I will present a case study of co-instantiated dynamical mechanisms from multi-alternative decision-making. The physical mechanisms involved in this cognitive capacity possess more complicated dynamical properties as a result of implementing multiple distinct dynamical mechanisms. Complex behavior can be decomposed into simple components, as I will now illustrate occurs in LIP during multi-alternative decision-making.

3.2.1 Divisive Normalization and the Encoding of Value

In order to illustrate how complex dynamics arise from the implementation of comparatively simpler dynamical mechanisms, a brief discussion of another dynamical mechanism is in order: divisive normalization. Neuronal activity in sensory areas, such as visual areas V1 or MT, often exhibits nonlinear characteristics, such as gain and contrast modulation for representing environmental variables (Andersen and

4 Many thanks to Steve Chang for detailed conversations on the following topics.
Mountcastle, 1983; Heeger, 1992; Heuer and Britten, 2002; McAdams and Maunsell, 1999; Treue and Martinez-Trujillo, 1999). In order to account for these effects, the divisive normalization (DN) mechanism has been proposed (Heeger, 1992; Carandini et al., 1997; Cavanaugh et al., 2002).\(^5\)

DN determines the activity of a particular neuron by dividing the drive from the portion of the stimulus field (the stimuli in space and time) that falls within the neuron’s receptive field (RF; the region of the modality which elicits a neuronal response) by an aggregate response field, pooled across the population. The basic form of divisive normalization is

\[
R = \gamma \frac{D^j}{\sigma^n + \sum_k D^k}
\]

for response of a cell \(R\), input to the \(j\)th cell \(D^j\), and normalization pool summed over the normalization input \(D^n\) (Carandini and Heeger 2012, p. 54). The parameters \(\gamma\), \(n\), and \(\sigma\) are fit to the data, with \(\sigma\) controlling how quickly the firing of a cell saturates, the semisaturation term. When \(\sigma\) is very large, the normalizing input has little effect on the firing rate, and when \(\sigma \sim 0\), the semisaturation term has little effect on the firing rate, and the rate is principally driven by the normalizing pool of activity. Relative to the

\(^5\) Chirimuuta has recently discussed the divisive normalization model, arguing that it is a form of optimality explanation, in conflict with Kaplan and Craver’s mechanistic approach to explanation in cognitive neuroscience (Chirimuuta 2014; Kaplan and Craver 2011). I disagree somewhat with this analysis; regardless, as she is concerned with explanation, I won’t be addressing her argument herein.
normalization input $Dv$, if $\sigma$ is roughly equivalent to that input, the cell becomes moderately driven by the normalization pool.

DN has been proposed for a number of different cognitive functions. For example, the DN model of attention (Reynolds and Heeger 2009) computes the activation of a neuron by multiplying the response to the stimulus by the attention field and then dividing by a suppression field composed of a convolution of the attention field and the stimulus field. The DN dynamical mechanism accounts for some of the effects of attention on visual responses, including response gain changes, contrast gain changes, and combinations of the two. In addition to attention, DN is an effective mechanism for motion integration (Simoncelli and Heeger 1998) and multisensory integration (Ohshiro et al. 2011), suggesting a general role for DN in mediating the implementation of various formal models.

The mechanisms underlying the encoding of cognitive variables like value remain unclear. A recent paper by Louie and colleagues investigated value encoding in the lateral intraparietal area (LIP) by varying the number of targets and their associated values (reward sizes) in a cued saccade paradigm (Louie et al. 2011). Presented with an array of one to three targets in one of seven different spatial configurations, the monkeys made a spatially cued saccade (multi-target cued-saccade task; figure 8, panel a; adapted from Louie et al. 2011, p. 10631). In order to assess firing rates under different aggregate
value conditions, the reward amounts for the targets inside and outside the RF (\(V_{in}\) and \(V_{out}\), respectively) varied systematically across blocks. The effects of value manipulation can be assessed in two epochs, the target presentation epoch (1000 ms from targets on until the cue) and the pre-saccade epoch (500 ms from cue until saccade) (Louie et al. 2011, their Figure 1, Panel A and Figure 2, Panel A).
Figure 8: Task and neural data from the multi-target cued saccade task. Adapted from Louie et al. 2011, p. 10631.
In a two-target version of their task (cued-saccade task), during the target presentation epoch, for a fixed $V_{out}$, firing positively correlates with $V_{in}$, while for constant $V_{in}$, firing rate inversely correlates with $V_{out}$ (Louie et al. 2011, their Figure 1). For the three-target version, total value modulates LIP responses, whether the neuron’s RF contained the cued target or not, with LIP firing inversely correlated with total value ($V_{tot} = V_{in} + V_{out}$) for fixed $V_{in}$ (figure 8).

What models might explain the modulation of LIP neurons by total value during the target presentation epoch? Louie and colleagues fit four different value models to the recorded LIP activity, comparing neuronal responses modeled as linear responses weighted by either fractional values ($V_{in} / V_{out}$) or differences in value ($V_{in} - V_{out}$), or modeled as multiplicative gain-modulation of maximum neuronal response with the gain either a simple divisive normalization ($V_{in} / (\sigma + V_{in} + V_{out})$) or a full divisive normalization ($(V_{in} + \beta) / (\sigma + V_{in} + V_{out})$). The full divisive normalization model achieved the best fit to the neural data (Louie et al. 2011). Similar to the suppression field in the DN model of attention, greater total value led to greater suppression of activity. Louie et al.’s full DN model computes the firing rate $R$ of LIP neurons as a function of the values of the targets:

$$R = R_{max} \frac{V_{in}+\beta}{\sigma + V_{in} + V_{out}}$$ (1)
where $R_{\text{max}}$ is the maximum response gain, $\sigma$ is a semisaturation term, and $\beta$ is a baseline variable. Greater total value increases the normalizing quantity in the DN model, which decreases the modeled firing rate of a neuron, because the fraction representing the value drive for the neuronal response is decreased.

Louie et al.’s data reveal other effects of value on decision-related activity during the pre-saccade epoch. For the three-target task, trials with more targets show greater excursion, that is, larger differences between perisaccadic and evidence-onset (i.e., instruction cue) responses (figure 8, panel b, c). Louie et al.’s data also show that greater value contrast, $V_c = V_{\text{in}} / (V_{\text{in}} + V_{\text{out}})$, drives greater excursion (for constant $V_{\text{in}}$). The larger excursion for greater $V_{\text{out}}$ was not present in their simpler two-target version of the task, where the correct target was color cued (Louie et al. 2011, their figure 1, panel b, d), suggesting that evidence conveyed by the reduction in the number of targets helps mediate excursion. In addition to greater excursions, the Louie et al. data exhibit ‘inversion’. In the presence of a target in a neuron’s RF, the order of average firing by total value in the pre-saccadic epoch is in the reverse of the order seen during the targets presentation epoch (figure 8, panel b, c). The population data also indicate that the inversion occurs before the ‘go’ signal, initiating around the time of the instruction cue (figure 8, panel c).
3.2.2 Integrate-to-Bound and Divisive Normalization: A Case Study of Concatenated Dynamical Mechanisms

The DN model proposed by Louie et al. accounts for the initial effects of value on the firing rate of LIP neurons during the target presentation epoch. But their study also reveals several dynamical effects, such as inversion and differences in excursion, which appear to be related to the initial number of targets and the value contrast $V_c$. These effects have been seen in other multi-alternative decision-making tasks before, such as a multi-alternative RDMT (figure 9; Churchland et al. 2008). Additionally, their data exhibit the stereotypical increase in activity prior to making a saccade that has been observed in many perceptual decision tasks (e.g. Barash et al. 1991; Churchland et al. 2008). DN as formulated to fit the observed changes in firing during the target presentation epoch is unable to accommodate these dynamics. How might the DN and integrate-to-bound mechanisms combine to describe the observed effects of value and target number on firing rates over the time course of decision-making?
Excursion is easy to explain by the concatenated implementation of a DN mechanism and an integrate-to-bound mechanism: supposing the same threshold for movement execution, a lower starting point due to a greater denominator in the normalization equation results in a longer period of integration until thresholding. Once movement execution dynamics begin to integrate, the signal must evolve over a greater portion of the state-space of the system. These longer integration periods correspond to the cases where lower activity for greater value was observed in LIP cells in the Louie et al. data.\(^6\) Thus, excursion results.

\(^6\) As well as corresponding to the cases with four targets, as opposed to two targets, in Churchland et al.’s (2008) data.
Inversion is harder to explain, and several different hypotheses present themselves. First, could inversion result from a higher rate of increase in order to reach the threshold? A faster rate of increase would not result in inversion unless the integration overshot the threshold. However, at least for the RDMT neural data, build-up rates were slower for the four-choice than for the two-choice, indicating both that there was a lower rate of increase and that in fact there was normalization affecting the exponentiation in some way (Churchland et al. 2008). This fact would suggest that the ramp-up is not overshooting as the ramp-up was slower; a slower ramp-up does not entail no overshoot, but such an overshoot would be more surprising, especially if the ramp-up is partly due to recurrent excitation, as has been hypothesized (Wang 2002).

For the cued-saccade task, monkeys were faster for higher $V_{in}$ but showed no dependence for $V_{out}$; that is, for non-error trials where the saccade target was in the cell’s RF, saccades to those targets were quicker for higher $V_{in}$, but showed no dependence on $V_{out}$ (Louie et al. 2011). For the multi-target cued-saccade task, monkeys were faster for the medium total value context than for low or high. However, unlike the cued-saccade version, $V_{out}$ was the only value manipulation in the multi-target cued-saccade task, and RTs did depend on $V_{out}$ (Louie et al. 2011). All the correlations of RTs with value

\footnote{The RT data also speaks to the speed of ramp-up, and suggests slower ramp-up for the four-target version. There is a difference in the RTs for the two versions of the RDMT; the four-target version was slower than the two target version (Churchland et al. 2008).}
conditions were negative, indicating that as the value condition in a particular value context went up, the RTs went down (the monkey responded quicker) (Louie et al. 2011).

Thus, both the Louie et al. results and the Churchland et al. results suggest that higher value conditions, where there is greater reward present in the visual field, result in incursion, despite slower rates of integration in the Churchland et al. data. Greater total value alone could be affecting the threshold. The Louie et al. results are consistent with that interpretation: that the greater total value drives the threshold of the build-up higher. Number of alternative targets, another possible parameter driving threshold differences, is insufficient for the effect, since the Louie et al. results contain conditions with 1 or 2 targets outside the RF under comparable value contrasts (0.33 vs 0.28) that exhibit similar inversions. On balance, then, the build-up is either slower or the same and the threshold may shift. Examination of the neural data itself suggests the latter, at least at the level of individual neurons, as the neural activity rose to a higher threshold in the Louie et al. multi-target cued-saccade task (figure 8, panel b). Similar results were observed in the Churchland et al. data (figure 9, panel a; adapted from Churchland et al. 2008, p. 698), with higher thresholds for the four target trials than two targets on the RDMT.
How might the concatenated combination of a DN and integrate-to-bound mechanism result in both excursion and inversion in LIP cells? An integration of a time-dependent decision signal, implemented by an integrate-to-bound mechanism, together with DN, can explain the initial suppression of firing by total value, inversion, and excursion in LIP. I hypothesize that having more targets drives activity by acting as a contrast gain and response gain modulator, whose strength is modulated by either the probability of choosing a particular target conditional on the motion evidence in the multi-target RDMT, or the probability of initiating a saccade given an instruction cue in the multi-target cued-saccade task under a specific value contrast. This gain term can written as

\[ \gamma = \frac{k\delta(x,n,t)}{V_c} \] (2)

where \( V_c \) is the value contrast, \( k \) is a constant, and \( \delta \) is the decision gain, a function of the target locations (\( x \)), the number of targets (\( n \)), and the time (\( t \)). The response \( r \) of a single LIP neuron at location \( x \) for a stimulus with \( n \) targets is then proportional to

\[ r_{x,n}(t) \propto r_{max} \frac{\int_0^t \gamma dt + V_{in} + \beta}{V_{tot} + \gamma + \sigma} \] (3)

where \( r_{max} \) governs the maximum response, \( t_o \) is the time of onset for the instruction cue in the multi-target cued-saccade task or the onset of motion for the multi-target RDMT.

---

\(^8\) What follows is a sketch of a how-possibly model for the combination of these two dynamical mechanisms that accommodates both excursion and inversion.
and the integral is a function of the target location and the time and the other terms as defined above. A similar model can account for the population activity as well.

In order to accommodate inversion and excursion, the decision gain evolves over time, modulated by the number of targets, $n$. The decision gain is not simply added to the numerator of the DN model; it must also effect the semisaturation term $\sigma$, in order to modulate the height of the integration period, as observed in the neural responses. Importantly, the height of the integration period need not reflect a higher decision threshold, as in the multi-target cued-saccade task, response times were not different between the conditions as might be expected if the threshold itself were modulated (though it was so modulated on the multi-target RDMT). Figure 10, panels A and B show the simulated excursion and inversion, and evolution of $\delta(x,n,t)$ over time using Eq. 3, respectively. This evolution is approximated as a step-function for one location, with $\delta(n,t) = 1$ before the onset of evidence or instruction cue. (The use of a step-function is inessential; in fact, a sigmoid may be more accurate (and perhaps more biologically realistic).) Importantly, the supposition of a decision gain amounts to hypothesized input into the integrate-to-bound mechanism, resulting in the integrative activity prior to saccade onset.
Figure 10: The evolution of LIP activity and the decision field D. A: LIP activity is modulated by the value inside the cell’s RF and total value outside the RF (VIN and VOUT, respectively). For constant VIN and VOUT, LIP activity is suppressed following target onset, but exhibits inversion and greater excursion, both as a function of increasing number of targets. For constant value VIN but variable
VOUT, activity exhibits suppression only. B: Decision field D(n,t) for RF containing a target. The decision field strength rises as a function of the number of targets (n), time-locked to evidence onset. Two downward arrows shown inside the plot correspond to the time of activating decision field (D(n,t) > 1) and deactivating decision field following a decision just prior to saccade onset (D(n,t) = 1), respectively.

At the time of presentation of the targets, δ = 1, and firing is suppressed by larger $V_{tot}$ or smaller $V_{in}$. Upon instruction cue onset, $\lambda$ is still small due to low probability, but $\delta$ rises to a constant value > 1 and proportional to the number of targets, decreasing the effect of semisaturation and hence approximating the proportional response as

$$r_{x,n}(t) \propto r_{max} \frac{V_{in} + \beta}{V_{tot} + \gamma + \frac{V}{\gamma}}.$$  \hspace{1cm} (4)

An increase in $\delta$ results in increased firing and inversion. The probability grows over the course of the instruction epoch (figure 10, panel a), preserving the inverted ordering. Near the time of saccade, however, $\lambda$ approaches one and $r$ approaches saturation such that equation (3) is approximated by

$$r_{x,n}(t) \propto r_{max} \frac{r + V_{in} + \beta}{V_{tot} + \gamma + \frac{V}{\gamma}}.$$  \hspace{1cm} (5)

Thus, a larger number of targets or a lower value contrast results in higher firing, accommodating greater excursion.

While I called this decision-related process the result of changes in a ‘decision gain’, attention, motor planning, or some other process may provide the input to the integration. The combination of an integrate-to-bound mechanism integrating decision
gain input with a DN mechanism accounts for excursions and inversions of activity present in the temporal dynamics prior to choice; larger excursions with more choices; and the absence of these effects when there is no target in the RF of a neuron. The combination of these two mechanisms occurs in a superadditive fashion, as incursion results only if the inputs into the DN mechanism modulate the integrative activity. This simple example illustrates how more complex dynamics in a system can be obtained from the combination of simpler dynamical mechanisms.

Recall that I discussed two objections to the decomposability hypothesis. First, the epistemic objection held that the physical systems underlying cognitive phenomena are too complex for us to learn of or determine the components of these systems. In reply, I argued that we should be open to the attempt to devise components, and that we can use a synthetic strategy to decompose these systems, selecting possible components from a pre-existing theoretically-motivated set. Second, the systemic objection held that the effect of system-level properties on the functioning of components resulted in the extrinsic determination of how those components function, violating the conditions for decomposition listed by Bechtel and Richardson. In reply, I argued that more complex dynamics could arise from the hypothetical implementation and combination of different simple dynamical mechanisms. I illustrated such a claim by discussing the combined implementation of an integrate-to-bound and a divisive normalization
mechanism, in the discussion most recently concluded. This is not a full reply to the systemic objection, but it is suggestive. It suggests that more complex dynamics which are described by power laws, order parameters and so forth at the system level may in fact arise from the implementation of many dynamical mechanisms.

A final note. While Anderson rejects componentiality, he is open to a revised form of componentiality that is very close to my view. He argues that

The only way to hold on to the idea of componentiality in a system like [the brain] is to imagine that the “components” are transiently (and reproducibly) assembled in real time…. We could call the transient reproducible local networks of the brain “components”, but they have functional and temporal properties and interrelations that differ significantly from the kinds of components we normally imagine…. It appears that in the brain we have Transiently Assembled Local Neural Sub-systems (TALoNS). TALoNS—which are the constituents of large-scale functional networks—exhibit temporary, reproducible functional selectivity, but do not have the normal functional characteristics of components…. TALoNS… have… sets of causal properties useful in various circumstances. Which of these properties is most relevant, and contributes most to overall functional outcome, is determined by the totality of those circumstances and the interactions they engender….“ (Anderson 2014, p. 70-72).

Anderson’s TALoNS are certainly in the spirit of my account. My account hypothesizes the existence of dynamical mechanisms that compose cognitive systems. These dynamics can vary and change over time, just like Anderson’s TALoNS, in an on-line fashion. However, Anderson does not substantiate the claim that these are components in some significantly different sense, and I maintain that such systems are decomposable.
3.3 Two Arguments for the Distinctness of Dynamical Mechanisms

So far, I’ve illustrated how dynamical mechanisms execute formal models of processing; how different, more simple dynamical mechanisms could combine to give rise to more complex such mechanisms; analyzed the model execution relation; provided an analysis of the implementation relation; and argued that cognitive systems are decomposable into component mechanisms. However, the second principle above states that the physical and dynamical mechanisms of cognitive systems are distinct, with the physical mechanisms implementing the dynamical ones. The implementation relation defined above clarifies the second component of this principle. The first component of the principle, covering the distinctness of dynamical and physical mechanisms, remains to be discussed, to which I now turn. By claiming these mechanisms are distinct, I am claiming that individual instances of dynamical mechanisms and subsets of the properties of physical mechanisms are type distinct, though they are token identical. What justifies the inference that there are distinct dynamical mechanisms?

The argument by analogy to computer programs states that the inference to dynamical mechanisms distinct from their physical implementations is as justified as the inference to computer programs distinct from their physical implementations. The distinctness of computer programs can be inferred by examining the dynamic activity of
the physical hardware: what the inputs (e.g., keyboard input) cause in the hardware, what outputs the hardware provides (e.g., changes on a screen), and what internal changes occur in the input-state-output sequence. On the basis of these inputs, outputs and states, one can construct a machine table that corresponds to the functioning of the virtual machine that is a program. This virtual machine can be implemented in any physical device that has a causal structure that maps on to the machine table. Since this structure can have many different substrates, multiple implementation is possible, such as occurs when the same program (e.g. Microsoft Word) is run on different types of computer (e.g. a Mac and a PC, or, if you are John Searle, on the wall behind you (Searle 1992)). Despite this multiple implementation, the very same virtual machine is implemented on the different devices.

The argument by analogy suggests that, just as in the case of programs, dynamical mechanisms present in cognitive systems can be multiply implemented and the distinctness of the dynamical mechanism can be inferred from the existence of the same suite of dynamical properties in the diverse physical substrates of the dynamical mechanism. Recall above the discussion of the integrate-to-bound mechanism: it is implemented in different areas of the brain using diverse physiological mechanisms,

---

9 This implementation relation, surprisingly, has yet to be satisfactorily described, though many attempts have been made (e.g. Chalmers 1994). Until a formal analysis is provided, this notion may or may not be the same as the one invoked in describing the relationship between the physical mechanisms and dynamical mechanisms of cognition.
including a smooth increase in firing rates in LIP neurons and an increase in the
saltatory activation of ACC neurons for the integrative activity of the mechanism. The
dynamical properties of these different implementations are the same, despite their
diverse physical implementations. Just as in the case of computer programs, the
diversity of implementations conceals a commonality across them, the set of dynamical
properties all the implementations have in common. This commonality reflects the
implementation of the same program in different computers. Likewise, by analogy, the
commonality observed in physical mechanism’s dynamical properties reflects the
implementation of the same dynamical mechanism by different physical mechanisms.
The argument by analogy holds that the inference to the distinctness of the same
dynamical mechanism is as justified as the inference to the same program.

The first problem with this argument is that the metaphysical status of
dynamical mechanisms mirrors the status of computer programs. But the analogical
conclusion is consistent with the denial of any realism about computer programs. Why
should we believe in the distinctness of computer programs, apart from their
implementation in actual computers? There are similarities, it is clear, between two
instances of a program, such as a description of the structure of the implementing
devices. However, this does not *eo ipso* entail the distinctness of the computer program
and the physical implementing mechanism. This skepticism applies to the inference to
the distinctness of dynamical mechanisms as well: the description of the dynamical structure of the implementing devices does not entail the existence of the dynamical mechanism.

While the issue of the existence claim will figure prominently in a moment, there is a second, much worse problem with the argument by analogy to computer programs. Computer programs, including Turing machines, are all characterized by difference equations. Difference equations, however, are a class of differential equations with discrete time (Strogatz 1994; Giunti 1995). Recall that dynamical systems are described by system evolution functions characterizing how the system components and properties change over time or with respect to each other, and mathematically characterized by the state evolution equations of those same components. Such sets of difference equations look a lot like state evolution equations, and the input-state, state-output machine tables that characterize a given Turing machine look a lot like system evolution functions. In short, it appears that computer programs may in fact be dynamical mechanisms of the sort whose existence is being argued for by analogy to computer programs! Thus, the argument by analogy appears to beg the question.

However, the argument by analogy lays the groundwork for a second argument, the argument from regularities. In the argument by analogy, the widespread implementation in diverse physical devices of the same computer program provided
grounds for the type distinctness of the program from the physical implementations. The argument from regularities makes a similar claim. The argument from regularities states that the dynamical models of the physical substrates of cognitive functions imply the type distinctness of dynamical mechanisms in virtue of the repeated instantiation of the set of dynamical properties as indicated by the use of a dynamical model across physical systems. There is a set of physical mechanisms, each of which executes the same formal model of processing in some cognitive system. For every member of this set, some subset of the dynamical properties of each member executes the function. Each subset can be represented using a dynamical model of some type, such as a computational model or a graphical model. These different models all target the same subset of properties of the physical mechanism. From the use of some model to represent these subsets of the physical dynamics, in short the establishment of a pattern of repeated instances of a set of dynamical properties, the distinctness of the dynamical mechanism possessing those properties may be inferred. The presence of the regularity

---

10 More formally: for some cognitive function C, C is executed by the set of properties p_i of a token instance of a member of a set of types of physical mechanism M = {m_1, m_2, ..., m_n}. But not all such properties p_i are relevant. For all token instances of the elements of the set M, C is executed by a subset π_i of the properties p_i of that token of each physical mechanism type m_i ∈ M. The set π_i is a set of dynamical properties possessed by each token of m_i. For a group of properties p constitutive of a mechanism, if some subset π of p is present across different executions of C, and π satisfies M*, the definition of mechanisms provided earlier in this dissertation, then π constitutes a distinct mechanism. How is π inferred? For a set of dynamical models of type t, D_t, some element d_i of D_t models π. The types of dynamical models range over computational, graphical, linguistic, and other types. Call D the union over i of all d_i, \( \bigcup_i d_i \). Above some cardinality k, the size of \( D \), \( |D| \), establishes a pattern of repeated instances of a set of dynamical properties, viz. π. This last
is established by using some set of models for the dynamics of the physical mechanism, and these models have the same set of dynamical properties as their targets. The regularity consists in repeated instances of the same set of properties that satisfy the definition of a mechanism provided earlier; in short, the mechanism is inferred from the models,\textsuperscript{11, 12}

This argument entails that dynamical mechanisms are formed by taking a subset of the properties of a physical mechanism, specifically a subset of the physical mechanism’s dynamical properties. In particular, the argument from regularities provides a procedure for locating the different dynamical mechanisms: look at the

constraint evolves from a consideration of ‘one off’ models; not every instance of a model is evidence for the existence of a mechanism. Precisely how many models are required for the inference, or how widespread their use, is determined by the community of researchers and the standards of the discipline. The criterion k may not be a particular value, either; a fuzzy boundary may exist. From the specification of a set M, a function C, and the establishment of a pattern of repeated instances of a set of dynamical properties, the existence of a distinct mechanism possessing those properties may be inferred.

\textsuperscript{11} There are other arguments that militate for viewing these dynamic mechanisms differently from both their implementations and the formal models of processing. For example, one could infer that these are distinct mechanisms from differences in the way they fail and what counts as a failure. For a similar sort of inference, see Block 1997.

\textsuperscript{12} Note that the way regularities play a role in this inference is distinct from the role they play in mechanisms. Bogen and Machamer have both questioned the role of regularities in mechanisms (Machamer 2004; Bogen 2005). Their concern is with irregular mechanisms (see also DesAutels 2011 for discussion). For example, vesicle release of neurotransmitter into the synaptic cleft fails most of the time a change in membrane potential sufficient for release arrives at the pre-synaptic bouton, for as yet undetermined reasons (Kandel et al. 2000). However irregular or regular the functioning of a mechanism may be, however, the properties that are used to individuate those mechanisms may regularly reoccur in classes of systems, the relevant point for the argument based on regularities.
various physical mechanisms that serve as the executing substrate of the formal model and locate the overlap in the dynamical properties. Some subset of these dynamical properties constitutes the dynamical mechanism, an Enigma of the mind.13

Expounding and defending this argument will occupy the rest of this chapter. First, I will discuss the role of models in inferring the presence of a mechanism. I begin with a discussion of the role of models in scientific inference. I then discuss one way that such models have been used, to infer the existence of natural kinds (Weiskopf 2011). After criticizing this argument, I turn to analyzing how models provide individuation criteria for identifying mechanisms.

3.4 On Model-based Inferences: the Role of Models in Mechanistic Inference

What is the role of a model in scientific inference?14 Models are representations of entities, and scientific models are representations of scientific entities. If a model is accurate or true, it succeeds in referring to a mechanism that has some subset of the

---

13 Mechanisms must always be differentiated from the huge variety of properties present in any state-of-affairs in the environment (Machamer et al 2000; Craver 2007). Craver calls this “Glennan’s Law” after Glennan who first formulated the thesis (Glennan 1996).

14 Before launching into the argument, this use of ‘model’ is distinct from my use above of ‘formal model’ to describe the formal processing for adaptive behavior. When I mean the latter notion, I will speak of a ‘formal model’ as opposed to just a ‘model’.

172
properties of the model.¹⁵ Models come in many different types, such as computational, mathematical, graphical, physical, linguistic and others (for one typology of models, see Weisberg 2013). Dynamical models are models of the dynamics of physical systems, and can be instances of any of the types of model just mentioned; importantly, dynamical models need not themselves be dynamic, as graphical models of the dynamics of a system, for example, don’t themselves need to move or change over time.¹⁶ Furthermore, scientific models play a variety of roles in scientific reasoning: they suggest research questions; they are appealed to in adjudicating conflict between theories; they help formulate hypotheses; they aid in the discovery, search, and characterization of mechanisms (Darden 2006; Burnston 2014); they play a role in characterizing scientific kinds (Weiskopf 2011); amongst many others. Thus, I take a scientific model to be a type of representation of a scientific entity used in scientific inference. I will argue that, in the case of mechanistic inference, models play a role in providing individuating criteria for mechanisms.¹⁷

²⁵ And, of course, some properties that the model does not have. For advocates of the representational function of models, see Weisberg or Godfrey-Smith, amongst others (Weisberg 2007; Godfrey-Smith 2006).

¹⁶ Most graphical models, perhaps even all, that I’ve seen do only one thing: sit on the page.

¹⁷ This is a distinctly metaphysical role for models. Models are typically discussed in the context of explanation (e.g., see Weisberg 2013).
I will discuss two arguments based on a scientific inference from the widespread use of models. The first example of a scientific inference from the use of models to the distinctness of a set of entities I discuss is an argument from Weiskopf for the kind-hood of a group of functional properties (Weiskopf 2011). Weiskopf argues for grounds for inferring scientific kinds from the widespread use of models. The problems with this argument will helpfully inform the discussion of the argument from regularities, and specifically, what role regularities and models play in the inference. The second example of a scientific inference from the use of models occurs in the argument for the second principle of my theory of cognition. The second principle implies that the physical mechanisms and dynamical mechanisms are type distinct. But why do I infer that there are type distinct physical and dynamical mechanisms? Above, I argued that the repeated use of different types of models to represent the same set of mechanistic properties implies a set of individuating criteria for a mechanism underlying this set of properties. The argument from regularities captures what I take to be the grounds for such inferences; importantly, these are inductive, and not deductive, grounds, and in providing distinct individuation criteria for dynamical and physical mechanisms, the distinctness of the two types of mechanisms is implied.
3.4.1 Inferring Kinds from Models

Recently, Weiskopf has inferred the kind-hood of a set of properties from the use of a model in a wide class of explanations (Weiskopf 2011). Kinds are defined by the existence of regularities: whether a property “ψ is a kind or not depends on whether there is a sufficiently large and interesting body of empirical regularities in which ψ itself is implicated…. Kinds are non-accidental loci of inductive potential” (Weiskopf 2011, p. 235). The inference to kind-hood is licensed by the widespread appearance of a property type in different explanatory models because this widespread occurrence suggests an empirical regularity, and in particular, of one that is due to a mechanism.

The existence of special science kinds turns on the role that putative kinds play in explanatory models. On Weiskopf’s view, “functional categories are kinds when they are appropriate and useful for constructing explanations of how a system comes to exercise particular psychological and behavioral capacities” (Weiskopf 2011, p. 247). In the behavioral sciences, the capacities of creatures are explained “by constructing models of the internal control structures that” produce the capacities, where these models

“typically describe the creature as containing a variety of systems, structures, processes, resources, and sources of activity that interact in a predictable way to produce certain effects under a certain range of conditions. The capacity in question is explained by the presence of these systems operating under certain well-specified conditions…. The interesting functionally defined categories, then, constitute recurrent building blocks of cognitive systems…. [I]nsofar as such functional categories can usefully be applied to modeling cognition, they count as kinds” (Weiskopf 2011, p. 249).
If a model is useful for modeling cognition, where a model is constructed on the basis of the various properties imputed to a mechanism, then the properties used in the model count as kinds. Merely being useful in a model is not sufficient, however; in addition to this pragmatic criterion of the usefulness of a model, relative importance of the kind in the model and widespread applicability of the model are also required:

“On the present account, a class becomes a kind… in virtue of playing a role in a range of models of cognitive and behavioral capacities. Whether K is a kind is a matter of how important K is in these models, plus how widely applicable such models turn out to be (how widely these models can be projected across distinct types of cognizer)” (Weiskopf 2011, p. 250).

The kind must play an important role in the model’s explanation of a cognitive or behavioral capacity. Additionally, the model should apply to many different types of cognitive system. The repeated appearance of a set of important properties in models of the psychological and behavioral capacities of systems implies that the set of properties is a scientific kind.

What type of inference is this? On Weiskopf’s account, “kindhood… is predicated on relational characteristics of functionally grouped entities; roughly, how widely they can be used in our best explanatory models” (Weiskopf 2011, p. 250).

Scientists first construct a set of models of the capacities of cognitive systems, and then examine the models for regularities. These regularities are taken to indicate the presence of functional kinds, kinds that serve particular functions as laid out in the models. In particular, “explanation involves modeling the functional structures that interact to
produce cognition and behavior…. If kinds are fundamentally the real explanatory structures at work in a certain domain, the functional categories posited by well-confirmed models should qualify” as kinds (Weiskopf 2011, p. 252). Thus the inference is derivative of a type of abduction, an inference to the best explanation, as those best explanatory models reveal the regularity underwriting the inference to the kind.

The nature of the inference appears to be: construct a model of some group of phenomena; examine the model for the types of properties referred to in the model; apply the types of properties referred to in that model in many other models; infer existence of the kind of property denoted by the model. This last step has a bit of magical air to it, much as the infamous cartoon of a mathematician’s proof, where “…then a miracle occurs…” appears as a step in the proof. Why should the widespread occurrence of a type of property in a model provide grounds for the kind-hood of the property? The inference hinges on the explanatory success of the models. Still, this is no guarantor of kind-hood, and it does not follow deductively from models that there are any kinds. Furthermore, explanations may have an irreducibly epistemic component, concealed by an ontological veneer (Wright 2012). If the models accurately represented the world, then we could infer the kind-hood of the set of properties denoted by the model. But we don’t know if the models are accurate.
But, the inference is a start. Models are playing an important role, and the widespread use of a set of models gives insight into that role. In the case of cognitive phenomena, some properties of cognitive systems are executing formal models, resulting in the observed behavior. This set of properties is existentially licensed by the reverse inference to a cause of the behavior, and determining the identity of that cause occurs through the use of models. I argued above that we can infer the distinctness of a dynamical mechanism from the widespread utilization and success of such dynamical models. But why should the widespread utilization and success of such dynamical models guarantee the distinctness of such mechanisms? I contend that the widespread use of such models serves to establish a set of individuation criteria for determining the identity of the entity underlying cognitive phenomena. These individuation criteria are distinct from those for physical mechanisms, and so dynamical mechanisms and physical mechanisms are type distinct.

3.4.2 Inferring Mechanisms from Models

What is the nature of the argument from regularities? The argument from regularities grounds the inference to the distinctness of mechanisms in the repeated use of different types of models in reasoning about and investigating some set of phenomena. A diversity of models has been used to investigate the role of integrate-to-bound mechanisms in cognitive phenomena, such as perceptual decision-making in the
RDMT or foraging decisions in the patch-leaving task. Examining how models are used will allow me to substantiate the fundamental role for models in the inference to dynamical mechanisms: as providing individuation criteria for type-identifying the mechanisms that execute the formal models of cognitive processes.

3.4.2.1 Mechanism Bootstrapping

Scientists use parts of the modeled mechanism to infer other parts, be it the component entities or component activities (cf. Kauffman 1970 on articulation of parts explanations). They reason this way using various types of models of mechanisms. For example, Hayden et al. used a graphical model in investigating the integrate-to-bound mechanism implemented by ACC cells during the foraging task. Importantly, the graphical model cites only the dynamical properties of the mechanism, not any of the physical properties. Thus, they investigated whether the threshold changes, the baseline changes, or the integration is faster. This ignores the physical mechanism and focuses only on the entities and activities of the dynamical mechanism.
Figure 11: Graphical model illustrating different ways to modulate patch leave times by modulating dynamical components in an integrate-to-bound mechanism. Adapted from Hayden et al. 2011, p. 937.

In the patch-leaving task, the initial evidence of an increasingly larger series of transient activations was suggestive of an integrate-to-bound mechanism. To investigate whether the other components of the integrate-to-bound were present, such as a threshold or a reset, Hayden and colleagues used a graphical model to reason about the system (figure 11; Hayden et al. 2011, p. 937). To modulate decisions to leave a patch, the rate of rise to the threshold can change, the peak activation can be changed, or the baseline—the starting point for the integrative activity—can be changed (figure 11, panel a). Of these three possibilities, two were observed: faster rates of rise were observed for shorter travel times (figure 11, panel b and panel c) and the firing rate on leave patch trials rose with longer travel times (figure 11, panel d and panel e). These changes in the
firing rates correlated with the observed behavior. Recall from the presentation of this case study that longer travel times resulted in later patch exit times. These patch exit times were apparently modulated by both changes in the rate of integration and in the height of the boundary. This graphical model was used to formulate hypotheses about possible mechanisms for threshold modulation, corresponding to different patch leave times. I call this mechanism bootstrapping: using a model of a dynamical mechanism to investigate hypothetical parts of the target mechanism.

In deploying the graphical model, the neurophysiological details about how each mechanism entity or activity are ignored. Rather, the graphical model depicts only the dynamics that are determined by each way of modulating the thresholding of the mechanism, when the mechanism’s activation point reaches the boundary. A faster rate of rise results in a steeper slope for sooner boundary crossings, and a lower boundary results in integrative activity thresholding sooner for sooner boundary crossings. Neither of these dynamical possibilities refers to the neurophysiological details about how they are implemented. Furthermore, the examination of the data used to confirm these possibilities proceeded by looking at the changes in the firing rate of the recorded cells, an aggregate measure that ignores spike times, membrane potentials, and other
neurophysiological details, though these details are available to the researchers.\textsuperscript{18} The biophysiological detail is passed over in search of the dynamical details to confirm the dynamical model predictions.

3.4.2.2 Mechanism Prediction

In the case of the RDMT, tokens of many different model types have been used to investigate the integrate-to-bound mechanism, both at the behavioral and neural levels. For example, Gold and Shadlen have investigated the behavioral correlates of the integrate-to-bound model for perceptual decisions using mathematical and computational modeling (Gold and Shadlen 2002).\textsuperscript{19} They examined the effect of

\textsuperscript{18} This is a methodological point about the nature of the neural recordings made by Hayden and colleagues (and holds for many similar studies). These recordings involve an electrode situated in the neural area of study and recording the local electrical activity in the area around the recording contact on the electrode. These recordings result in a time series consisting of a series of local voltage measurements, typically around 40,000 such measurements per second, downsampled by computer software, with voltage waveforms reconstructed from the downsampled measurements. These waveforms are sorted to separate out individual action potentials, larger changes in voltage over very small periods of time that are taken to correspond to one type of electrical signals transmitted by neurons. The data recoverable from these recordings goes beyond computing the activation of a cell relative to some environmental or behavioral event. Specifically, the timings of individual action potentials are recorded (and often analyzed as well). Sometimes the local changes in voltage are also analyzed, as the voltages are also recorded. Finally, independently of the recorded action potentials, local field potentials are also recorded, variations in the local background voltage, and these serve as another source of data. All three of these alternative sources of data can serve both as additional dynamical properties of the physical mechanisms, by looking at how these properties change over time or with respect to each other, and as an additional source of information about the implementational substrate, in virtue of picking out spatiotemporal (or, generally, physical) details about the physical mechanism, such as specific voltages or timings of action potentials.

\textsuperscript{19} They also investigate the neural correlates of the model.
different threshold values on reward rates and accuracy. The probability of correct choice $p_i$ was computed using the following formula (Gold and Shadlen 2002, p. 303):

$$p_i = \frac{1}{1 + e^{-2B|\mu_i|}}$$

for the height of the barrier $B$, and the $i^{th}$ drift rate or degree of coherence scale factor $\mu_i$.

Different scalings by drift rate correspond to weights on the barrier, accounting for differences in the probability of correct choice for different levels of coherence. As the barrier value increases, so too does the probability of correct choice (figure 12, panel A).

![Figure 12](image)

**Figure 12**: Panel A: As the barrier value increases along the x-axis, the percentage of correct choices increases as well. Panel B: As the motion strength increases along the x-axis, so too does the probability of choosing correctly. Adapted from Gold and Shadlen 2002, p. 304.

Likewise, as the motion strength, the percent coherence of moving dots, increases, so does the probability of correct choice (figure 12, panel B). This matches the behavior observed during a response-time version of the task (figure 13, dark solid line; Roitman
and Shadlen 2002, p. 9479), where greater motion coherence (x-axis) was associated with
better performance. The computational and mathematical modeling was used to

![Graph showing the relationship between the percentage of coherently moving dots and the probability of choosing correctly.](image)

**Figure 13:** As the percentage of coherently moving dots increases, so too does the probability of choosing correctly. Adapted from Roitman and Shadlen 2002, p. 9479.

investigate how changes in model parameters, such as the boundary or percent coherently moving dots, modulates the behavioral of the model parameters. I call this mechanism prediction, the use of mathematical and computational modeling to determine how the hypothesized mechanism, including dynamical mechanisms, will augment the behavior of the system under different task conditions.

### 3.4.2.3 Mechanism Simulation

At the neuronal level, Mazurek and colleagues investigated the properties of the neuronal mechanism through the use of computational models of the integrative activity of LIP neurons during the RDMT (Mazurek et al. 2003). LIP neurons were modeled as
integrating input from area MT, an area hypothetically encoding motion direction, according to the following equation:

\[
r_{\text{right}}^{\text{LIP}}(t) = r_{\text{baseline}}^{\text{LIP}} + k \int_{\delta_{\text{MT}}}^{t-\delta_{\text{LIP}}} (r_{\text{right}}^{\text{MT}}(\psi) - r_{\text{left}}^{\text{MT}}(\psi)) d\psi
\]

where \( r_{i}^{j} \) is the firing rate for the \( j^{th} \) area in the \( i^{th} \) direction, \( t \) is the time, \( r_{\text{baseline}}^{\text{LIP}} \) is the baseline LIP firing rate, \( k \) is a scaling parameter for the dynamic range of the neuronal response, \( \delta_{\text{MT}} \) is the time from motion onset until a steady state response in the MT cells, \( \delta_{\text{LIP}} \) is the delay time between LIP and MT, and \( \Psi \) are the rate-determining parameters for MT cells, including time \( t \) and percent motion coherence \( C \) (adapted from Mazurek et al. 2003, p. S2). Integration ended once a threshold \( \theta \) had been reached. This mathematical model was used in a computer simulation of the neuronal data (figure 14; Mazurek et al. 2003, p.1261). Importantly, the computer simulations computed the activation of LIP neurons from the simulated coded input of MT cells, utilizing a simple integration function. The dirty details of the neurophysiology, such as membrane potentials, ion concentrations, synaptic weights, and so forth, were left out. Furthermore, the firing rates over time of the LIP neurons were computed, not individual action potentials or membrane potentials. The computer models targeted some of the dynamical properties of the physical mechanisms modeled, leaving out the biophysical details. I call this mechanism simulation, simulating the dynamical or other
properties of mechanism on a computer or other device to confirm that the hypothesized mechanism can recapitulate the observed properties.

Figure 14: Simulated and recorded LIP firing rates during the RDMT. Adapted from Mazurek et al. 2003, p.1261.

Consistent with the argument from regularities, the scientific investigation of the cognitive mechanisms of behavior invokes a diversity of models of mechanism. Many different models of the integrate-to-bound mechanism have been used to investigate the
mechanisms behind various behaviors, including perceptual decisions and foraging decisions. Similar lessons can be learnt from models of other dynamical mechanisms (some of which are discussed below). However, this diversity of types of models all denote the same dynamical mechanism, and these different token models are used to investigate the putative occurrence of this mechanism. Thus, this set of models is a family of models, all possessing the same target mechanism, that cuts across model types. The target mechanism is determined by which properties the models pick out in representing the mechanism. In the case of dynamical models of dynamical cognitive mechanisms, the models focus on the dynamical properties of their target systems. Furthermore, these models pick out dynamical properties that recur across different physical mechanisms and for different behavioral tasks. The models pick up on, hone in on, or lock on to an underlying regularity that is subsequently used to individuate and classify the mechanism.\textsuperscript{20} Regularities become key to determining the type of an instantiated mechanism.\textsuperscript{21}

\textsuperscript{20} Though not the focus herein, in addition to aiding in the individuation of mechanisms, each token model of the mechanism is used to reason about the underlying mechanism. In the case of the patch-leaving task, the graphical model was utilized to suggest hypotheses about how the threshold, integration, or baseline varied; the particular application of the model to the neural data was determined by the dynamics of the physical properties of the underlying neuronal mechanism, as recorded during the physiological experiments. Darden calls this ‘schema instantiation’: when a model of a mechanism, at some level of abstraction, is used to guide research about a mechanism underlying some phenomenon (Darden 2002). In the case of the RDMT, the mathematical and computational models were also utilized to assess the behavioral and neural evidence. However, the neural mechanisms of the foraging decision are distinct from those of the perceptual decision. Thus, a diversity of physical mechanisms exist, but that share a set of
3.5 Models, Mechanisms and Cognition

The evidence that I have cited for the existence of dynamical mechanisms of cognition consists in models of the dynamic activity of neural populations and electrophysiological recordings of neural activity. Recall above that the integrate-to-bound mechanism is most simply mathematically modeled as a first-order differential equation, corresponding to an exponentiation (Wang 2002; Carandini and Heeger 2012). Graphical and computational models have also been used to investigate the mechanism, as detailed above (e.g. Hayden et al. 2011; Mazurek et al. 2003; Usher and McClelland 2001), in addition to other types of models. The use of such models gives rise to two classes of objection to the argument from regularities. First, it might be objected that dynamical properties that make up a dynamical mechanism, and the models are used to both individuate and to suggest hypotheses about those underlying dynamical mechanisms.

21 This role for models in providing individuation criteria would seem to provide grounds for responding to a concern raised by Craver about classifying mechanisms (Craver 2009). In a discussion of natural kinds and mechanisms, Craver argues that “scientists describe mechanisms at different degrees of abstraction…. [W]ether two mechanisms are mechanisms of the same kind depends upon which grain of abstraction one chooses to describe them. If there is no objectively appropriate degree of abstraction for typing mechanisms, then judgments about whether two mechanisms are mechanisms of the same kind rely ineliminably on judgments by people (in concert) about the appropriate degree of abstraction required for the problem at hand. By extension, judgments about how to… refine a scientific taxonomy rely ineliminably on judgments by people (in concert) about the appropriate scope and precision required to predict, explain, and/or control for a particular purpose” (Craver 2009, p. 589). However, objectively appropriate degrees of abstraction can be located on the basis of the locus of a regularity in a set of token mechanisms, as indicated by the widespread successful use of different types of models in investigating those mechanisms.
these models target the physical mechanisms. Second, it might be objected that the
models used cannot support the proposed role in scientific inference attributed to them,
because these models are in fact mechanism sketches, mechanism schemas, mechanism
abstractions or mechanism idealizations. The way I’ve inferred the existence of
dynamical mechanism turns on the use of dynamical models, and both classes of
objection aim to show that dynamical models do not, or cannot, target dynamical
mechanisms.

3.5.1 Dynamical Models and Physical Mechanisms

Recall that dynamical mechanisms are subsets of the properties of physical
mechanisms. As a result, physical mechanisms can account for the observed behavioral
and biophysiological phenomena, including the dynamics, such as integrative activity in
LIP. It might be objected, then, that the models used in reasoning about these systems
target the physical mechanisms, not the dynamical ones. If this were so, then more
biophysically plausible models would supersede models that were less so, in virtue of
including more biophysical details.\(^{22}\) Whether the models have dynamical mechanisms
or physical mechanisms as their targets will depend on the reasoning and experimental

\(^{22}\) Unless, of course, the models were taken to be mechanism sketches or mechanisms schemata. Both of
these possibilities are discussed below.
practices of the scientists. I will now argue that the models in fact target the dynamical mechanisms. Note, though, that dynamical models could be used both as models for the physical mechanism and for the dynamical mechanism. I don’t deny this flexibility in the use of a model. My argument is meant to demonstrate that scientists are using these models to target the dynamical mechanism, not that they are only using these models in this way.

The objection arises from a consideration of the way that scientists are using models in investigating cognitive phenomena. The objection contends that as a matter of scientific practice and theory development in cognitive neuroscience, more biophysically plausible models compete with and are possibly better than ones that are less so. This in turn is taken to imply that the targets for those models are the physical mechanisms.

For example, consider the role of models of LIP activity during the RDMT. One prima facie biophysically plausible model for activation in LIP during perceptual decisions under noise is X.-J. Wang’s synaptic reverberation model (Wang 2002). In a series of articles, XJ Wang and colleagues have developed a neurophysiological mechanism-based explanation of the neurophysiological and behavioral data recorded during the RDMT (Wang 2002, 2008, 2012; Wong and Wang 2006; Lo and Wang 2006; Wong et al. 2007; Liu and Wang 2008). The account describes a mechanism based on NMDA-receptor mediated synaptic reverberation with a winner-take-all competitive
interaction modulated by feedback inhibition (the ‘synaptic reverberation’ mechanism; figure 15, from Wong and Wang 2006, p. 1316).²³ Wang talks of investigating whether the model is “a true neural integrator”, describing the model’s performance in certain conditions as “pure integration where evidence for A is added and evidence for B is subtracted in neural group A…” (Wang 2002, p. 961-962). Furthermore, when “the population activity…has reached a threshold, the network settles in an attractor state…” and a decision is reached (Wang 2002, p. 962). The model then shows “… the two steps of a decision-making process…: initial integration of inputs by ramping neural activity and categorical decision choice by the attractor dynamics” (Wang 2002, p. 961-962). The model of this mechanism exhibits many of the physiological characteristics observed in LIP neurons during the RDMT, including a slow ramp-up in activity that is steeper with greater motion coherence, divergence between the two populations permitting a binary

²³ For the cognoscenti: the model consists in two groups of simulated neurons, each representing a possible direction of motion, with endogenous Poisson spiking characteristics, excitatory connections within themselves, and inhibitory connections to each other. The two groups receive a noisy input signal representing the stimulus (the random dot motion). These recurrent connection dynamics ensure that even in noisy environments (like low motion coherence), one group will ‘win’ by showing persistent elevated levels of activity and suppression of the other group.
choice to form, variants that incorporate modulation of a decision threshold, and the triggering of choice by threshold crossing.

Figure 15: Diagram illustrating the synaptic reverberation dynamical mechanism. Adapted from Wong and Wang 2006, p. 1316.

I will now argue that Wang’s model targets a dynamical mechanism. Wang’s model is biophysically plausible, possibly targeting the physical mechanism and standing as a prima facie competitor of the integrate-to-bound model for that mechanism. However, it also targets a dynamical mechanism that stands as a competitor to the integrate-to-bound dynamical mechanism. Upon closer inspection, the inferred mechanism is in fact another example of a dynamical mechanism, proposed as playing an implementing role in a number of cognitive functions.

---

24 For a nice experiment demonstrating that LIP neurons can categorize stimuli into two classes, suggesting that perhaps this categorization is the correct formal model implemented in LIP, see (Freedman and Assad 2006).

25 The mechanism also accounts for the differences in firing rates between correct and error trials, and for the behavior of the monkeys in the task, replicating for example the linear relationship between decision time and the log of the percent coherence of motion. The model, published in 2002, even predicts the fall off in performance for long stimulus durations as Kiani et al. confirmed in 2008.
The objection maintains that the dynamical models cited in the argument from regularities for individuating dynamical mechanisms are, in fact, models of the physical mechanism. On the basis of what counts as a competing possibly better model, specifically one that includes more physical detail, the objector infers that the target of the models are the physical mechanism. On one interpretation, the model is a good candidate for describing a biophysically realistic mechanism for decisions under noisy perceptual conditions. This view of Wang’s model makes his proposal somewhat

---

26 Is Wang’s model a competitor to the formal model of the behavior, the DDM? No. His model should be seen as proposing an implementing biophysiological mechanism for the DDM, not as a competitor to it. Wang describes the model in terms of integrating evidence until a bound and the resultant categorical choice due to attractor dynamics. Thus, Wang’s synaptic reverberation model based explanation of the RDMT (and by extension similar tasks) can be seen as implementing the DDM formalism. Wang himself sees his proposed mechanism as answering a particular biophysical challenge: “[s]low temporal integration and categorical choice are two indispensable aspects of a decision process, which seem to be difficult to be accomplished by a single network…. [N]eural correlates of the both processes are usually observed in the same cortical area”, which presents the specific biophysical problem of constructing a mechanism that possess both features (Wang 2002, p. 964).

Wang’s model could become a true competitor if it were to establish a formal model that accounts for more of the behavior than the DDM. Nothing in my account prevents this form occurring. In that case, a different formal model would have to be proposed which accounts for the choice behavior. In fact, Wang’s analysis often sounds like he is developing a dynamical system that is meant to replace the DDM as a processing model. I leave a discussion of that project—which, in my opinion, is potentially very fruitful—for another place. Note, though, that as far as optimality is concerned, any proposal for the formal model of processing non-equivalent to the DDM would be suboptimal.

In support of this, the synaptic reverberation model makes several novel predictions, borne out in subjects’ behavior on the RDMT. The synaptic reverberation model predicts different behavior than neural implementations of the diffusion model (Wang 2008, p. 222). The problematic predictions that result from applications of the DDM to behavioral data include a failure to predict longer response times in error trials than correct, unbounded performance with longer duration viewing (as falsified by Kiani et al. 2008), and an inability to accurately account for the effect of additional evidence at different time points during a trial, all aspects of the behavior captured by the synaptic reverberation model. However, the situation is more complicated than this, since humans show right-tailed response time distributions in the task, which can be modeled by the DDM considered as a biophysical mechanism but not by the synaptic reverberation model.
banal: the synaptic reverberation model just targets the physical mechanism that implements some dynamical mechanism. The objector, however, takes this to indicate the target for models of LIP activity are the physical mechanism.

A physical mechanism as defined earlier, consists in an organized collection of spatiotemporally characterized entities and activities. These entities and activities possess various dynamical properties. Some subset of these properties is an instance of a (part of a) particular dynamical mechanism iff the physical mechanism implements the dynamical mechanism. If the target were the physical mechanism, then Wang would focus on those spatiotemporally characterized entities and activities. Does the science reflect this? In particular, does Wang take the model’s target to be the physical mechanism only?

In a 2008 review in the journal Neuron, Wang assesses the synaptic reverberation mechanism and its relation to the integrate-to-bound mechanism. In discussing the difference between the integrate-to-bound mechanism and his alternative, the synaptic

Discussants in this debate occasionally treat the DDM as specifying a type of mechanism (Wang 2002; Usher and McClelland 2001). Seeing the DDM as a proposed mechanism is conceptually distinct from seeing it as the optimal formal model of the behavior. Regardless of the viability of the DDM as directly describing a mechanism, the DDM is provably the optimal formal model for making noisy decisions, and Wang’s own model could be seen as implementing the DDM through the same mapping relation that was invoked earlier for the integrate-to-bound implementation of that same formal model. One literally cannot do better as far as a formal model of the behavior is concerned (Bogacz et al. 2006).

In that article, Wang calls this mechanism the DDM.
reverberation mechanism, Wang notes that, when analyzing the model dynamics, the synaptic reverberation model is “inherently two dimensional” whereas the DDM is one-dimensional (Wang 2008, p. 221). The two hypothetical mechanisms are distinct in virtue of possessing distinct dynamics; the synaptic reverberation mechanism is two-dimensional, as a result of the two hypothetical pools of neurons in the model, whereas the integrate-to-bound mechanism is one-dimensional. This is precisely the level at which to assess the dynamical mechanism: the dimensionality of the state-space arises from the state evolution equations and does not require specification of the implementation of the dynamical mechanism by a physical one.

Models of physical mechanisms also can be described using state-spaces, from their state evolution equations. Wang’s focus on the dynamical differences does not alone suggest that the model is meant to target the dynamical properties of the hypothetical mechanisms. However, the use of the synaptic reverberation mechanism to

---

28 Additionally, the synaptic reverberation model predicts different behavior than neural implementations of the diffusion model (Wang 2008, p. 222). This might prima facie suggest the target is the physical mechanism, except that every dynamical mechanism is token identical to a physical one. Appeals to behavior don’t distinguish between model targets.

29 Seeing as the dimensionality of the state evolution equations for the synaptic reverberation mechanism and the integrate-to-bound mechanism are different, the target is a distinct dynamical mechanism. Matching the dimensionality of the systems is not sufficient, however, for type identity of dynamical mechanism. Recall earlier that dynamical systems, and a fortiori dynamical mechanisms, are characterized by their system evolution functions, the specification of how the system components and properties change over time and with respect to each other. Type identity of dynamical systems requires matching these system evolution functions.
explain a wide range of behavior and processing during many different tasks for
different brain areas, including the model’s initial development to explain persistent
activity in the dorsolateral prefrontal cortex during working memory tasks, militates
against this interpretation (working memory: Wang 1999; Compte et al. 2000; Wang
2001; Brunel and Wang 2001; Miller et al. 2003; Constantinidis and Wang 2004; Wei et al.
sensorimotor mapping: Fusi et al. 2007; Vasilaki et al. 2009; rule-based task switching:
Rigotti et al. 2010; Ardid and Wang 2013). The application of the model to different brain
areas and for different formal models of processing suggests that the target for the
model is a dynamical mechanism.

In reply, the objector could maintain that the same type of physical mechanism
could recur across different neural locations. While there are different types of cells in
different neural areas, the details cited in the Wang model are quite generic. For
example, the buildup in activity in the mechanism is mediated by N-methyl-D-aspartate
(NMDA) receptors, one the main ionotropic glutamatergic receptors in the brain,
implicated in the control of synaptic plasticity and memory (Purves et al. 2008). These
receptors occur in many different brain areas. Similar considerations could apply to
other aspects of the mechanism. And, likewise, the same type of physical mechanism
could be utilized to execute different formal models of processing. The use of the mechanism to execute a range of formal models of cognitive functions, including working memory, executive control, and others, can occur if the model’s target is taken to be a physical mechanism as much as a dynamical one.

3.5.2 Dynamical Mechanisms and Regularities of Cognitive Science

The first objection leveled the charge that the targeted mechanisms were physical, not dynamical. In reply, I argued that one example of an extant neurally plausible model actually reinforced the conclusion that dynamical mechanisms are the targets of these dynamical models. These dynamical mechanisms execute many different formal models and are implemented by physical mechanisms in different areas. The objector, however, argued that the neural details are sufficiently general so as to make the putative implementations in different areas irrelevant to identifying the intended target mechanism for the model. Furthermore, the appeal to many different formal models being executed by the same mechanism is also insufficient to identify a dynamical mechanism as the intended target. Now, I will argue that there are such diverse physical implementations that only a dynamical mechanism can be the intended target.

A striking example provided by the synaptic reverberation mechanism illustrates the dynamical nature of the targeted mechanism in virtue of the diversity of
implementations of the dynamical mechanism. In the original model, during a two alternative forced choice decision, activity from two competing neurons, or two neural populations, will cross-inhibit, or reduce the activity of each other, and give rise to the two-dimensional dynamics of the mechanism. In the synaptic reverberation mechanism, the neural populations encoding each option send excitatory signals to a central inhibitory pool of interneurons, decreasing activity in both populations (figure 15). Recurrent excitatory connections ensure that the reflected cross-inhibition does not eliminate activity while still dampening activity of the competing neural pool. Cross-inhibition is a basic component activity out of which more complicated dynamical mechanisms can be built, such as Wang’s synaptic reverberation mechanism.

Decision-making by collective hive minds utilizes cross-inhibition, dramatically illustrating that the target of the modeling is the dynamical mechanism. Bee swarms must select their nesting site, a process that exhibits many analogies to decision-making mechanisms in the brain, including attention, memory and selective choice (Passino et al. 2008). Bee swarms integrate evidence for the relative desirability of a nest site on the basis of signals from dancing bee scouts, and then implement a thresholding process, similar to that seen in integrate-to-bound or synaptic reverberation mechanisms, to select a site from its competitors (Passino and Seeley 2006). Cross-inhibition is present in the implementation of the thresholding process in bee swarms, as recently reported by
Seeley and colleagues (Seeley et al. 2012). Scout bees will produce a stop signal, consisting of a vibrating signal that is transmitted physically by the sender butting her head against the dancers who communicate information to the swarm (Nieh 1993; Seeley et al. 2012). Dancers ceased dancing soon after receiving a large number of signals, and stop signals were strongly represented at the end of the dance, indicating that dances end because stop signals exceed a threshold (Seeley et al. 2012). When only one choice was present, dancers received significantly less inhibitory signals during the “decision phase” than when two choices were present (Seeley et al. 2012). However, during the “implementation phase”, when the swarm selects the nest site, the difference in stop signal transmission between one and two choice conditions was not significant, indicating that these inhibitory signals play a modulatory role when deciding which nest site to select (Seeley 2010). Different populations of scouts will cross-inhibit each other, in proportion to the quality of the potential nest site, and this cross-inhibition is present only during the decision phase of nest site selection by the swarm. Furthermore, in analyzing the decision process of nest site selection, the dynamical properties of the interactions of the two populations were the focus of the modeling, and in particular the role of cross-inhibition in the decision process.

This remarkable instance of cross-inhibition vividly illustrates that these dynamical mechanisms appear across physical systems. Neural cognitive mechanisms
implement cross-inhibition using a pool of GABA-ergic inhibitory interneurons that inhibit the firing in neural populations. Bee swarms implement cross-inhibition in swarm cognitive mechanisms using direct physical contact between conspecifics that are advocating for different options in the decision process. The physical mechanism of inhibition is markedly different. For neural populations, neurotransmitters make quiescent the neural populations encoding different options. For bee swarms, physical interdiction of conspecifics implements the inhibition. Nonetheless, the same type of dynamical mechanism is being implemented in both cases.30

A similar conclusion arises from considering the integrate-to-bound mechanism. Recall earlier the presentation of the integrate-to-bound mechanism and its diverse neural implementations. The mechanism is hypothetically implemented by neurons in LIP, where the implementation of the mechanism in turn implements a critical mathematical component of the DDM, an integration of evidence to a decision threshold. But we saw that it also is hypothetically implemented in ACC, where the implementation of the mechanism in turn executes a formal model or component

30 Of course, it can be objected that bee swarms are no more cognitive systems than the rock outside my window, the book on my desk or perhaps the computer on which I type. I won’t be concerned with defending swarm cognition herein, though there is much evidence to suggest that swarms produce adaptive behavior that reflects local environmental contexts much like a cognitive system. If swarms are taken to exhibit adaptive behavior and their behavior is driven by physical mechanisms that implement dynamical mechanisms, with those dynamical mechanisms executing formal models that describe the adaptive behavior, then on my theory swarms would have cognitive properties. Providing that full account is beyond the scope of this discussion. The swarm example was meant only as a particularly vivid illustration of dynamical mechanisms for adaptive behavior and how scientists target the dynamical mechanism.
thereof, perhaps the MVT, for determining a critical decision relevant variable, such as foregone reward or a comparison of instantaneous and average reward rates. The integrate-to-bound mechanism can be implemented in different ways, in different places, with different physical mechanisms and at different timescales. The implementation of the same dynamical mechanism by a diverse set of neuronal dynamics also militates for the dynamical mechanism as the target of the modeling.

A separate argument can be made for seeing dynamical mechanisms as the target of the scientists. Since dynamical mechanisms occur repeatedly albeit for different tasks, in different areas, and for executing different functions, the mechanisms appear to capture some mechanistic regularity regarding cognitive phenomena. If the target of the modeling activities were the physical mechanism, then these regularities would be lost, as the set of physical implementing mechanisms is not unified. Insofar as a science is meant to capture the regularities in its target domain, sticking to the physical mechanisms results in missed regularities in cognitive systems and hence an incomplete science.

31 This underscores an important point that harkens back to the heyday of artificial intelligence in the 1970’s. One of the hallmarks of formal systems is that they are medium independent, capable of being implemented in many different physical substrates (Haugeland 1989). The formal models capturing adaptive behavior are medium independent. This medium independence applies not just to the formal model describing the processing that various systems must execute, but also to the dynamical mechanisms implemented by those systems in executing these formal models, as illustrated by the integrate-to-bound and synaptic reverberation mechanisms discussion above.
I have argued that as regards capturing mechanistic regularities in cognitive systems, dynamical mechanisms carry the burden. The objection from physical mechanisms could be pressed, however, by noting that in virtue of the dynamical mechanism being constituted by a subset of the physical mechanism’s properties, the physical mechanism plays the crucial theoretical role. Perhaps, the opponent concedes, there are such dynamical mechanisms; regardless, they aren’t central to our scientific theories of cognitive phenomena.

Keeping only the physical mechanism, however, is insufficient to account for the regularity. Suppose we knew all of the molecular and physiological details about the activation of neurons in LIP, that is, about this particular implementation of the integrate-to-bound mechanism. Without the dynamical mechanism, however, there is no way to capture the apparent underlying mechanistic regularity that exists in the dynamics of the different physical implementations of the integrate-to-bound mechanism. Suppose we remove all of the physical information about the mechanism irrelevant to the cognitive explanation, with the result that we have a sparse though still physical mechanism. Would this description capture the mechanistic regularity in cognitive systems, while lacking the dynamical mechanism? This description would be insufficient for the regularity, because there may be implementations that have no physiology at all, such as computers, or a non-earthly physiology, such as aliens.
Likewise, there may be diverse biophysiological implementations, as implicated above for integrate-to-bound mechanisms and synaptic reverberation mechanisms. Suppose we remove all the physiological specifics in order to attempt to cover this diversity of mechanism. It’s hard to see how the result of that isn’t just the dynamical mechanism. Thus, for dynamical mechanisms, either the regularities escape our science when we retain information regarding the specific physical mechanisms, or if we cover those regularities, we end up with the dynamical mechanism itself. The physical mechanisms underlying the dynamics are unable to describe the mechanistic regularities in cognitive systems, and so dynamical mechanisms are critical for understanding cognitive systems.

I contend on the basis of the foregoing that the target phenomenon for the dynamical models present in cognitive science and cognitive neuroscience, such as Wang’s model, are in fact the dynamical mechanisms implemented by the physical mechanisms. Wang’s model posits a mechanism explaining much of the biophysiological evidence, in a way that preserves the mapping between the DDM and the physiology, by providing a dynamical model. The elements of the DDM guide the construction of the dynamical mechanism by detailing the aspects of the physical mechanism’s dynamics that are relevant to cognitive processing. In turn, these dynamics guide the investigation of the biophysical mechanism, by establishing what aspects of the physiology are relevant to the dynamical mechanism. Other physiological facts
about the components of neural mechanisms and the way they function are also relevant to the physical mechanism, but can be ignored when investigating the execution of the formal model. Wang’s mechanism accounts for important aspects of decision processes by constructing a dynamical mechanism that executes formal models and that is in turn implemented by the neurophysiology, consistent with the theory being outlined herein.32

In sum, in reply to the objection from physical mechanisms, scientific practice and the existence of regularities both favor recognition of a distinct class of dynamical mechanisms. As a matter of scientific practice, cognitive neuroscientists are in the business of constructing models of dynamical mechanisms that can in turn receive physically plausible descriptions of their implementation. And as a matter of describing regularities present in the systems that cognitive science investigates, dynamical models can describe those repeated patterns of mechanistic organization only if they target the dynamical mechanism. Nonetheless, a class of objections remains, maintaining that dynamical mechanisms are just incomplete or idealized models of physical mechanisms.

3.5.3 Mechanism Sketches, Schemas, and Abstractions

Recall that the argument for the distinctness of dynamical mechanisms from physical mechanisms turned on the use of dynamical models to identify repeated

32 For a discussion of another such model, Cisék’s proposal, see Appendix C.
instances of a suite of dynamical properties of physical mechanisms. The objections from models argue that inferring the existence of dynamical mechanisms from dynamic models commits a fundamental blunder: the models are merely mechanism sketches, incomplete models of mechanisms that await further physical details to be complete, or mechanism schemata, relatively complete but abstract mechanisms.

The objection from mechanism sketches proceeds by identifying the dynamic models as incomplete models of physical mechanisms lacking certain details. Mechanism sketches are “[i]ncomplete models—with gaps, question-marks, filler-terms, or hand-waving boxes and arrows…. Mechanism sketches are incomplete because they leave out crucial details about how the mechanism works” (Piccinini and Craver 2011, p. 292). Mechanism sketches play a role in mechanism discovery. As Darden notes,

“…a mechanism sketch cannot (yet) be instantiated. Components are (as yet) unknown. Sketches may have black boxes for missing components whose function is not yet known. They may also have gray boxes, whose functional role… is known or conjectured; however, which specific entities and activities carry out that function in the mechanism are (as yet) unknown. The goal in mechanism discovery is to transform black boxes (components and their functions unknown) to gray boxes (component functions specified) to glass boxes (components supported by good evidence)…” (Darden 2008, p. 966-967).
A mechanism sketch serves as a map to further research, to fill in the missing gaps (or, in line with the metaphor, to make the boxes more transparent). The objection from mechanism sketches maintains that these dynamical models are actually sketches of physical mechanisms, incomplete models merely missing physical detail, reflective of the presently inchoate state of the science regarding the precise nature of the
neurophysiological mechanism. If these dynamical models are nothing more than incomplete sketches of the neural functioning, filling in the neural details results in the relevant mechanism supporting such adaptive behavior, and the model has as its target a physical mechanism. But then, the inference to a distinct class of dynamical mechanism is blocked: the different models used to adduce a regularity in the form of a dynamical mechanism in fact possess distinct targets. In such a case, the dynamical models constitute a functional analysis of the physical mechanism, assessing how the physical mechanisms changes with respect to time or some other variable, and the various components and activities are nothing more than black boxes (or, at best, gray ones) awaiting further research. Such functional analysis directly constrains mechanistic explanation “…if and only if the functional properties described by a functional analysis restrict the range of structural components and component organizations that might exhibit those capacities…” (Piccinini and Craver 2011, p. 14). By restricting the range of components and organizations that exhibit the capacities targeted for explanation, functional analyses are mechanism sketches, to be filled in with details about the mechanism. As Piccinini and Craver insist, “a complete constitutive explanation of a phenomenon in terms of functional properties requires identifying the structures that possess those functional properties—that is, it requires fitting the functional properties within a mechanism” (Piccinini and Craver 2011, p. 8). Functional analysis becomes a
mere step in constructing a representation of a mechanism. In virtue of constituting a
type of functional analysis, dynamical models are only a step in constructing a more
complete model of a physical mechanism, and thus their targets are physical
mechanisms, not dynamical ones.

There are two replies to this objection. First, as it stands, it is an argument
*ignoratio elenchi*: it can both be true that dynamical models constrain the construction of
models of physical mechanisms, and that they are relatively more complete models of
dynamical mechanisms. A model can have many targets, and its use by one scientific
community for a particular purpose does not preclude another use by the same or
different community. And, as I argued above, these models are used to describe
dynamical mechanisms. This does not prevent their being used as well to guide research
into the implementing physical mechanism. In order for the mechanism sketch objection
to be effective, the proponent of the objection needs to hold that models of physical
mechanisms can’t also be models of dynamical mechanisms, for example either because
models of mechanisms don’t have the right intrinsic properties to represent dynamical
mechanisms, or because models can have only one function assigned to them, where
that function is to represent the physical mechanism. But both of those claims are far too
strong: against the first formulation, there is no reason to think that intrinsic
representational limitations prevent modeling dynamical mechanisms, and against the
second formulation, the multifarious uses of models adumbrated above illustrate the many different functions that models can have. Perhaps, the objector maintains, there can only be one representational function for these models, but the science plainly indicates otherwise.

Second, in reply to the objection that such models are mechanism sketches, functional analysis of the type present in the dynamical models do not restrict the range of components and organizations in physical mechanisms that could exhibit the capacities targeted for explanation. Individuating the components and activities possessing the functional properties is part of a complete explanation of a cognitive phenomenon, where the relevant functional properties are the dynamical ones described by the model. But the individuation of components and activities is not the same as determining the relevant functional properties within a physical mechanism. Such a determination merely establishes an input-output isomorphism. For example, the dynamics specified by the model can occur either in virtue of the physical mechanism’s components executing the very same function specified in the dynamical model or by the components executing some other function whose output isomorphic to that function (across some range). The dynamical mechanism can be implemented by the physical mechanism’s components executing any number of different functions so long as the dynamics of the executed function map on to those described by the dynamical
mechanism. This limitation placed on the functional properties need not explicitly restrict what components are included in the mechanism, which activities the components engage in, or how they are organized, so long as the input-output isomorphism exists.

In sum, in reply to the objection that dynamical models are merely mechanism sketches for physical mechanisms, I argue that as it stands, the objection is true but irrelevant. In order for the objection to have force, there needs to be reason for thinking that these dynamical models can only have physical mechanisms as their targets. I argued that no such reason was forthcoming. Furthermore, these models do not play the role that mechanism sketches play, because they do not restrict what physical components or activities are present in the physical mechanism.

An alternative, related objection is that these models are mechanism schemas. A mechanism schema is “a truncated abstract description of a mechanism that can be filled with more specific descriptions of component entities and activities…” (Darden 2002, p. S356). Such schemas “are abstract frameworks for mechanisms. They contain placeholders for the components of the mechanism (both entities and activities) and indicate, with variable degrees of abstraction, how the components are organized. Often these placeholders characterize a component’s role in the mechanism. Discovering a mechanism involves specifying and filling in the details of a schema, that is,
instantiating it by moving to a lower degree of abstraction" (Darden and Craver 2002, p. 4). Schemas also play an important role in mechanism discovery by providing a functional framework for possible component entities and activities to fill. These abstract representations of mechanisms “typically specify roles, black boxes, at varying degrees of abstraction and with more or less detail specified. The schema terms can then be filled with the mechanism’s entities and activities as they are hypothesized and discovered” (Darden and Craver 2002, p. 20). As Darden elaborates,

“[a] less schematic description of a mechanism shows, with more or less detail, how the mechanism operates to produce the phenomenon in a productively continuous way and satisfies the componenty, spatial, temporal, and contextual constraints. A goal in mechanism discovery is to find a description of a mechanism that produces the phenomenon, and for which there is empirical evidence for its various componenty and organizational features. A mechanism schema can be instantiated to yield such an adequate description of a mechanism” (Darden 2008, p. 966).

Mechanism schemas embody the general knowledge of how a mechanism works, detailing how the organization and sequence of these steps understood, with each step in the production of the mechanism’s output outlined. The objection from mechanism schemas maintains that dynamical models are mechanism schemas: abstract descriptions of mechanisms, to be filled in with neuronal descriptions (or, more generally, physical descriptions) of the component entities and activities.

In reply, first note that the same concerns about the relevance of the objection as stated for the case of sketches applies to the case schemas. It could be the case that these models are schemas of physical mechanisms, while also being models of instantiated
dynamical mechanisms. For the objection to have bite, the objector needs to argue that these models are just schemas of physical mechanisms.

Mechanism schemas are abstract models of mechanisms. What does it mean to call a model abstract? According to Thomson-Jones, abstraction first and foremost involves “the omission of a truth” (Thomson-Jones 2005, p. 3). Models involve “abstraction in a particular respect only when it omits some feature of the modeled system without representing the system as lacking that feature” (Thomson-Jones 2005, p. 15). However, not all such representational oversights are abstractions, as relevance and simplicity considerations are key. For Thomson-Jones, “simplification would seem to be an automatic consequence of omission, in that, of two models of a given system, the one which omits mention of a certain feature of the system will thereby be the simpler model, ceteris paribus…. [A]bstractions always contribute to simplicity…. ” (Thomson-Jones 2005, p. 20). Since some omissions make models wildly complex, such as removing inhibitory interactions from models could lead to chaotic behavior, not all omissions are abstractions for that reason as well. But do all abstractions result in simplifications? Do the dynamical mechanisms of cognition, which are here charged as being abstract, simplifying in some sense? Ultimately a simplicity metric would need to be supplied.

______________

33 Thomson-Jones stresses that his definitions of ‘abstraction’ and (below) ‘idealization’ do not provide necessary and sufficient conditions. For the purposes of our discussion, I will treat them as being perhaps more rigorous than he wistfully conspires.
And, for Thomson-Jones, “models do sometimes abstract away from relevant features of the systems they model.... [W]e might choose to stipulate that abstraction should only be spoken of when what is omitted is a relevant feature of the system” if we wish to emphasize the importance of relevance (Thomson-Jones 2005, p. 22). In such a case, the omission must be an omission of a relevant feature, where what’s relevant is determined pragmatically. Thus, a model is abstract iff (i) it omits a feature of the target system; (ii) the omission contributes to the simplicity of the model; and (iii) the omission is of a relevant feature.

In the case of models of dynamical mechanisms, are they abstract in this sense? Abstract models are abstract sensu Thomson-Jones if they omit details of the target system, where the omission is simplifying in some way and concerns a relevant feature. But this objection equivocates on different types of ‘abstraction’. First, the notion of abstraction developed by Thomson-Jones does not seem to be the sense in which dynamical models of dynamical mechanisms are abstract. These models need not omit details of the dynamical mechanisms (though, like all scientific models, some, perhaps even most, presumably do). Perhaps it could be replied that they do omit the physical

\[\text{\textsuperscript{34}}\text{Thomson-Jones is admirably open to abstracting away irrelevant features as well. I have chosen to emphasize the relevance version of the definition of abstraction because I find it more substantive. My points stand if we drop the relevance criterion.}\]
details of the implementation. But, in rebuttal, such details are not relevant to the
dynamical mechanisms represented by the dynamical model.

The diversity of physiological systems that are taken to implement the integrate-
to-bound mechanism constitutes a family of families (or a family of types), and so are
abstract in a different sense than mechanism schemas. Furthermore, dynamical
mechanisms, unlike mechanism schemas, abstractly but fully characterize the dynamics
of the system. This sense of abstraction, then, is different from the sense in which
omission plays a central role, such as in Thomson-Jones’ approach. Abstraction here
means something more like not concrete or not specifying the implementing media. So
in opposition to the objector who maintains that dynamical models purporting to
describe dynamical mechanisms are really just incomplete models of physical
mechanisms, the reality is that the models of these mechanisms may be incomplete as
explanations of the cognitive phenomenon, but the mechanisms themselves are not
incomplete owing to a lack of physical details. In virtue of dynamical mechanisms being
abstract, they need not specify physical mechanisms.

Both of the objections, from sketches and from schemas, hold that these
dynamical models are incomplete models of physical, specifically physiological,
mechanisms. There are two respects that these models are claimed to be incomplete:
details are claimed to be missing in the case of sketches, and the details are abstract in
the case of schemas. But these models are not defective in either of these ways: they are not just relatively incomplete models of physical mechanisms, but rather relatively more complete models of dynamical mechanisms; and they are not just providing abstract roles to be filled in, but rather are abstract in that their physical implementations are left unspecified. Of course, they are incomplete in various respects and degrees, but so is every mechanism, dynamical, physical, or otherwise. Furthermore, they get more complete by adding dynamical mechanistic parts, such as functionally defined entities and activities, not by adding implementational details.

In reply to the objection that dynamical models are mechanism schemas for physical mechanisms, I argue that as in the case of the claim that these models are mechanism sketches, the dynamical models can be both physical mechanism schemas and models of dynamical mechanisms. In addition, mechanism schemas are abstract. But on at least one analysis of abstraction, Thomson-Jones, does not entail that these dynamical models are abstract, schemas of physical mechanisms.

I have addressed two objections against the claim that the targets of dynamical models are dynamical mechanisms in cognitive systems. The dynamical models do not appear to be physical mechanism schemas, as they are made more complete by including more dynamical detail, nor do they serve only as guides to the underlying components and activities that comprise their physical counterparts. They also do not
appear to be mechanism sketches, as they do not constrain how the physical details are filled in beyond the specification of an input-output mapping. Importantly for both objections, dynamical models can represent multiple different kinds of mechanisms. As argued above, the dynamical mechanisms of cognition are often the targets of these modeling practices, especially when considering the patterns of dynamical properties that reoccur across different types of system.

3.6 Conclusion

In sum, the dynamical mechanisms of cognition execute the formal models of processing and are in turn implemented by the physical mechanisms of cognitive systems. The nature of current scientific practice and the existence of mechanistic regularities suggest that the dynamical models scientists are constructing for cognitive systems target dynamical mechanisms, a level of mechanism distinct from the physical implementing substrate. The dynamical models appear to be models of the dynamical mechanisms at work in cognitive systems, the Enigmas of the mind.
4 Cognitive Recycling

The third principle of my philosophical theory of cognition states that the mechanisms of cognition are used over and over to execute different formal models of processing.

Principle 3: The same dynamical mechanisms are repeatedly used to execute different formal models of processing.

This principle concerns the repeated use, or reuse, of the mechanisms of cognition for different processing goals, which are accomplished in virtue of executing different formal models. The concept of reuse occurs in many different theories of cognition (e.g., Hurley 2005, 2008; Dehaene 2005; Gallese and Lakoff 2005; Gallese 2008; Wolpert et al. 2002), though philosophical analysis of the concept of reuse is rare (though see Anderson 2007a, 2007b, 2007c, 2008, 2010, 2014). The reuse hypothesis I will defend herein claims that dynamical mechanisms are reused in cognitive systems to execute different cognitive functions. I will be defining cognitive reuse, and differentiating several different reuse hypotheses, below.

4.1 Criteria for a Theory of Reuse

What are the criteria for a satisfactory theory of reuse? There are two sorts of such criteria, theoretical and philosophical. The theoretical criteria are those that any
particular theory of cognition that features reuse must satisfy in order to justify the reuse in the theory. For example, Anderson lists three such criteria. According to Anderson, “supporters of recycling and redeployment need to provide at least three things: specific models of how information could flow between reused circuits; particular examples of how different configurations of the same parts can result in different computations; and a more complete discussion of how (and when and whether) multiple uses of the same circuit can be coordinated” (Anderson 2010, p. 262; cf. Anderson 2014, p. 78). Specific models of information flow between reused circuits must be an empirically driven research enterprise. As such, the models offered will vary between theories, but all these models may have some notion of reuse in common. Analyzing the concept of reuse is a job for a philosophical theory. Likewise, providing specific examples of reuse is an empirical affair. However, ensuring some putative example is a true instance of reuse will have to be adjudicated against a definition of reuse, a philosophical concern. The coordination of reuse and the control of different circuits in being reused is also a request for a particular feature in some proposed cognitive architecture. Without a philosophical analysis of reuse, however, saying what precisely is being controlled or coordinated becomes unclear.

Philosophical criteria are those that a philosophical theory of reuse must satisfy in order to give an analysis of the concept of reuse. The first philosophical criterion is to
specify what, precisely, is being reused. Are the physical mechanisms of the cognitive system being reused? The dynamical mechanisms? The computations? Are the organizational features, such as the circuits or infrastructure, of the system being reused? And are these different instances of reuse instances of the same concept of reuse? Any particular theory of cognition may have multiple objects of reuse, such as entailing that the physical mechanisms are reused and the representations are reused. Different objects of reuse may have different reuse conditions, the properties of the object that determine what counts as reuse. Thus, different objects of reuse may require stipulation of different notions of reuse.

The second philosophical criterion in analyzing reuse, be it neural, cognitive, or other, is to define what it is for something to be used or put to a use. Use can come in different orders. Some object of use can be used multiple times in a first-order fashion: it can be used to execute many different functions. This is just multi-use. The concern is that instances of reuse will turn out to be instances of the more mundane multi-use (Jungé and Dennett 2010). Multi-use denotes some particular mechanism (representation, infrastructure, domain, etc.) having multiple uses. In order to establish reuse, instead of the weaker multi-use, the “original use of components” must be specified (Jungé and Dennett 2010, p. 278). But an object can also be used in second-order fashion as well, executing the same first-order use for many different functions.
This is just reuse. Use can occur in third-order fashion as well; that is, the reuse of a particular object can itself be reused. This occurs when the same first-order use becomes used for another function, and that reuse in turn becomes used for another function. In a simple extension of the notion, objects can be reused repeatedly at higher and higher orders. The second criterion for a philosophical theory of reuse is to define use in a way that differentiates reuse from multi-use and makes sense of the adicity of use.

Third, a philosophical theory of reuse should keep reuse distinct from redundancy and duplicate parts. Redundancy occurs when there are backup mechanisms for some function served by a mechanism in a system. These backup mechanisms are distinct from the primary mechanism, but they may be of the same type and perform the same function. However, we do not want to count those mechanisms as being reused, as intuitively, a redundant mechanism is not one that is being reused. Duplicate parts cases occur wherein some mechanism is composed of two distinct components, both of which are on their own capable of performing some particular function. In this case, the function is overdetermined for the system. However, again intuitively, duplicate parts overdetermining an outcome should be conceptually distinct from reuse. Keeping these concepts distinct is the third criterion for a philosophical theory of reuse.

1 Note that backups may not actively perform the function F in the system S, but that they do have the capacity to perform F in S.
Finally, a philosophical theory of reuse should be empirically adequate. Given the application of the concept in various theories of cognition, a philosophical account should strive to cover as wide a range of theories as possible. Furthermore, given the accepted examples of reuse in various cognitive phenomena, a philosophical account should strive to accommodate the range of examples. This is not to say that scientists, or cognitive science, or theories of cognition, will always be right about what’s being used, multi-used, or reused. Often a philosophical analysis will result in certain judgments that are counter-intuitive or counter-culture to the accepted or denoted instances of a concept, and so on some philosophical analysis of reuse, some putative instances of reuse may be judged as instances of multi-use. But a philosophical theory of reuse that results in fewer such reclassifications is, ceteris paribus, better than one that results in more. Likewise, a philosophical account should strive to find the most coherent concept of reuse consistent with the science. Again ceteris paribus, a philosophical account that distinguishes a core concept of reuse—or fewer core concepts of reuse—is preferable to one that distinguishes multiple, or more, core concepts of reuse.

In sum, a philosophical analysis of reuse should satisfy four basic criteria. First, the analysis should specify what is being reused. Second, the analysis should define use, and consequently reuse, differentiating reuse from multi-use and accounting for the adicity of use and reuse. Third, the analysis should keep the concept of reuse distinct
from redundancy and from duplicate parts. Fourth, the analysis should be empirically adequate, accommodating the evidence for reuse as best as possible while also unifying the uses of the concept of reuse as best as possible. In the following, I will develop a theory of reuse, starting with a definition of the use of a mechanism for cognition. This definition is perfectly general for determining the uses of mechanisms in cognitive systems. However, as my theory of cognition emphasizes the role of dynamical mechanisms, I develop a notion of reuse that is particular to dynamical mechanisms. This concept of reuse utilizes the definition of use that I develop for the use of a mechanism in a cognitive system.²

4.2 Use and Reuse

I have argued that an adequate theory of reuse for cognitive systems will pick out what is being reused, will differentiate use, reuse, and multi-use as well as accounting for the different orders of reuse, will keep reuse distinct from redundancy and duplication of parts, and will account for the empirical evidence of reuse and the different types of reuse.

For the theory being developed herein, a definition of reuse will also accommodate dynamical mechanisms. I will now develop such a positive account. My

² See Appendix E for an extended discussion and critique of concepts of reuse as they appear in other theories.
account is motivated from the empirical evidence of reuse of dynamical mechanisms, but defines a notion of reuse that is broadly applicable.

The key to understanding reuse is to start with a definition of the use of a mechanism. On my approach to cognition, the component mechanisms of cognitive systems execute formal models. Thus, a mechanism is used by a system when that mechanism either executes or helps to execute a formal model. How does that formal model get executed? The mechanism performs some function F for a system S, such that performing that function totally or in part executes the model M. We can now define use as

(U): A mechanism M is used in a system S iff M is a component of S and M performs a function F in S such that M’s performing F in S, possibly together with other M'-components of S performing functions F', where M' ≠ M and F' ◁= F, execute some formal model of processing P. 3

(U) states that a mechanism is used in a system if, in addition to being a component of the system, the mechanism’s function(s) (help) execute a formal model of processing.

Earlier, model execution was defined as weak equivalence: there is a mapping between the inputs and outputs of the formal model and the properties of the executing system or mechanism. To say that M’s performing F in S helps to execute the model is to say that ascribing a function F to M in S denotes certain properties of M that are weakly

3 ‘◁’ means is possibly identical to and ‘≠’ means is not identical to. The formal model must be a formal model of processing. See Appendix D for more discussion.
equivalent to (parts of) the formal model. Not every property of M, whether M is a
dynamical or physical mechanism, is relevant to executing some formal model; ascribing
the function F picks out which of those properties are relevant.

This definition of use clearly ties the M’s functions to the formal model P
executed in part by M’s function. Using ‘use’ to pick out use in the sense of (U), reuse is
then

(R): The use of mechanism M is an instance of reuse iff M is used and there exists
some mechanism M′ that is also used such that M is not token identical to M′, M
and M′ are of the same type of mechanism, M and M′ perform the same function,
and the formal model P for M ≠ the formal model P′ for M’.  
Critically, note that the variables binding P in (U) are existential, not universal. For
reuse, the formal model P1 executed (in part) by M and the formal model P2 executed (in
part) by M′ must not be identical. In fact, this very flexibility makes the notion of reuse
so appealing: the same thing described mathematically as performing the same function
for some system is utilized to execute different formal models of processing.

The above definitions of use and reuse are explicitly functional, and concern the
causal role that a mechanism plays in a particular system. This is not the cognitive
function of the system, with a corresponding formal model executed by the system’s
components, in response to some processing problem, that is, the formal model of
processing executed, in whole or in part, by the mechanism. Rather, the causal role here

4 The non-identity clause prevents trivial cases of reuse from any instance of use.

223
refers to the mechanistic function of the system’s components, which, on my theory, are dynamical mechanisms.\footnote{This distinction between cognitive function and mechanistic function appears in both the cognitive and biological literature. For example, Bock and von Wahlert (Bock and von Wahlert 1965) distinguish form, the description of the material composition and arrangement of some biological feature, from function, the description of the causal properties of some feature arising from its form, and define the faculty as the combination of the form and function. Different faculties can result from the same form possessing different functions, prima facie outlining instances of reuse.

Anderson and Bergeron draw a similar distinction (Bergeron 2007; Anderson 2010). Bergeron argues that functional specification can proceed in two modes. The first mode regards the specification of a component’s cognitive role, “the function of a particular component… specified relative to a cognitive process, or group of such processes, in which that component is thought to participate” (Bergeron 2007, p. 181, italics in original). The second mode regards the specification of a component’s cognitive working, “the component’s function… specified relative to a cognitive process, or group of such processes, that it is thought to perform” (Bergeron 2007, p. 181, italics removed).

However, unlike my distinction between cognitive function and mechanistic function, which are both meant to be nonteleological, Bergeron assimilates the distinction between roles and workings to the distinction between teleological and nonteleological analysis. A teleological analysis “is one in which the functional aspect of a structure (or process) is specified relative to a certain end product (or goal) that the structure helps to bring about”, while in nonteleological analysis, “the functional aspect of a structure is specified relative to the particular working that is associated with this structure in any given context” (Bergeron 2007, p. 182, italics removed). As applied to cognitive phenomena, teleological analysis reflects cognitive role, and nonteleological analysis reflects cognitive working.

The distinctions between role and working, and teleological and nonteleological, prima facie seem orthogonal. There’s nothing in the concept of a cognitive working, that of a component’s performing a function, that eo ipso conflicts or contradicts with a teleological analysis. In particular, the cognitive working is in part determined by which formal model of processing the system implements, as there may be a range of dynamical properties that are not included when specifying the dynamical mechanism out of the set of dynamical properties of the physical mechanism. Insofar as these mechanisms were selected for in virtue of the role they play in implementing these models, a cognitive working can be given a teleological analysis. Likewise, cognitive roles can be given a nonteleological analysis. In particular, the role of a mechanism in implementing an information processing operation can be specified in terms of its mechanistic contribution to that implementation.

Anderson (Anderson 2010) agrees that there is a distinction between the functions of local circuits, or circumscribed neural areas, and cognitive functions, invoking the language of “workings” and “uses” from Bergeron. Anderson notes “brain regions have fixed low-level functions (“workings”) that are put to many high-level “uses”...” (Anderson 2010, p. 265) and that “[n]eural reuse holds that the “workings” of local neural circuits are put to many different higher-level “uses”, and that the flexibility and variety of our cognitive repertoire results in part from the ability to put together the same parts in different configurations to achieve different behavioral outcomes” (Anderson 2010, p. 295).}
What notion of function is this? One plausible option is Cummins’ concept of causal role functions, where the function F of M in a system S is to have the capacity to F in S. This notion of function, as Cummins eloquently argues, is relative to an “analytical account” of some capacity of the system, that is, relative to an analysis of some one of the system’s capacities in terms of a number of other capacities of the system and their interrelations (Cummins 1975, p. 759ff). As Cummins defines it, “x functions as a ϕ in s (or: the function of x in s is to ϕ) relative to an analytical account A of s’s capacity to ψ just in case x is capable of ϕ-ing in s and A appropriately and adequately accounts for s’s capacity to ψ by, in part, appealing to the capacity of x to ϕ in s” (Cummins 1975, p. 762).

On this understanding of function, capacities are understood as causal powers.

This definition of reuse allows the formulation of a response to an in-principle objection to the distinction between reuse and multi-use from Anderson. The ubiquity of neuromodulation leads Anderson to a novel objection to the philosophical criterion of differentiating reuse from multi-use. He argues that though

“[i]t might seem that the role of neuromodulation should instead lead us to strongly distinguish neural reuse from neural multi-use, reserving “reuse” for the case when a single neural element (neuron or network) is reused in the same state for multiple purposes, and “multi-use” for the case when the element moves into a different functional configuration. But… these situations are somewhat difficult to disentangle even in principle. Because there is reuse at multiple spatial scales, what is reuse at… one level of organization can be multi-use at another. Using the same neurons in a different configuration or when modulated by genetic or chemical factors is reuse of neurons, but multi-use in the local network. Reuse of a local region that cooperates with different partners is reuse of the region, but multi-use at the level of the large-scale network” (Anderson 2014, p. 28, italics in original)
Anderson is pointing out that relative to one level of organization, an object is reused, but relative to another, an object is multi-used. This relativity, for Anderson, stands as a principled objection to asserting that any individual instance of use is reuse or multi-use: from one perspective it may be reuse, but from another, multi-use.

We can now see how reuse occurs for mechanisms and formulate a response to the multi-use objection as stated by Anderson. While it is true that mechanisms can interact with one another, resulting in novel dynamical and other properties, that does not mean that the implementation of multiple mechanisms which result in those novel properties endangers the assignment of use, as defined by playing a role in the execution of a formal model. Properties of more complex mechanisms can result from the combination of more basic mechanisms in ways that are predictable from and explicable by the way those simpler mechanisms are combined or concatenated. What would result in multi-use, as follows from the above definition of use and reuse, would be the use of some other subset of the mechanism’s properties for implementing some other formal model. In that case, the mechanism is the same, even though different aspects of that mechanism play the role of executing the formal model.

Thus, mechanisms can be multi-used on this definition of reuse. If a mechanism performs different functions for executing the same or different formal model, then it is multi-used:
(MU): The use of a mechanism M is an instance of multi-use iff M is used at least twice, and for at least two instances of use of M, the function $F_1$ for one use and the function $F_2$ for another use are such that $F_1 \neq F_2$.

We can now see how multi-use and reuse are distinct for mechanisms. Both turn on the mechanism aiding in the execution of a formal model by performing some function. Precisely which function is performed matters for distinguishing multi-use from reuse. In multi-use, different functions are performed by the mechanism, regardless of whether or not the mechanism is used for executing the same, or for different, formal models. So long as different functions are performed by the dynamical mechanism, then the mechanism is being multi-used *sensu* (MU). In reuse, the mechanism is performing the same function but for executing different formal models. Thus, in response to Anderson’s concerns about the relativity of reuse, there are definitions of reuse that stand independent of the perspective of analysis.

These definitions satisfy the criteria for adequacy for an analysis of reuse adumbrated above. Recall that the first criterion for adequacy required that a theory of reuse specify what precisely is being reused. In the definitions of use and reuse above, mechanisms are being reused, and more specifically below, dynamical mechanisms. Thus the first criterion is satisfied.

Recall that the second criterion for adequacy required that use should be conceptually distinct from multi-use, and that the adicity of use should be accounted for. As just discussed, multi-use and reuse are defined differently on the present theory. As
for the adicity of use, that follows fairly straightforwardly from iterated instances of a mechanism standing in the use relation (U). That is, a mechanism may be used in virtue of performing a function that helps in the execution of a formal model. Executing that particular formal model might be part of executing a more general formal model, corresponding to a more complex cognitive capacity, and so on. Thus, the account also accommodates the adicity of use, and the second criterion is satisfied.

The third criterion required that the concept of reuse should be distinct from redundancy and duplication of parts. Redundancy occurs when a back-up mechanism waits to take over should the original malfunction. Duplication of parts occurs when an active mechanism consists of two or more duplicate parts that perform the same function; these duplicate parts would seem to be instances of reuse for each other. Both concepts are distinct from reuse as defined above. The duplicate parts explicitly perform the same functions for executing the same model. On (R), this is no longer reuse. And redundant parts likewise perform the same function for executing the same model, and so are no longer classified as instances of reuse on (R). Both objections arise from considering different kinds of overdetermination of model execution. In the case of duplicate parts, the functions are performed occurrencely, whereas in the case of redundant parts, the functions are performed successively, and for both, the same
formal models are being executed. But that sameness of formal model execution violates the definition of reuse.

Finally, the fourth criterion requires empirical adequacy for the theory. Empirical adequacy must here be understood relative to the target of the reuse hypothesis. In the case studies above, dynamical mechanisms executed cognitive functions, and so dynamical mechanisms would need to be the entity of interest. To demonstrate this fourth criterion will require a specific formulation of the reuse hypothesis in terms of dynamical mechanisms, to which I now turn.

### 4.3 The Reuse of Dynamical Mechanisms in Cognitive Systems

Having laid out a general approach to use and reuse, I will now apply those definitions to my dynamical mechanism approach. On my view, the relevant entities of reuse are the dynamical mechanisms of cognition. Earlier, I defined the use and reuse of a mechanism in terms of that mechanism performing a function for a system, where performing a function was understood in terms of the possession of causal powers. However, granted that the relevant entity of reuse is a dynamical mechanism, a problem arises from considering the causal powers of such mechanisms.

What is a causal power? We might think of a causal power as the ability for something to stand in the cause-effect relation to another thing, specifically serving in the cause position of that relation. But this leads to a straightforward objection: dynamical
mechanisms can’t stand in causal relations \textit{qua} dynamical mechanism, as dynamical mechanisms are token identical to sets of dynamical properties of physical mechanisms, exclusive of the physical entities whose dynamics are being analyzed. But, the objection runs, only those physical entities can stand in causal relations.\footnote{This may not be true on some analyses of causation, such as Woodward’s interventionist approach (Woodward 2003). I think that, in this case, the objection carries force.}

Depending on how an analysis of causation runs, dynamical mechanisms may or may not be able to stand in causal relations. If they can, then there is no problem. If they cannot, then in reply I contend that this problem arises for just about every function ascription where the entity to which a function is being ascribed is not at the level of the fundamental constituents of nature, supposing such a level exists. Consider the case of a heart, discussed previously. The function of hearts in organisms is to circulate blood. Understood as a causal role function, hearts have the causal power to circulate blood. But this causal power is also possessed by the aggregate organized constituent components of hearts, the ventricles, atria, veins and so forth that compose a heart. Hearts of course are identical to this organized set of entities and activities, but that does not change the facts about causal powers. And similarly, the ventricles, atria, veins and so forth are composed of more fundamental physical entities that will also possess the causal powers of the entities they compose. Causal powers ‘drain away’, to use Block’s
phrase, and the causal powers of hearts (or ion channels or wings or...) will drain away to their fundamental physical constituents.

I contend that dynamical mechanisms are in the same conceptual position as physical mechanisms, like hearts. Just as in the case of hearts, dynamical mechanisms are identical to some set of physical properties, *viz.* dynamical ones. And, just as in the case of hearts, the more fundamental physical substrate possesses the causal powers corresponding to the ascribed function. So, while it may be the case that dynamical mechanisms do not consist in the right sorts of properties to stand in causal relations, even if they do, they are in the same conceptual position as other sorts of systems to which I ascribe functions, like hearts, albeit for different reasons. The problem has not gone away; however, it has been assimilated to a different, more general problem, and I’ll consider that progress enough.

We can now specify the reuse hypothesis as it is relevant to my theory:

(RH): Cognitive systems reuse the same dynamical mechanisms to execute formal models.

---

7 I think that a careful conceptual analysis of the relation between function ascription and causal role ascription might be fruitful here. In particular, the physical mechanism might possess the causal powers, but nevertheless the function is ascribed to the dynamical mechanism. Compare the case of wings and flight: the causal powers might devolve on to the fundamental physical constituents of wings; nonetheless, the function of providing lift (or what-have-you) is ascribed to the wings. The difference, of course, is that wings (or hearts) are physical mechanisms, that is, mechanisms that in part denote physical entities, whereas dynamical mechanisms are not. The intuition is that this difference does not matter for ascribing functions.
(RH) states that dynamical mechanisms are the relevant type of entity being reused.

There are two general types of reuse as applied to dynamical mechanisms, depending on whether the same or different spatial locations are being considered. First, spatially distinct tokens of the same type of dynamical mechanisms can execute the same or different formal models. We can formalize this as

(RH1): Cognitive systems reuse the same dynamical mechanism in different locations at the same or different times.

The two case studies presented earlier, of noisy perceptual decision-making and of strategic decision-making, illustrate (RH1). The same dynamical mechanism, the integrate-to-bound mechanism, was implemented in different areas for different formal models, the RDMT for the perceptual problem and the MVT for the foraging problem.

The first reuse thesis (RH1) states that different tokens of the same type of dynamical mechanism can exist at different locations, that is, the system is spatially mechanistically adaptive. This thesis applies in two different ways. The first application of (RH1) distinguishes tokens of the same type of dynamical mechanism owing to the instantiation of different tokens of the same type of physical mechanism. Since the dynamical mechanism is identified as a subset of the dynamical properties of the physical mechanism, token distinct instances of the same physical mechanism can yield token distinct implementations of the same dynamical mechanism. A very clear example of this is the integrate-to-bound mechanism. Integration of evidence and a concomitant
integration function in neurons has been taken to underlie a variety of perceptual
decisions, including disparity discrimination (Uka and DeAngelis 2003, 2006), olfactory
discrimination (Uchida and Mainen 2003; Kepecs et al. 2006; Uchida et al. 2006), and
vibrotactile perceptual decisions (Romo et al. 2002, 2004; Hernandez et al. 2002),
amongst others. The repeated instances of the integrative activity at the neuronal level
suggest a similar biophysical mechanism and, as a result, the implementation of the
same dynamical mechanism. Of course, there are biophysical differences between
neurons in different areas, so the physical mechanisms are similar but not identical.
These similar mechanisms, though, are able to produce dynamics resulting in the
implementation of the same dynamical mechanisms.

This is perhaps not surprising. The problem facing the organism is similar, even
if the information is coding for a different modality (sound, vision, somatosensation,
etc.). This cognitive functions requires the sampling and integration of evidence, albeit
different kinds of evidence, over time. Is this the same or different formal model
executed? This depends on how finely we individuate our formal models. In the
discussion of the execution of formal models above, I stated that the formal model must
be interpreted. Insofar as such an interpretation assigns different referents to the
encoding variables mathematically describing the integration of evidence, then different
functions are being executed, and the integrate-to-bound mechanism is being reused.
Importantly for the first reuse hypothesis, the mere reoccurrence of a physical mechanism does not entail the implementation of the same dynamical mechanism. Physical mechanisms possess many dynamical properties, and different dynamical mechanisms can be implemented by the same physical mechanism in virtue of the inclusion of different dynamical properties into the dynamical mechanism. Granted this sort of flexibility, simply detailing the physical mechanism doesn’t provide all the details on what, cognitively, the physical mechanism contributes to the system. That contribution has to be determined by looking at the formal model the system is executing and then establishing how that execution occurs via the system’s dynamics. Different formal models can be executed by different dynamical mechanisms, even if the same physical mechanism is present.

The second application of (RH1) distinguishes tokens of the same type of dynamical mechanism implemented by token distinct but type identical physical mechanisms. The case studies of perceptual decision-making under noise and of strategic foraging decisions illustrate this second application of the first reuse hypothesis. In those case studies, different tokens of the same type of dynamical mechanism, the integrate-to-bound mechanism, is implemented by tokens of distinct types of physical mechanisms, a temporally continuous increase in firing rates in LIP during the RDMT and a temporally discontinuous increase in firing rates in ACC during
the patch-leaving task. This represents a qualitative degree of physical difference greater than in instances of the first application of RH1, though there is in reality merely a range of differences in the physical mechanisms, from lesser to great similarity. The integrate-to-bound mechanism has also been seen in prefrontal cortical areas at the level of the neuronal population, with the population exhibiting integration through its state space (Monte et al. 2013). Thus, different physical mechanisms, at the level of the single neuron or neuronal population, implement the same dynamical mechanism. This application of (RH1) occurs in virtue of a different physical substrate.

Second, different collocated tokens of the same type of dynamical mechanism can execute the same or different formal models. This sense of reuse will be illustrated below, but amounts to the same type of dynamical mechanism being implemented at different times for the same or different formal models. The second reuse hypothesis runs

(RH2): Cognitive systems reuse the same dynamical mechanism in the same location at the same or different times. Implementing a dynamical mechanism at different times for the same formal model is not surprising. For example, an example of type identical dynamical mechanism being implemented at different times occurs in the RDMT case study presented previously, as on each successive trial, the neurons in LIP implement an integrate-to-bound mechanism. Similarly, area LIP implements that dynamical mechanism on different
days, for each experimental run. Instead of focusing on the execution of the same formal model by the same dynamical mechanism at different times, I will be focusing on the much more interesting case of type identical dynamical mechanism being implemented for different formal models, at the same or different times. A more elaborate presentation of such a case occurs below.

The second reuse thesis (RH2) refers to the repeated implementation of the same dynamical mechanism, in the same place and possibly by the same physical mechanism but at the same or different times, for executing different formal models. Note that a particular dynamical mechanism might play a role in executing multiple formal models simultaneously. The physical mechanism need not be identical because different physical mechanisms might be collocated, in that they occupy at some appropriately large spatial grain the same spatial location, and both may give rise to the same dynamical mechanism. There may be different physical mechanisms in the same location because what we may individuate different physical mechanisms succeeding each other in the same location while each gives rise the implementation of the same type of dynamical mechanism. For example, if there are two different neuronal processes at work, then these distinct processes may individuate token distinct physical mechanisms. But these two processes might possess identical dynamical properties, and
thus may implement the same dynamical mechanism. In such a case, it is incidental that
the location of the dynamical mechanism is the same across implementations.

An example of the reuse of a dynamical mechanism consistent with (RH2) is in
order. I will present evidence that integrate-to-bound dynamical mechanisms play a
central role in saccade generation, which is described with the use of a different formal
model than that utilized during the RDMT, described previously. Recall that LIP
neurons exhibit a stereotypical increase in their firing rate over the duration of the
presentation of evidence during the RDMT, thresholding at a common value for
different strengths of evidence. LIP neurons generally exhibit a diverse array of
properties in their dynamics; however, one stereotypical pattern of firing commonly
seen is a ramp-up in firing prior to executing a saccade (e.g. Barash et al. 1991). In the
delayed-saccade task, the animal fixates a centrally presented target while a peripheral
target flashes on the screen for a variable amount of time and then extinguishes. Before,
during, and after target presentation, the animal must maintain fixation on the centrally
presented fixation target. The animal is cued to make a saccade by the disappearance of
the central fixation target (the ‘go’ cue). Subsequent to the ‘go’ cue, ramp-up activity,
similar to what is seen during the RDMT, is observed in LIP cells (Figure 16, from
Barash et al. 1991, p. 1100). However, this dynamical mechanism is modulated in the
case of the RDMT, whereas in the delayed-saccade task, there is only a single ramp-up
condition, as there is only a single trial type: the delayed-saccade type. Here, following the “Go” signal, this LIP cell exhibited a stereotypical increase in firing rate prior to the saccade, whose average horizontal and vertical traces can be seen below the activity line for the cell.8

![Activity Line](image)

Figure 16: Pre-saccadic increase in activity in an LIP cell. Adapted from Barash et al. 1991, p. 1100.

The integrative activity prior to a saccade seen in the delayed-saccade task is common for a subset of LIP cells (examples abound; see, e.g., Platt and Glimcher 1997; Gottlieb et al. 1998; Louie et al. 2011; and many others). But what is the processing problem the organisms faces? Fundamentally, the system needs to make a decision

8 Note that for the example cell for this task there is also a transient increase in activity following the presentation of the target.
about when to execute an action, in some particular task context. Different tasks contexts will present different processing demands on the system. In the delayed-saccade task, the system must remember the location of the target and then initiate the movement. The processing demands involve movement initiation. One formal model for eye movement initiation is a rise-to-threshold model, also called the LATER (linear approach to threshold with ergodic rate) model (the model can be applied to other types of movements as well; Carpenter and Williams 1995; Schall and Thompson 1999; Reddi and Carpenter 2000). The LATER model “postulates a decision signal $S$ associated with a particular response. When an appropriate stimulus appears, $S$ starts to rise linearly from an initial level $S_0$ at a rate $r$; upon reaching a pre-specified threshold $S_T$, the saccade is triggered…” (Reddi and Carpenter 2000, p. 827). The integrate-to-bound dynamics observed in LIP for many different tasks can be taken to execute the LATER model for action initiation.

The integrate-to-bound mechanism is reused for the RDMT and the delayed-saccade task. In the RDMT, the integrate-to-bound dynamics in LIP are taken to execute the DDM. In the delayed-saccade task, the integrate-to-bound dynamics in LIP are taken to execute the LATER model. A basic objection, however, argues that these are not in fact different information-processing models. This is simply false. The integration period in the DDM has a clear evidential interpretation that is absent in the integration period...
in the LATER model; as discussed previously, the processing model is determined by the formal mathematical model as well as the interpretation of the variables, which external or internal variables are being encoded. There is no evidential integration period in the processing problem facing the animal in the cued-saccade and delayed-saccade paradigms. Thus, the formal models executed to meet the processing demands on the system are different models.

Regardless of the identity of the physical mechanisms that implement the dynamical mechanisms, type identical dynamical mechanism may execute different formal models at different times. The formal model being executed is determined by the processing demands facing the organism: what behaviors need to be produced given the ethological challenges confronting the organism. Different formal models can be executed using the same machinery, however. The same patterns of dynamical activity are observed in LIP, for example, as discussed earlier for basic execution of saccades (Barash et al. 1991; Platt and Glimcher 1997; Gottlieb et al. 1998; Louie et al. 2011), but also for categorization (Freedman and Assad 2006) and other processing operations as for perceptual decision-making. Furthermore, these are not instances of the same formal model, just for superficially different task environments, as was argued previously.
4.4 Conclusion

I have argued that an acceptable theory of reuse in cognitive systems will specify the objects of reuse, will conceptually distinguish reuse and multiuse as well as account for the adicity of use, will conceptually distinguish reuse, redundancy, and duplication of parts, and will be empirically adequate. I defined use in terms of a mechanism’s function in a cognitive system in executing a formal model and reuse in terms of the same type of mechanism being used in executing distinct formal models. This analysis of reuse satisfies the first three criteria. I then presented a specific application of this definition of reuse, in terms of dynamical mechanisms, and showed how empirical case studies illustrated different instances of reuse so-defined. Thus, my analysis of use and reuse satisfies all four criteria for an adequate theory of reuse.

To conclude, there are three different types of reuse that occur in my theory (table 1). First, there is cognitive reuse: reuse of the same cognitive function for different domains. Cognitive reuse occurs when the same formal model of processing is utilized for different domains, that is, to accomplish different processing goals for the organism.\(^9\) An interesting taxonomical question arises from such reuse: are the functions, that is, the processing problems, really different between the different uses? Second, there is cognitive recycling: reuse of the same dynamical mechanism for executing different

---

\(^9\) See Appendix D for examples of this sort of reuse.
cognitive functions. Examples of this class include the discussion of the integrate-to-bound mechanism above. Cognitive recycling, the focus of the foregoing, occurs when the same dynamical mechanism, resulting from either the same or distinct physical mechanisms, is used for the same or different processing problem. This amounts to the possibility of recycling the dynamical machinery of cognition for different goals of the system. Finally, there is cognitive composting: reuse of the same physical mechanism for different dynamics or different formal models. Such reuse has been extensively discussed elsewhere, but is evident in several examples that have been discussed above, such as in area LIP for both evidence integration and movement generation.

**Table 1: Cognitive Reuse, Recycling, and Composting**

<table>
<thead>
<tr>
<th>Type of Reuse</th>
<th>Same Physical Mechanism?</th>
<th>Same Dynamical Mechanism?</th>
<th>Same Formal Model?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional - “Reuse”</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Cognitive - “Recycle”</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Physical - “Compost”</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
5 The Dynamics of Minds

In this dissertation, I have developed a componential dynamical mechanisms theory of cognition. This theory is empirically motivated, but addresses several philosophical concerns related to cognitive systems, analyzing the concepts of implementation, mechanism, and reuse. I argued for this theory by arguing for three distinct theses:

1. Cognitive systems’ dynamical mechanisms execute formal models of processing.
2. Cognitive systems’ dynamical mechanisms are distinct from and implemented by physical mechanisms.
3. Cognitive systems repeatedly use the same dynamical mechanism to execute different formal models of processing.

While these three theses are empirically motivated, taken together they represent only one vision for a componential dynamical systems theory of cognition. One purpose of the dissertation was to illustrate the viability of such an approach, and I hope that, in virtue of laying out these theses and the evidence for them, I have illustrated this viability.

In the first chapter, I argued that there are two types of dynamicist approaches to cognition, componential and systemic. The heart of the difference between the two regards the decomposability of cognitive systems into component dynamical systems. Accepting that such components can be found results in a distinct approach to cognitive systems.
In the second chapter, I empirically motivated the componential approach by presenting evidence from two case studies from electrophysiological cognitive neuroscience. Both the case study of perceptual decision making under noisy condition and the case study of strategic decision making in a foraging task revealed tokens of the same type of component dynamical system. The case studies also motivated the relationship between those components and the formal models of cognitive processing. Components of cognitive system execute those formal models in virtue of a weak, or input-output, equivalence between the component activity and the formal model. These empirical case studies constituted an empirical argument for the first thesis. In order to make sense of the distinction between the dynamical components and the physical mechanisms, I also defined the implementation relation. Physical mechanisms implement dynamical systems in virtue of the subset of dynamical properties of the physical mechanism being selected for because those dynamical properties are weakly equivalent to a formal model of processing, and that subset of dynamical properties is token identical to the dynamical system under investigation.

Next, I presented a definition of mechanism drawn from recent discussion of mechanisms in scientific explanation. I argued that this definition excludes those component dynamical systems. However, those component dynamical systems of cognitive systems do exhibit many mechanistic properties. On the basis of this
discussion, I developed a more liberal notion of mechanism that includes those component dynamical systems as dynamical mechanisms.

In the third chapter, I argued for the second thesis, that the component dynamical mechanisms of cognitive systems are distinct from though implemented by physical mechanisms. I defended this componentiality claim against two objections to the possibility of decomposition: the epistemic and systemic objections. In reply to the epistemic objection, I noted that while it may be the case that we can not know that cognitive systems possess dynamical mechanisms of components due to the complexity of the dynamics, this does not entail that they are not so composed. Furthermore, I argued that adopting a motivated theoretical stance that indicates possible components potentially helps us in decomposing that complexity. In reply to the systemic objection, I argued that while complex systems may be difficult to understand, the complexity of their dynamics can as a matter of fact arise from the implementation of simpler dynamical mechanisms. To illustrate this point, I proposed a how-possibly model for the simultaneous implementation of tokens of multiple type-distinct dynamical mechanisms that recapitulates observed, more complex dynamics in LIP neurons during decision making tasks.

Next, I discussed the grounds for inferring the type distinctness of dynamical and physical mechanisms. Dynamical models, models of the dynamical properties of
systems, play an important role in inferring the existence of cognitive mechanisms. These models provide the individuation criteria for identifying the mechanism at work during various cognitive tasks. On the basis of the widespread use of dynamical models that target dynamical mechanisms, dynamical and physical mechanisms are distinct types of mechanism.

In the fourth chapter, I argue that dynamical mechanisms are reused to execute different formal models of processing. First, I provided four distinct criteria for a theory of use and reuse: the theory should specify the entities that are used and reused; the theory should distinguish between use and multi-use, as well as account for the adicity of use; the theory should distinguish between reuse, redundancy, and duplication of parts; and the theory should be consistent with the experimental evidence. After defining the concept of use as a system component that performs a function for the system, I defined reuse as the use of a system component for executing distinct formal models. I next developed the definition of reuse into several derivative principles, and illustrated these with several examples drawn from the literature.

The case for this componential dynamical mechanisms theory of cognitive systems was largely implicit in this discussion. Most explicitly, I described how the theory was supported by and consistent with much empirical research. I gestured toward other work in systems biology that is very much in the same spirit as this
approach, suggestive of a unification of these diverse fields. Finally, I hinted that there are repeated patterns of functional organization across cognitive systems with diverse physical compositions, mechanistic regularities that are explained by my theory. All three of these arguments require careful exposition that goes beyond what I have attempted here, and I hope to elaborate on these arguments elsewhere.

The main purpose of this essay has been something like a proof of concept. By constructing the barebones outline of a componential dynamical approach, I hope to have illustrated the viability of the approach, and how the theory has certain strengths in virtue of both the dynamical focus and the componential focus. What is fascinating is that, in the end, we end up with something like a massively complex Enigma machine, where mental processes result from dynamical mechanisms operating according to certain stereotyped principles to produce adaptive behavior.
Appendix A. Two Case Studies in Cognitive Neuroscience

Two case studies in electrophysiological cognitive neuroscience, of perceptual decision-making and strategic decision-making, exemplify the theory of cognition constructed herein. The same dynamical system, called an integrate-to-bound mechanism, is implemented when making perceptual or strategic decisions. The discussion of these case studies will lay the foundation for understanding how neuronal dynamics execute the formal models relevant to adaptive behavior, that is, how the dynamics of neural systems result in cognition.

A.1. Case Study 1: Cognitive Mechanisms of Perceptual Decisions

The first case study concerns the neural mechanisms of perceptual decision-making under noisy conditions, and illustrates the distinction between the formal model of behavior, the dynamical system that is a dynamical mechanism executing the formal model, and the physical mechanism that implements the dynamics. In noisy perceptual decision-making, animals make perceptual decisions with equivocal sensory evidence. Such perceptual decisions are rife, and include deciding if the rustling in the bushes portends a predator, deciding whether or not the approaching conspecific is a friend, deciding if the patch of color hanging from the tree is a ripe and delicious fruit, and similar sorts of decisions. Noisy perceptual decisions are good examples of cognitive
behavior because such decisions exhibit adaptive responses that reflect the environment’s properties. Neuroscientists have developed behavioral tasks designed to probe this cognitive function and have made some revelatory and startling discoveries.

A well-studied noisy perceptual decision-making task in rhesus monkeys (*macaca mulatta*) is the random-dot motion task (RDMT). The RDMT presents to a subject a field of moving dots, and the subject’s task is to determine the correct direction of motion. Two targets appear on a screen followed by a random-dot motion stimulus, consisting of a field of dots, some percentage of which move coherently in the same direction towards one of the two targets (“motion strength”; see Figure 17, panel a; adopted from Gold and Shadlen 2007). The animal selects the target in the direction of perceived motion. Monkeys adapt their responses to the quality of the motion signal, making fewer errors and faster decisions as the motion strength increases (Figure 17, panel b).
Figure 17: The random dot motion task. Adapted from Gold and Shadlen 2007, p. 548.
A mathematical model known as the drift diffusion model (DDM) describes the behavior on the RDMT (Gold and Shadlen 2001, 2002, 2007). In brief, the model has three components: a starting point, an evidence sampling process, and an ending point. The starting point of the decision process is determined by the prior odds of each hypothesis being true. The weight of evidence, the evidence accumulated during perceptual sampling, is recalculated for each piece of evidence and added to the value of the priors to arrive at the current support for a particular hypothesis, the posterior odds. As evidence continues to accrue, this assessment of support for each hypothesis is updated, with the model determining which of the two hypotheses is better supported by the total amount of evidence to date. To make a decision, the model’s evidence

---

10 For a comprehensive yet brief discussion of the model, after which the following explanation is modeled, see (Gold and Shadlen 2002). The DDM is formally equivalent to a statistical decision procedure known as the sequential probability ratio test (Bogacz et al. 2006; Wald and Wolfowitz 1948).

11 More formally: the prior odds are equal to the log of the ratio of the prior probability of hypothesis $h_1$ to the prior probability of hypothesis $h_0$, or

$$\log \frac{P(h_1)}{P(h_0)}$$

This starting point requires the prior probabilities for $h_1$ and $h_0$ to be well defined and non-zero. The weight of the evidence is the log of the ratio of the conditional probabilities of the evidence given each hypothesis, or

$$\text{Weight of evidence} = \log \frac{P(e|h_1)}{P(e|h_0)}$$

Here, ‘e’ stands for the evidence that was collected during the observation process. In the case of the random dot motion paradigm, ‘e’ would stand for some arbitrary time window over which perceptual information was integrated (Shadlen and Newsome 2001). The posterior odds is the log of the ratio of the probabilities of each hypothesis conditional on the evidence:

$$\log \frac{P(h_1|e)}{P(h_0|e)} = \log \frac{P(h_1)}{P(h_0)} + \text{weight of evidence}.$$
accumulation process stops under some termination rule, such as reaching some threshold, upon which the process terminates and a decision is made.\textsuperscript{12} This evidence sampling process is a formal model that describes the variables relevant to the goal of the organism, such as the prior odds of the direction of the dots or the weight of the evidence, and the relations between those variables, such as computing a ratio.

This simple model also aptly describes the observed behavior of the animal. With better evidence in the form of a greater percentage of coherently moving dots, the weight of evidence is larger, and supposing the animal stops accumulating evidence upon crossing some arbitrary boundary, the model predicts faster decision times and hence faster response times. This is precisely what is observed: response times decrease as the percentage of coherently moving dots increases (Figure 17, panel b, bottom plot). The model also predicts that error rates should increase as quality of the evidence degrades, as the noise in our evidence will occasionally support the selection of the alternative hypothesis. Just as predicted, the number of errors increases as the quality of evidence...

\textsuperscript{12} Here are two other possible stopping rules. First, assuming a finite amount of evidence, the process may terminate once all the evidence is in, in which case the sign of the posterior odds indicates which hypothesis is more likely, while its magnitude indicates the how much more likely it is than the alternative. Second, there may be an arbitrary temporal boundary beyond which further evidence is discarded. Upon reaching the temporal boundary, the sign once again indicates which hypothesis is more likely given the evidence and the magnitude indicates the strength of that evidence.
the evidence degrades (Figure 17, panel b, top plot).\textsuperscript{13} The first element in our account is in place: a formal model of the processing that is consistent with the behavior. But how do the animals make this sort of decision? That is, what are the neural mechanisms of perceptual decision-making, such as in the RDMT?

Electrophysiological recordings from neurons in the lateral intraparietal area (LIP) reveal some of the neural mechanisms of such perceptual decisions. The system needs to integrate motion information, encoded in V5/MT, from the noisy visual stimulus and provide that information to eye control centers. LIP is centrally situated to perform such an integrative function, receiving projections from V5/MT (Britten et al. 1992) and sending projections to the frontal eye fields (FEF) and superior colliculus (SC), two brain regions central to the control of eye movements (Ferraina et al. 2002). Furthermore, LIP is implicated in eye movement (saccade) preparation and decision-making (Platt and Glimcher 1999). When presented with noisy motion stimuli such as a field of moving dots, the firing rate of MT cells, as measured by the number of signals (action potentials) transmitted per second, increase in proportion to the coherence of the stimulus and in proportion to the preferred direction of motion for that neuron (Britten et al. 1993; see Figure 17, panel c, left plot inset for illustration). For different motion coherences, MT cells’ activity serves as an index to the strength of the stimulus. Thus,

\textsuperscript{13} The model makes a number of other predictions, including dictating optimal speed-accuracy tradeoffs. See Gold and Shadlen 2002 for more details.
LIP receives a coded motion signal from MT, carrying the information about the
direction and strength of the motion stimulus.

The data from neural recordings in LIP neurons reveals three notable facts
(Figure 17, panel c). First, when the motion is towards a target in the visual receptive
field\textsuperscript{14} of the recorded cell, the increase in the firing rate of the cell is proportional to the
strength of the motion evidence. Second, for those trials, the firing rates for different
motion conditions converge at a common value about 50 ms prior to a saccade toward
the relevant target. Third, there is a stereotyped dip just prior to the stereotyped increase
in neuronal activity. The stereotyped increase in neural activity in area LIP during
perceptual decisions is proportional to the strength of the evidence, starts at a common
firing rate, and culminates in a common firing rate across different strengths of
evidence.

The presence of this integrate-to-bound pattern of neural activation in LIP maps
onto the DDM.\textsuperscript{15} The three main components of the model, the prior odds, the weight of
the evidence, and the termination rule, are all taken to have neurophysiological

\textsuperscript{14} A neuron’s visual receptive field is the area of the visual field within which a neuronal response can be evoked.

\textsuperscript{15} Technically, the DDM is an injective homomorphism (a monomorphism) of the activity in LIP. It is not
isomorphic since there are many details about the electrophysiological activity in LIP that do not map onto elements of the DDM. In other words, there is a lot of physiological detail that is irrelevant to the implementation of the DDM, lacking a corollary in the DDM.
correlates in the integrate-to-bound pattern in LIP. First, there is an initial phase, just after motion onset, with a stereotyped ‘dip’ in the firing rate for each motion condition, interpreted as an initial setting of the probabilities or an encoding of the prior odds. Second, there is the evidence accumulation phase, corresponding to the rise in activity, proportional to the noisiness of the evidence. Third, and finally, there is a common firing value reached just prior to the time of decision, indicative of crossing a threshold in the drift diffusion model. In short, LIP plausibly executes the DDM in virtue of implementing an integrate-to-bound process.

**A.2. Case Study 2: Cognitive Mechanisms of Strategic Decisions**

Strategic decisions, decisions related to exploration or exploitation, allocation of behavior or resources across time, and the like, play an essential role in the fitness of animals. In particular, animals face a ubiquitous set of strategic foraging problems in the course of survival, all of which share certain mathematical features that fall under the purview of formal ecology. These foraging problems directly impinge on the fitness of the animal through the quest for alimentary resources, sexual encounters, information intake, or any other of a number of fitness-relevant variables, all of which can be mathematically represented as rates. Maximizing these rates, parsimoniously described by ordinary differential equations, contributes to the animal’s fitness by helping it achieve optimal behavior.
A specific example of a foraging decision is the patch-leaving problem. Granted an environment where resources are scattered in discrete clumps (‘patches’), when should a foraging animal leave a patch to explore a new one? As animals forage in a patch, food items in the patch are consumed, leaving fewer items to choose from, of lower quality and harder to consume, resulting in a decreased patch value. A monkey at a berry bush may begin by eating the ripest, plumpest berries, but as the monkey continues to consume, the berries become fewer and farther between, and the berries that are left are the mushy or half-ripe ones. When should the monkey vacate the berry bush to travel down the road to the fruit tree?

Like the case of perceptual decisions, a formal model describes the processing necessary for such patch-leaving decisions: the marginal value theorem (MVT) (Charnov 1976). The MVT determines the energy intake rate as a function of the value associated with the food item, the handling time for consuming the item, the average travel time between patches, and other environmental variables.\(^{16}\) Maximizing this rate results in a simple decision rule for leaving a patch: exit a patch when the instantaneous energy

\[ E_n = \frac{\sum P_i g_i(T_i) - t E_t}{t + \sum P_i T_i} \]

The numerator is the sum of the probabilities of encountering a patch of type \(i\) \((P_i)\) times the average energy intake from foraging for some time \((g_i(T_i))\) in patches of type \(i\), minus the energy expended traveling the average travel time \(t\). The denominator is the average travel time \(t\) plus the sum of the probabilities of encountering patches of type \(i\) times the average time spent foraging in a patch of type \(i\). Differentiating with respect to time allows one to formulate a simple first-order ordinary differential equation and then compute optimal leave times (Charnov 1976).
intake rate falls below the average rate for the environment. If the current patch is providing worse energy intake than the average across the entire environment, the animal should leave the patch to forage elsewhere. This average intake rate can be estimated from the animal’s experience foraging in its environment or from the animal’s genetic inheritance.

**Figure 18: The simulated patch-leaving task. Adapted from Hayden et al. 2011, p. 934.**

While we might not expect animals to follow the MVT perfectly, a vast array of species from bees to worms to humans surfing the internet obey the theorem in a noisy fashion, implying the execution of the MVT by decision mechanisms in the different species. In order to investigate the neural mechanisms of patch-leaving decisions and how the MVT is executed by the brain, Hayden and colleagues (Hayden et al. 2011) devised a simulacrum of the patch-leaving problem suitable for neural recordings. In the task, the subject is presented with two choices: a small blue rectangle and a large gray
rectangle (Figure 18, adapted from Hayden et al. 2011, p. 934). Upon selection of the small blue target, the target shrinks and a squirt of juice is delivered once it disappears. On the next trial, should the animal select the same target, a smaller squirt of juice will be delivered, and as the animal successively selects the same blue target (‘stay in patch’ choice), smaller and smaller juice rewards are delivered, mimicking patch depletion. At some point, the rewards will decrease sufficiently to motivate selection of the large gray bar (‘leave patch’ choice). Once the gray target is chosen, the animal must wait for the whole bar to shrink to nothing (at the same rate as the blue bar), without reward. Once the next trial starts, the locations of the two targets have switched sides, the reward associated with the stay in patch option has reset to the full value, and a new travel time is selected, as represented by the height of the gray bar.

**Figure 19:** Behavior on the patch-leaving task. Adapted from Hayden et al. 2011, p. 934.
Figure 20: Example neuron from the anterior cingulate cortex, recorded during the patch-leaving task. Adapted from Hayden et al. 2011, p. 935.

Rhesus macaques show a significant if slight effect of travel time on patch residence time (Hayden et al. 2011). On average, animals tended to stay longer in a given patch for longer travel times to a new patch, although the behavior is very noisy (Figure 19, adapted from Hayden et al. 2011, p. 934). This behavior represents a slight deviation from the optimal strategy as determined by the MVT, though the animals appear to treat each patch as though it has been drawn from its own environment. Given that patch residence time varies with travel time to a new patch, Hayden et al. searched for a neural signal that encoded patch residence time and that varied with the travel time.

Recording from the anterior cingulate cortex sulcus (ACCs), a medial prefrontal cortical structure, Hayden and colleagues uncovered a different neurophysiological

---

17 This suboptimal behavior may be due to a noisy implementation of the MVT, or it may suggest that some other model would better capture the behavior. Wang (2012) interestingly suggests that deviations from optimality might result from the neural mechanisms used to implement optimal models.
implementation of an integrate-to-bound system. As the animal forages within a patch over the course of many trials, these increases in patch residence time are encoded by an increase in the peri-saccadic peak response in ACC neurons (see example neuron in Figure 20, adapted from Hayden et al. 2011, p. 935). The average firing rate over a window centered on the time of choice rose over the course of the patch, indicating the presence of an integrate-to-bound mechanism. The firing rates in those neurons also rose to a common threshold for different patch leave times, reaching crescendo at roughly the same firing rate for different patch exit times. Furthermore, the initial firing rates at the beginning of the patch for a given travel time were the same. All three elements, of baseline, integration, and threshold, present in the LIP implementation of an integrate-to-bound system are also present in the ACC data collected during the foraging task, suggesting the same system is implemented in both regions.18

These two case studies illustrate the repeated dynamical systems the brain implements to execute cognitive functions, in these cases, decisions. Though different formal models characterize the behavior in the foraging and perceptual tasks, and

18 Despite both areas seemingly implementing the same dynamical mechanism, the physiological mechanisms used to do so differ. In the case of LIP and the RDMT, the integrative activity occurred on the timescale of hundreds of milliseconds and was implemented by an increase in firing by individual neurons. This is a subject of some debate, but for the time being let’s suppose the trial-wise activity of individual neurons reflects the average firing rate traces in the plots discussed above. In the case of ACC and the patch foraging task, the integrative activity occurred on the timescale of tens of seconds and was implemented by an increase in the peak of a choice-locked transient increase in activity in individual neurons. Physiologically, it is just not possible to drive these two effects with the use of identical physiological mechanisms.
different processing demands are present for the system, the brain appears to execute these models by implementing the same dynamical system, albeit with different neurophysiological mechanisms. In the case of LIP, the system executes an integration of evidence function. The specific function executed by the dynamical system in ACC remains unclear, though assessments of opportunity cost or a comparison of instantaneous reward intake rates to average reward intake rates are among the possibilities. Both areas implement an integrate-to-bound dynamical system, albeit by different physiological mechanisms. These dynamical systems are repeatedly used in the execution of the formal models of behavior. In the case of the integrate-to-bound mechanism, the dynamics of the system appear to execute the DDM in the case of noisy perceptual decision-making and the MVT in the case of patch-leaving. I contend that these reused systems appear in a wide variety of cases (divisive normalization, linear filtering, synaptic reverberation and center-surround inhibition are other examples) and constitute a fundamental level of functional organization for the system, playing a prominent role in explaining the capacity that cognitive systems have to behave flexibly. The behavior is flexible in two distinct senses: first, that the system’s functioning reflects local environmental conditions, and second, that the same type of system can be utilized for different functions. This flexibility in the use and functioning of dynamical systems supports the adaptive behavior characteristic of cognition.
Appendix B. Markov Structures and Dynamical Systems

Above, systems and their evolution were defined in continuous terms. There is an equivalent definition in discrete terms, which I will present here. We can describe the system with a Markov structure, which lists the states and their connections and the transition probabilities between them. For our example, the Markov structure would connect the state \( \{0, 0\} \) with \( \{1, 0\} \) with some probability \( p_{12} \) and the state \( \{1, 0\} \) with \( \{1, 1\} \) with some probability \( p_{23} \) (and in general, any two states \( i, j \) are connected with probability \( p_{ij} \)).

Recall the light detecting system from the main text above, with two switches, one of which flips at a lower amounts of light in the environment than the other. The probability of transitions between states for our little system would be partly determined by the nature of the light in the system’s environment, whose statistics are currently undescribed. The state \( \{0, 0\} \) and \( \{1, 1\} \) could never be connected, receiving probability \( p_{13} = 0 \), while the probability of any state transitioning into \( \{0, 1\} \) is 0, and likewise there is no probability of moving from \( \{0, 1\} \) to any state. In fact, by the way the situation has been described, there is no way for the system to inhabit \( \{0, 1\} \); this is represented in the Markov structure by the zero probabilities. Furthermore, while without transition probabilities we noted that there are four states for the system to inhabit, once we incorporate the different ways to transition between states, the complexity in our system have vastly increased. Suppose we define the property of
symmetry for such structures as if a system being able to transition from state a to state b implies a transition possible from state b to state a (although perhaps with a different, non-zero probability). The number of transitions for n states is $\sum_{i=1}^{n} 2(i-1)$ without reflexivity (permitting a state to transition to itself) and with sharing, and $n^2$ with reflexivity and with sharing. Without reflexivity, in our little switch system, moving from {0 0} to {1 0} is different from moving from {1 0} to {0 0}, and likewise for every possible transition. Thus our four state system has 12 possible transitions without and 16 possible transitions with reflexivity. We can now define a Markov structure:

A complete Markov structure is a matrix whose dimensionality is the number of states of the system for which it is a structure, and whose entries correspond to the probability of transition from the state corresponding to the x-coordinate of an entry to the state corresponding to the y-coordinate of that entry. If every given transition between states

\[ \text{Markov structure} \equiv \text{a complete Markov structure } M \text{ for some system } \Sigma \text{ is} \]
\[ \begin{align*}
\text{i.} & \quad \text{a set of states } \mathcal{S} \text{ of } \Sigma \text{ and a set of entries } \mathcal{M} \text{ of } M \text{ such that } \forall x \in \mathcal{S} \quad \exists y \in \mathcal{M} \quad I_{xy} \quad \forall z \in \mathcal{M} \quad (z = m \in \mathcal{M} \quad I_{xz} \quad y = z) ; \\
\text{ii.} & \quad \text{for all } s \in \mathcal{S}, \text{ if } s_i \text{ is reachable from } s_j, \text{ then there is some number at the location } (i, j) \text{ of } M \text{ that is the probability of transition from } s_i \text{ to } s_j.
\end{align*} \]

A complete Markov structure is a matrix whose dimensionality is the number of states of the system for which it is a structure, and whose entries correspond to the probability of transition from the state corresponding to the x-coordinate of an entry to the state corresponding to the y-coordinate of that entry. If every given transition between states

\[ \text{There is some vagueness of terms here: do we count each reflexive transition, such as the transition from state a to state a, as one or two transitions? By the property of symmetry, for every transition, it’s mirror transition is also a possible transition. But then, for transitions to state a from state a, there will be another transition to state a from state a. Since these are the same states, we might conclude there is only one transition here; or we can go with the formal answer due to the symmetry property, and count it twice. If we count it twice, then a system of n states will have } n + n^2 \text{ transitions.} \]

263
receives a probability (even a zero probability), then the Markov structure is for that
system is complete; otherwise it is partial.

We’ve seen now that a relatively simple system, consisting of two switches with
input from a photodetector, can show complex state transitions. Certain states of the
system are reachable from other states, and certain states are not; and other states aren’t
able to be occupied by the system at all. We can think of this system geometrically, as
carving out different states in some space. The state space of the system, despite its
simple construction, is rather complex. These simple systems grow exponentially as the
number of switching elements is increased, with the number of states that can be
theoretically occupied (that is, that are possible for the system, independently of
considerations such as blocking off in virtue of the input characteristics established in
the system’s construction, like our light level switches) by the system growing at $2^n$,
where $n$ is the number of switches. The system just described has binary switches that
are either ‘on’ (1) or ‘off’ (0). Imagine, instead, that we have multi-stage switches.
Suppose we start with switches that can be in three positions: -1, 0, or 1. If our system
has only one switch, it can be in any of those 3 states. If it has two switches, it can be in
any one of 9 states. In general, then, the number of states that a system can inhabit will
be $X^n$, where $n$ is the number of switches and $X$ is the number of states that each switch
can inhabit.
In order to capture the central feature of functional circuits, their dynamics, the way the system changes with regard to some variable (usually time), the concept of a state transition function is required. If we go back to our two switch simple system, then the dynamics emerge as a result of the changes in the environment and changes in the system, in particular changes in the amount of light in the environment which forces the system to evolve as the light changes, with the detection of more light flipping more switches. By specifying the states and the probabilities of transition between different states, these dynamics are captured in part by the Markov structure. In dynamical systems theory, this relationship between the input to the system and the system’s state variables is captured by what is known as the phase-plane analysis of the system. Phase-plane analyses plot how the state variables, in this case the state(s) of the switch(es), change as a function of the input to the system. Phase-plane plots can depict the relationship over time between the environment and the states of the system, how the system’s states evolve over time given the input to the system. This evolution of the system over time reveals the importance of the distinction between the states and the state transition functions.

Let’s return to the two switch example to illustrate this distinction. The switches were the basic components of the system, able to occupy one of two states, switching between those states as determined by the photodetector. Thinking of the switches as
being in one of two states and of being able to transition between states with some probability does not specify how the switches switch between the states. The transition might be an instantaneous flip in state, or the switch might smoothly transition between the states. These two possibilities aren’t so far apart; the binary switch is a smooth transition that takes zero time. Suppose the switch starts in state 0 at time 0 and ends in state 1 at time 1. Suppose this transition takes some time, such that time 1 > time 0 (equivalently, time 1 - time 0 > 0). What should we say about the state of the switch at time 0 + dt < time 1? The switch may be in no state, corresponding to something like a binary, discontinuous transition. Or, the switch may be in a state that is in the interval [0 1]; this case is the smooth, continuous transition function, with a state space that is mapped on to the real number line. Or, the system might be somewhat in state 0 and somewhat in state 1. This is the smooth, discontinuous transition function, as the system can be in state 0, state 1, or some mix of the two. These are all different state transition functions, specifying how the system transitions between states. State transition functions allow us to capture the dynamics of the system, by describing how the states of the system transition over time (or over some other variable, such as the input). (Note that the dynamics now result from a temporal sequence of physical states.) We can define a state transition function:

\[
\text{State transition function } F \equiv \text{ for some state } s_i \in \mathcal{S} \text{ and state } s_j \in \mathcal{S},
\]

\[
i. \quad s_j \text{ is reachable from } s_i;
\]
ii. \( F(s_i) = s_j \);

iii. \( F \) defines some intermediating set \( S \) of ordered pairs \( <\alpha, \beta> \) such that:

a. For all elements \( \eta \) of the ordered pairs of \( S \), \( \eta \in \mathcal{S} \);

b. For all \( \alpha, \beta \), \( \beta \) is reachable from \( \alpha \);

c. In \( S \), there is at least one ordered pair with \( s_i \) in the first position and \( s_j \) in the last position.

What this definition of \( F \) says is that \( F \) is a mapping from states of the system to other states of the system that can have any number of intermediating states.

So far, I have defined our little machines in terms of states and transitions between them. However, earlier I stressed the importance of the dynamics of cognition. What gives? There are two critical points here. First, any dynamical system defined in terms of differential equations, as is usually done for dynamical systems, can be captured by the formalism I’ve just briefly outlined. Every function \( F \) of order \( n \) defined over the real number line can be described as an uncountably infinite subset of \( n \)-tuples of the \( n \)th Cartesian product of the real number line on itself, \( \mathbb{R}^n \). Thus a line is some subset \( L \) of ordered pairs \( (x, y) \) where \( L \subset \mathbb{R} \times \mathbb{R} \). Since we can translate every function, including every differential equation, into sets and subsets, nothing is lost by defining a dynamical system in the way I have. Second, defining a dynamical system mathematically as a collection of states mapped onto subsets of the real number line is not the same as defining the relevant mental entities for an analysis of cognition to be static, enduring things. While it may be true that one can take the zero-limit time derivative of a dynamical system to arrive at parameter values and thus a ‘state’, which
will be included in the Markov structure and possibly one or more of the state transition functions for that system, the existence of such a state does not *eo ipso* make it the relevant unit of analysis for cognitive systems. The relevant unit of analysis, rather, is the collection of such states, collections that the definitions above permit to be any arbitrary number of states (including uncountably many, since the definitions range over \( \mathbb{R} \)). In particular, it has to do with relations between sets of states, relations that capture the dynamics of cognition.
Appendix C. A Competitor to the DDM and the Integrate-to-bound Mechanism for Perceptual Decisions

An illustrating contrast can be drawn with another competing model, Paul Cisék’s urgency-gating model (Cisék 2006, 2007; Cisék et al. 2009). Cisék’s model stands as a challenge to both the formal processing model and the dynamical mechanism used to execute it. The DDM describes the cognitive model a system needs to execute for adaptive behavior, whereas Wang’s biophysical model targets the dynamical mechanism that is used to implement the cognitive model. Cisék, in contrast, explicitly targets both the processing model, utilizing a mathematical, Bayesian approach to the decision processes in sensorimotor transforms (that is, the sensation-to-response loop), and the dynamical mechanism that executes the processing model, utilizing a competitive interaction dynamical mechanism (Cisék 2006, 2007; Cisek et al. 2009). The fundamental idea is that decision making and action planning are encoded in the same neural populations as a distribution of possible action plans that have differential activity and compete for execution (figure 21, panel c, from Cisék 2012, p.2). This competitive interaction between the plans constitutes the decision process as well as a specific proposal for the executing mechanism. Cisék calls this the “affordance competition hypothesis” (Cisék, 2007).
Figure 21: Three different processing sequences for selecting an action. Adapted from Cisék 2012, p. 2.

Cisék and colleagues apply this alternate model to the RDMT. They point out that, in the typical version of the task (described above), elapsed time is confounded with stimulus information, because the sensory stimulus changes while time elapses over the course of that presentation. If we hold constant the amount of information present in the stimulus, neural activity “may be the product of current stimulus information and a growing signal related to the urgency of making a choice” (Cisék et al. 2009, p. 11560). By keeping within-trial presentation of choice information constant but varying the amount of choice information presented between trials, the putative urgency signal, or “motor-initiation buildup signal”, may be separated from the stimulus information signal (Cisék et al. 2009, p. 11571). The ramp up in activity can be seen as driving the system to settle on an action in a unified decide-and-act framework, instead
of implicating an evidence integration process. Cisék calls this the urgency-gating model.

One target of Cisék’s model is the cognitive phenomenon of decision making under noisy conditions. Cisék offers an alternative formal model of processing that must be executed for the production of behavior. The focus on the computational problem being solved makes Cisék’s model a competitor to the DDM. Furthermore, Cisék and colleague’s argument is bolstered by experimental human behavioral data, not by electrophysiological recordings in LIP (Cisék et al. 2009). The behavior serves as a probe for the cognitive state of the system, with differences in the behavior establishing further cognitive phenomena that differentiate between the different models. If Cisék’s model were a physical implementation, or even merely a dynamical mechanism execution, the simulation of LIP activity would have been paramount. Instead, they tried to establish evidence for the model by looking at modulations in the behavior of subjects in a variant

---

1 This particular concern with the classical RDMT design meshes with Cisek’s proposal for decision making and action preparation and selection; the DDM tackles the decision making aspect separately from action selection, whereas in Cisek’s model, action selection and decision making are simultaneously accomplished.

2 Cisek et al. also argue against the DDM on the grounds that the model is unable to execute a time-accuracy tradeoff, as must be done in animals who do not have unlimited time to act. The threshold signal in the RT version of the RDMT is higher than in the fixed duration version of the task as well. Why should the threshold change in a task-dependent manner in this way? Cisek’s model accounts for this difference, by allowing the monkeys to tradeoff accuracy and time through differential modulation of by the urgency signal. This complaint amounts to noting an additional element being mapped between Cisek’s model and the observed neural data that is absent in the DDM.

3 Consequently, one could hold that Cisék’s processing model is in fact implemented by either the integrate-to-bound mechanism or the synaptic reverberation mechanism.
of the RDMT, evidence that lies at the cognitive level of explanation. In virtue of this focus on the choice behavior, Cisék’s model serves as a true alternative to the DDM.

Figure 22: Cisék’s competing proposal for a dynamical mechanism for action selection. Adapted from Cisék 2006, p. 9763.

In keeping with the dynamical mechanism approach, Cisek also proposes a dynamical mechanism for executing his affordance competition hypothesis. The dynamical mechanism is composed of a set of units encoding different action plans that are in competitive interaction, inhibiting dissimilar action plans with cross-inhibition and exciting similar action plans with cross-excitation. For example, suppose that the action to be decided on is a reaching motion to a particular elevation and azimuth (figure 22, panel a; adapted from Cisék 2006, p. 9763). A set of elementary units, such as neurons, will exhibit a preference for some combination of elevation and azimuth (figure 22, panel b). This population can encode single actions (figure 22, panel c, top plot) or
multiple actions simultaneously (figure 22, panel c, bottom plot). When there are multiple actions, the different, dissimilar plans inhibit each other, in virtue of cross-inhibitory connections between action plans, and similar plans excite each other, in virtue of cross-excitatory connections (Cisék 2006, p. 9762, their figure 1). The result is a competitive interaction with a winner-take-all (WTA) mechanism, that is, one action plan remains excited while the other plans fade out.

Cisék’s model illustrates both a challenge to the processing model in the RDMT, the urgency-gating model, and to the dynamical mechanism implemented by LIP neurons, the WTA competitive interaction mechanism. The proposed processing model challenges the RDMT because it does not propose an integration of evidence in arriving at a decision. The dynamical mechanism is distinct from the integrate-to-bound mechanism and the synaptic reverberation mechanism because of the novel dynamical components, including entities such as cross-inhibitory connections, cross-excitatory connections, and the use of many units to encode actions, and including activities such as cross-inhibition, cross-excitation and winner-take-all.4

---

4 To be fair, the synaptic reverberation mechanism could be complicated by adding certain components, and then iterated, to arrive at the WTA competitive interaction mechanism. Nonetheless, the addition of these entities and activities, by the definition of a mechanism (M*) above, results in a distinct mechanism.
Appendix D. Cognitive Functions and Cognitive Mechanisms

Attention, decision making, perception, counting, planning, remembering—these are all oft-cited examples of cognitive functions. But what is a cognitive function? In this chapter, I will discuss some general aspects of cognitive functions that arise from considering the case studies presented earlier. While I won’t be advocating a particular notion of function, such as causal role (Cummins 1975), selected effects (Neander 1991), or some other notion, I will be highlighting the role that formal models of processing play in determining how cognitive systems act and adapt to their environment.¹

The first principle of my componential dynamicist theory of cognition states that cognitive systems execute formal models of processing, conceptually distinct from the mechanisms that execute them. The formal models of processing, mathematical formulae that detail the variables the organism must track and the interrelations between the variables, are distinct from the dynamical and physical mechanisms of cognitive systems, and are executed by the dynamical mechanisms. Arguing for these

¹ I have dropped the ‘information’ and speak only of processing. This sort of processing could be information processing, as commonly described in philosophy of mind and cognitive psychology. The processing could also be signal processing, as commonly described in electrophysiological neuroscience, which may or may not be distinct from information processing. At any rate, arguing that this processing is of a particular sort is part of a conceptual analysis beyond the scope of this appendix. Marr explicitly discusses information processing, so in the context of the discussion of Marr below, I will speak of information processing (Marr 1982).
two conjuncts will be one goal of this chapter. In addition to arguing that the formal models of processing are distinct from the mechanisms that execute them, I argue that these formal models of processing are cognitive functions. They correspond to processing challenges that the environment poses for cognitive systems, and which cognitive systems must execute if they are to behave adaptively. I also argue that cognitive functions are distinct from the role functions of the mechanisms that execute them. Cognitive functions correspond to executed formal models, whereas the execution of those formal models by the mechanisms of cognition is distinct: there need not be a one:one mapping between formal models and the mechanisms that execute them. Finally, I will illustrate how a consideration of the composition of such cognitive functions can result in a reinterpretation of the way that cognitive systems behave in a complex and adaptive fashion.

**D.1 What is a Formal Model of Processing?**

Cognitive systems are faced with a range of behavioral challenges that reflect the particular exigencies they face in navigating and manipulating the world. Different challenges face different organisms, and the particular challenges faced by an organism reflect the ecological context within which the organism is acting. An animal in the forest will face the need to detect predators occluded by foliage, whereas an animal in the desert will face different predatory dangers. Humans have perhaps the most diverse
environments to navigate. Most generally, perception, decision and other sorts of
cognitive processing require the organism to keep track of objects in the environment,
the various rewards and uncertainties associated with actions, and so forth. Different
objects will require different sorts of processing, different rewarding contexts will place
different processing demands on the organism, different perceptual contexts will require
different attentional commitments, and so forth.

This view of what cognitive functions consist in is a rationalist one. A rationalist
perspective is one that does not require an empirical analysis of a particular problem;
rather, the features of the environment dictate the problem. Analyzing cognitive
functions requires analyzing the behavior of cognitive systems; as Simon notes, “[i]f we
wish to know how an intelligent person will behave in the face of a particular problem,
we can investigate the requirements of the problem. Intelligence consists precisely in
responding to these requirements” (Simon 1990, p. 6). However, cognitive systems
rarely attain optimal performance due to intrinsic limitations of time, computing power,
and other mechanistic constraints. So while optimal, rational analyses of the demands
facing an organism can result in the optimal solution, this may not be the solution
settled upon by the system.

Formal mathematical models describe the processing necessary to overcome
these challenges. These formal models specify the variables that the system must
encode, the mathematical relationships between these variables, and the way that these encoded variables must be transformed in order to accomplish the goal of the organism, to respond to the behavioral challenge (providing such an account Marr calls a “theory” of the relevant cognitive domain; Marr 1982; Shagrir 2010). In the case of the RDMT, animals integrate evidence from the environment in order to make a perceptual decision, and in so doing they implement the DDM. In the case of patch foraging, animals keep track of average and instantaneous reward rates, travel times between patches, and other variables, and in so doing they implement the MVT. In either case, there is a mathematical formalism that describes the variables in the environment the animal must keep track of and the relationships between those variables.

Prima facie diverse behaviors can be unified under the same formal model. The formal model describes some general characteristics of the processing problem the system faces, and the theory behind the model states the goals the system needs to accomplish to overcome those problems. The processing problem refers to the properties of the behavioral problem facing the system. The theory that accompanies the problem outlines what the system needs to accomplish in its processing in order to behave adaptively. Furthermore, this outline of the problem determines what counts as a resolution to the processing problem. Take optimal foraging theory, discussed above in the context of foraging for alimentary resources in a patchy environment. This formal
theory applies to a wide range of behavioral problems, not just searches for food. In such problems, time needs to be allocated across resources that are distributed in some spatial context in a stochastic fashion, modeled as searches through a state-space with distributed resources (Hills 2006; Hills et al. 2008; Hills et al. 2010). A wide variety of cognitive tasks satisfy this description, including visual search (Cain et al. 2012), free recall (Hills et al. 2012), completion of sub-goals (Wilke et al. 2009), voluntary task-switching (Payne et al. 2007; Janssen et al. 2011; Farmer et al. 2011), study time allocation (Metcalfe and Jacobs 2010), and problem solving (Hills 2006; Payne and Duggan 2011).

Any adaptive behavioral problem involving the allocation of time between distinct resources can be captured by the optimal foraging formalism. Similar considerations apply to other formalisms for solving problems of adaptive behavior. What serves to unify, in this case, is the formal model, which allows seemingly diverse behaviors to fall under the same problem class.

These formal models are distinct from though executed by the mechanisms that compose cognitive systems. The dynamical properties of the system result from physical processes in mechanisms that transform signals and execute the formal models of processing. The accumulation of evidence present in the DDM is executed in the integrate-to-bound dynamical mechanism, which is in turn implemented in part by the increasing activity of neurons in LIP. Similarly, a mathematical function of comparing
instantaneous and average reward intake rates or tracking opportunity cost is executed in the integrate-to-bound dynamical mechanism implemented in part in ACC. The dynamics of the physical mechanisms execute the formal model.

How do the dynamics execute the formal model? Processing models are executed by dynamical mechanisms iff there is a mapping between the elements of the model and the elements of the mechanism. If there is such a mapping, I will say that the formal model is equivalent to the dynamical mechanism. There are two types of equivalence: strong and weak. Weak equivalence occurs when the input/output description of the formal model maps on to the dynamical mechanism, such that there are distinct states of the dynamical mechanism for distinct inputs to the formal model, distinct states of the dynamical mechanism for distinct outputs of the formal model, and a counterfactual

\[\text{(equation)}\]

\[^2\text{Note that this is a prospective claim, as the causal role of the ACC in this process has not yet been demonstrated. Nonetheless, for the purposes of illustrating the features of the theory, the case study will serve.}\]

\[^3\text{I have intentionally chosen these elements corresponding to the integrate-to-bound mechanism. A prima facie objection is that while these dynamical processes work well e.g. for the transformations in the mathematical model, they do not play any role in the encoding of the variables present in the formal model, such as the travel time or average reward rates in the MVT or the prior probabilities or likelihood ratios in the DDM. This, however, is not true. Variables are encoded in neural systems using a number of different dynamical mechanisms, including monotonic rate codes, where the firing rate of neurons encodes the value of the relevant variables, place codes, where the firing rate of a neuron relative to its neighbors encodes the relevant variable, or population codes, where the firing pattern of a neuronal population encodes the relevant variables. In all of these cases, however, the dynamical activities of the neuronal processes are the relevant aspect for encoding. Though I won’t argue for this view herein, this implicitly assimilates neuronal encoding mechanisms into the class of dynamical mechanisms that compose cognitive systems on my theory. Incidentally, these encoding mechanisms, I would contend, are prima facie candidates for representational vehicles on my view.}\]
mapping such that the sequence of dynamical mechanism states are arranged to preserve the input-output mapping that characterizes the formal model. Strong equivalence occurs when the input, output, and mathematical relations between the inputs and outputs are mapped on to the dynamical mechanism. In cases of strong equivalence, not only are there distinct dynamical mechanism states for the inputs and outputs, but the way that the dynamical mechanism transforms the input into the output can be described using the very same mathematical relations that occur in the formal model.

The mathematical functions that describe the mechanisms need not correspond to the functions that are present in the formal model. Integrative activity such as is seen in LIP during the RDMT can be described using different mathematical functions, such as integration, exponentiation (a simple first-order differential equation), or even using a prototypical ‘squashing’ function (Mazurek et al. 2003; Usher and McClelland 2001; Wang 2002, 2008; Churchland 2012). None of these functions are precisely the same function as the sequential probability ratios calculated in the DDM. Thus, strong equivalence is not required for a dynamical mechanism to execute a formal model of

\[ F \in \mathbb{R}^n \]

4 Insofar as we think of a mathematical relation as simply a set of ordered n-tuples \( F \) such that \( F \subseteq \mathbb{R}^n \) where \( \mathbb{R}^n \) is the \( n \)th Cartesian product of the set of real numbers on itself, then there is no distinction between the two. However, this is only the case if our domain is the entire set of reals. While this is perhaps true for the formal model executed by the system (and even then, only for some), this is simply not the case for mechanism states, unless there are uncountably many such states to map on to \( \mathbb{R} \).
processing. The formal model is executed so long as the dynamical mechanism’s activity maps onto some range of the formal model’s mathematical input/output sequence; that is, so long as the dynamical mechanism is weakly equivalent to the formal model.

Though the dynamical mechanisms execute the formal model, there is no necessary connection between any particular formal model and a dynamical mechanism. A formal model will be executed by some set of dynamical mechanisms, but which dynamical mechanisms execute that model may vary from system to system or within a system over time. Likewise, as amply illustrated by the two case studies discussed above, the same dynamical mechanism can execute different formal models. The integrate-to-bound mechanism in LIP executes the DDM, whereas the integrate-to-bound mechanism in ACC executes some aspect of the MVT. Thus, there is a many:many mapping between formal models and dynamical mechanisms.

Though weak equivalence is all that is required for the execution of a cognitive function by a set of mechanisms, there is a counterfactual constraint on cognitive function execution as well. Not only must it be the case that the mechanism is weakly equivalent to the cognitive function, but it must also be the case that if the inputs to the system were different, such that the formal model dictates a different set of values for the variables should obtain, then the mechanism would also be in a different set of states corresponding to those different values. The mechanisms of cognitive systems will be
weakly equivalent to any number of different formal models; enforcing the
counterfactual constraint restricts the range of alternate models to which the
mechanisms are weakly equivalent.

**D.2. Justifying the Ascription of Cognitive Functions**

What are the grounds for ascribing cognitive functions to systems? There are two
sorts of such grounds: grounds for cognitive function ascription in general, and grounds
for ascribing specific cognitive functions. The former presumably depends on the latter;
if all grounds for ascribing specific cognitive functions do not obtain, then the assertion
that a system is cognitive would be empty. But what sort of warrant must there be for
ascribing a particular cognitive function to a system? Justifying the ascription of
cognitive functions requires an account of the processing challenges facing an organism,
and how the formal model that is executed by cognitive mechanisms responds to those
challenges. Responding to those processing demands results in the execution of a formal
model. But to establish that a particular formal model is executed by a system, there
must be a theory that connects the processing problem, the formal model of the
processing, and the mechanisms that execute that formal model. Providing this theory in
turn grounds the ascription of a cognitive function to a system.

---

5 Excepting a position that maintains that ascription of cognitive function is amenable to elimination while retaining the notion of a cognitive system.
This account of cognitive functions draws on David Marr’s work on explanations in cognitive science. Marr’s seminal 1982 work Vision laid out a clear program for explanations of cognition, or as he framed it, understanding complex information processing systems. Marr distinguished three levels of explanation for information processing systems: the computational, the algorithmic, and the implementational. A complete cognitive explanation is filled in at these three levels, linking together the description of the computational problem the system solves with its proposed formal solution and the mechanism that implements that solution in the system. The first level of explanation is the “computational level”; this level corresponds to the level of the processing demands central to my account of cognitive functions. While Marr is focused on the structure of explanation in cognitive science, his analysis of the computational level is readily adapted herein as part of an analysis of cognitive functions. Insofar as the account being developed herein is not a theory about how or why explanations in cognitive science explain, but rather an attempt at providing conceptual grounds for constructing a theory of cognition, it might appear peculiar to discuss Marr, who emphasizes the nature of explanations of cognitive phenomena. The reason I am

---

6 Emphasis on adapted: I am not interested in a historical reconstruction of the Marr position, but rather in utilizing his account of the computational level to assess what must be included in an adequate theory of a particular cognitive function, and thus in an adequate theory of cognitive functions generally.
turning to Marr is that his account of explanation in cognitive science also provides justificatory grounds for ascribing cognitive functions.

In a series of articles, Shagrir has presented a sophisticated analysis of Marr’s computational level, detailing what he calls the appropriateness problem (Shagrir 2006, 2010a, 2010b). Marr approaches the problem from an information processing perspective, which he characterizes as a “mapping from one kind of information to another” (Marr 1982, p. 24) or as “mapping from one representation to another” (Marr 1982, p. 31). As Shagrir notes, Marr “does not provide a detailed account of what is meant by representational content or information…. Marr apparently identifies these terms with some sort of selective response to stimuli or with some reliable causal correlation between the activity of cells and certain types of stimuli” (Shagrir 2010a, p. 479). Shagrir follows Marr in using representational terms in discussing the account of cognitive functions. On my account, however, representation need play no role, and often it is argued that selective responsiveness or reliable causal correlation is insufficient for representation (Dretske 1986). I will emend Marr’s and Shagrir’s discussion in the following to reflect this: instead of representation, I will speak of encoding, in the sense of selective responsiveness or reliable causal correlation.

---

7 While Shagrir appeals to brain states, the account is easily appropriated for dynamical mechanisms; simply replacing the specification of brain states with dynamical mechanism states suffices for the same purpose, of executing a cognitive function via weak equivalence. I will speak of system states instead of brain states in order to generalize the argument.
In Marr’s theory of the analysis of cognitive systems, the computational level contains “separate arguments about what is computed and why”: the ‘what’ and the ‘why’ (Marr 1982, p. 23). A complete theory at this level will outline the problem the system must solve, including the processing goal. The ‘why’-element specifies the processing problem the system faces, the formal model that describes the relevant variables and relations inter alia, and provides a theory, in Marr’s terms, of why executing that formal model will solve that processing problem. First, a theory of a cognitive function must specify the processing problem posed by the environment and faced by the system. This problem specification will also specify the goal of the processing. In perception, the processing problem is to determine the properties of the object in the world. For example, Marr was concerned with an analysis of vision in particular, which faces the problem of determining how vision “…tells about shape and space and spatial arrangement” of objects in the environment from the sensory impingements on the peripheral sensory epithelia (Marr 1982, p. 36). The goal there is to determine what is where in the environment. Or, consider the problem of perceptual decision making under a noisy sensory stream, such as that faced by monkeys in the RDMT. The goal of the system is to identify the distal percept and select the appropriate action from a wide range of possible behavioral plans. Specifying a cognitive function starts with identifying what problem the system faces and how to solve that problem.
Second, a theory of a cognitive function will detail a formal model that describes the variables to be encoded and how they are to be transformed to reach the goal detailed in stating the processing problem. A formal mathematical model specifies both the variables relevant to accomplishing the processing goal and the mathematical relations between those variables. In the case of the RDMT, the formal model is the DDM; in the case of the patch-leaving problem, the formal model is the MVT. For Marr, the theory specifies the “resulting operation [that] is defined uniquely by the constraints it has to satisfy” (Marr 1982, p. 36). Specifying the relevant variables and their relations will result in a unique way to accomplish the processing goal of the system. In the RDMT, this unique solution is provided by the DDM because it is the optimal solution to a sequential evidence sampling process. Similarly, the MVT determines the optimal leave times for departing depleting patches. However, while there may be a unique specification in the sense of an optimal solution to the processing problem, the cognitive system may not implement the optimal solution. In the case of decision making, for example, different formal models exist, only some of which optimize a particular variable, and which variable is optimized can vary between models. Classical decision theory optimizes the overall utility of a choice set, defined as some transformation over the sum of the possible rewards accrued from a particular option multiplied by the probability of obtaining that particular reward. But heuristic decision making, such as
satisficing, does not optimize long-term utility in this fashion, rather settling for results that are good enough (Simon 1957). Regardless of the uniqueness or optimal nature of a particular formal proposal for a particular cognitive function, a theory of a cognitive function will always contain some formal proposal or other that encodes variables relevant to the processing problem faced by the system, dictating how those variables are to be transformed such that the system can behave adaptively and response to that problem.

Finally, the theory behind the formal model states why executing that particular model will solve the processing problem facing the system; merely stating a formal model of some sort is not good enough. For example, in defining one of the problems faced by foragers in environments where resources are distributed in discrete clumps, the Marginal Value Theorem optimizes the energy intake rate, one relevant variable for determining when a foraging system should cease exploiting a depleting resource and explore elsewhere (Charnov 1976). In Marr’s own theory of vision, the system computes second-crossings: where the second derivative equals zero for changes in contrast (Marr 1982). This formal model captures what external or internal variables are relevant to the system’s goals, and how the system’s processing must transform these variables for adaptive behavior. Note that the formal model is related to the mathematical characterization of the environment via this theory behind the formal model. In
foraging, the mathematical characterization of the environment will provide details such as the probability of encountering a patch of food of a certain type (e.g., bananas, apples, etc.), the average energy accrued from a patch of each type, the handling time (time to obtain, munch, swallow and then process for caloric benefit) for each food type, and similar variables. The MVT, which computes optimal leave times for environments containing patches of food, is a simple first-order differential equation that operates over these environmental variables. The formal model is thus related to but not identical to the mathematical characterization of the environment. Justifying the ascription of a cognitive function involves all three components: a statement of the processing problem facing the organism, the formal model stating the relevant variables and their relations inter alia, and a theory stating why that formal model resolves the behavioral challenges facing the organism.

The ‘what’-element specifies the internal processing operations present in the system when executing the formal model. An account of the structure of the system, including a mapping of the system states to the formal model, must be provided in order to establish that a particular formal model is executed. As argued above, a cognitive function is executed in virtue of a weak equivalence between the states of the system and the formal model that characterizes the goal of the system’s processing in response to the environmental problem facing the system. Since this weak equivalence
does not entail that the very same mathematical relations hold between the system states as hold between the variables standing for the relevant properties in the environment, the ‘what’-element describes what mathematical relation does hold between those system states. The grounds for ascribing a cognitive function require detailing both the ‘what’-element and the ‘why’-element behind the cognitive function.

The basic model for justifying the ascription of some cognitive function requires both ‘what’- and ‘why’-elements. Suppose that some system state B₁ encodes some object or property of the environment W₁. The ‘why’-element will pick out which objects or properties the system encodes, how those variables are related to each other, and how they are transformed. Since executing that formal model requires only weak equivalence and, as I argued above, as a matter of fact distinct mathematical relations do obtain between the encoding system states and the encoded world states (i.e., object or properties in the environment), specifying the ‘why’-element is insufficient. The ‘what’-element mathematically describes the causal relations between the system states.

In describing a system as an information processing system, Shagrir notes that the following question arises: “Why does the mapping process that starts from [an encoding] of [object/property] W₁... lead to [an encoding] of W₂?” (Shagrir 2010b, p. 278) We start with the encoding of W₁ and proceed causally from there; but nothing about that causal process guarantees that we end up with an encoding of W₂ as opposed to
some other property of the world. Shagrir calls this the “appropriateness” problem (Shagrir 2010a, p. 486ff), “as the task is to explain why the mathematical function f, which describes the relation between [the encoded input and output states], is appropriate for the information processing task, which is defined in terms of” the world states \(W_1\) and \(W_2\) (Shagrir 2010b, p. 278). For Shagrir, “[t]he role of the What element is to characterize, in mathematical terms, what is computed”, that is, the formal, mathematical model being computed, and the “… Why element… demonstrate[s] the basis of this mapping function in the physical world… or the appropriateness and the adequacy of this mapping to the information-processing task…” (Shagrir 2010a, p. 487).

So what is the appropriateness problem?

“The problem, in its most general form, is that of explaining why a process that starts from [an encoding] of \(W_1\) (i.e., \(B_1\)) ends up with [an encoding] of \(W_2\) (i.e., \(B_2\)) and not with [an encoding] of something else. How is it that the \(B_1\)-\(B_2\) relations track or mirror the external \(W_1\)-\(W_2\) relations? After all, we start the process with \(B_1\), which [encodes] \(W_1\), and then proceed from \(B_1\) to \(B_2\) through a mechanical (causal) process that takes place in our brain. We thus should wonder what is it about the causal relation between \(B_1\) and \(B_2\) that guarantees that the neural [encoding] of \(W_1\) will lead to a neural [encoding] of \(W_2\)” (Shagrir 2010a, p. 487).

In terms of the system states \(B_1\) and \(B_2\), the formal model specifies the relation between those states, the function being computed. But that’s not enough to answer the appropriateness problem, for now “we ask why the \(f\)-relation between \(B_1\) and \(B_2\) is appropriate for the information processing task that is defined in terms of \(W_1\) and \(W_2\): Why does the process that starts from a representation of \(W_1\)… and computes a mathematical function \(f\) end up with a representation of \(W_2\)?... Why is it that the \(f\)-
relations between $B_1$ and $B_2$ mirror... the $W_1$-$W_2$ relations?” (Shagrir 2010a, p. 487-488)

Take, for instance, the case of the RDMT. The dynamical properties of LIP cells are weakly equivalent to the DDM, in that there is a period of integration that maps on to the sequential summation of log-odds ratios present in the DDM. Assume that area MT encodes motion and the motion signal encoded in MT causes this integration in LIP.\(^8\)

Granted these two assumptions, why does LIP activity encode the log-odds ratio of the dots moving right or light, that is, the evidence in the noisy sensory stream? MT, after all, encoded something entirely different: the direction and speed of motion in the environment; it did not encode evidence, or log-odds, or anything of that sort. So why should a causal process driven by a system state the encodes one environmental variable, speed and direction of motion, result in a distinct system state that encodes a variable in the formal model? Does a computational process that starts with one encoding and executes some function over that encoding end up with the corresponding encoding that precisely matches the world?

Shagrir wants a more substantive answer than that the two different relations—between the world states inter alia, and between the system states inter alia—happen to be correlated. The appropriateness problem demands to know “what it is that facilitates this correlation.... Why is it that the [system state] relations can be correlated with the”

\(^8\) The physiological story is a good deal more complicated in reality, but suppose for the sake of discussion that this is accurate.
world state relations? (Shagrir 2010b, p. 278, italics in original) Shagrir notes that one answer is similarity: the system state relations are similar to the world state relations, and this similarity grounds the correlation. The similarity is not physical, but mathematical, “namely, that the [system state] mathematical relations between the [encoding] states are similar to the [world state] mathematical relations between the [encoded] features…” (Shagrir 2010b, p. 278). The very same mathematical relation that holds between the world states also holds between the system’s internal states. In virtue of this sameness of function, the system state output is mapped onto the world state.

A second similarity grounds the conclusion that the resulting system states encode the relevant properties and objects in the world. The similarity at the level of the mathematical relations between system states and world states is not the only similarity that grounds the appropriateness of the formal model. That similarity, after all, is due to the same mathematical connection between the system states as holds between the world states. But why should recapitulating that similarity be relevant to the problem faced by the organism? This is where the ‘why’-element plays a crucial role. The physical features in the environment serve as natural or physical constraints for the system’s processing, and the formal model defined in terms of the environmental variables and the constraints explain why the mathematical relations between the world states are defined the way they are. The mathematical relations between the system
states correspond to the relations in the world defined by the formal model, thus connecting the system’s causal processes to the processing problem the system must solve. Shagrir notes that this sort of account is ecological, since it refers to features of the environment within which the system processes information. However, the explanation is not adaptive, never invoking learning or evolution, though clearly the way the system comes to have the states it does, and the relations between those states, is due to learning or evolution. Evolution, or learning, simply doesn’t occur in an account of the appropriateness of a particular formal model for the processing demands on the system. Finally, though Marr’s explanation “appeals to similarity between the internal mapping relations and external relations between the features that are being [encoded]…. [t]he similarity is at a more abstract level of mathematical properties” (Shagrir 2010a, p. 489). The similarity can’t be physical, since the properties of the system aren’t like those of the targets of the encoding. The similarity must then be mathematical. In the final account, two similarities are present. First, there is the similarity in the function between the encodings and the mathematical relation holding between the states of the world. And second, there is a “similarity of mathematical structures” which grounds the first similarity and provides the explanatory resources for the ‘why’-part of the computational level theory (Shagrir 2010a, p. 489).
Consider again the RDMT and the DDM. The environmental demands placed on the organism, to determine the correct direction of motion and then indicate this with an eye movement, require that the system sample from a noisy perceptual stream to choose between the two possible directions. The processing problem faced by the organism matches the conditions for the class of problems covered by the DDM and for which the DDM is a provably optimal solution (for mathematical details, see Bogacz et al. 2006). The structure of this sampling process matches the mathematical structure of the environmental challenge, to determine the nature of the distal source of noisy sensory stimulation. The nature of that source is a signal embedded in noise, and that signal can be recovered from a sampling process, in particular the one precisely and optimally described by the DDM. The description of LIP activity in terms of integration or exponentiation does not just characterize the input to LIP from MT. It also describes how environmental input driving MT neuronal responses, and hence LIP neuronal responses, changes over time in a way that is isomorphic to the way that the DDM optimal sampling process proceeds. Thus, there is similarity not just at the level of the functions connecting the different types of states in the ‘what’ and ‘why’, but also in terms of the structure of these descriptions.

What is the structure of the response to the appropriateness problem? Shagrir provides the following schema:
1. $i(B_1) = W_1$, where $I$ = encodes.
2. $f(B_1) = B_2$.
3. $f(W_1) = W_2$.
4. $f(i(x)) = i(f(x))$.
Therefore: $i(B_2) = W_2$ (cf. Shagrir 2010a, p. 492; Shagrir 2010b, p. 278).

The first premise states that some system state $B_1$ encodes some world state $W_1$.

Applying this to the example of the RDMT and the DDM, different firing rates of MT cells encode the strength and direction of motion resulting from the percentage of dots that are moving coherently, how long the scene has been displayed, and so forth (Britten et al. 1992). The second premise states that there is a mathematical relation $f$ between $B_1$ and system state $B_2$. In the RDMT example, the firing rates of LIP cells are related to MT cells by an integrative, exponentiation, or similar sigmoidal function. This is a mathematical description of a causal relation between two distinct system states. The third premise states that in the formal model, there is the same mathematical relation $f$ between $W_1$ and world state $W_2$. In the DDM, this is a serial summation of log-odds ratios corresponding to the fact that the evidence for or against a particular direction of motion is driven by features of the environment such as the percentage of coherently moving dots. Such a premise must be justified by providing a theory for the cognitive function as discussed above. The fourth premise states a general isomorphism between the encoding function $i$ and the mathematical relation $f$, such that the structure of the

---

*Below, I will critique the claim that this is the same function $f$. 

295
encoding relations maps on to the structure of the mathematical relation $f$. For LIP cells and the DDM, the input driving MT neuronal responses, and hence LIP neuronal responses, changes over time in a way that is isomorphic to the way that the evidence sampling process operates in the DDM. That is, there is a general isomorphism between the integrative (or exponentiation or...) activity in LIP cells and the sequential summation of log-odds ratios in the DDM. The end result is a derivation of the connection between the resulting system state and the resulting world state: the initial system state encodes some world state, and once transformed in accord with the mathematical function $f$, results in a distinct system state that encodes another world state connected to the initial world state via the very same mathematical function $f$. Thus, the same function connects the two system states and the two world states. The argument schema provides grounds for claiming that the system state that results from the transformation in fact encodes the corresponding world state, and thus that the encoding transform present in the system corresponds to the relation between the states in the world. This in turn grounds the claim that the system is in fact executing some particular cognitive function, a function that is mathematically characterized by the formal model in the ‘why’-element of the theory for that particular cognitive function.

The similarity in the mathematical structure between the causal sequences of system states and the relations that reflect the operations defined in the formal model
over objects and properties in the world, together with the same transform \( f \) over the system states and the world states, provides the sort of justificatory grounds that makes the processing done by the system appropriate for the processing problem. The gist of the explanation, Shagrir says, is that

"an information processing task is mapping from one \([\text{encoding}]\) to another…. The goal of a computational analysis is to explain why the process that starts from \([\text{an encoding}]\) of \( W_1 \) (i.e., \( B_1 \)) ends up with \([\text{an encoding}]\) of \( W_2 \) (i.e., \( B_2 \)) and not with \([\text{an encoding}]\) of something else. A computational-level theory answers this question by pointing to a similarity at the more abstract level between the internal \( B_1-B_2 \) relations and the external \( W_1-W_2 \) relations. It states that the mathematical \( B_1-B_2 \) relations are “isomorphic”…. to the mathematical \( W_1-W_2 \) relations. The role of the What element is to specify the mathematical function that is being computed, namely, the \( f \)-relation between \( B_1 \) and \( B_2 \). The role of the Why element is to demonstrate that the \( W_1-W_2 \) relation is also an \( f \)-relation” (Shagrir 2010a, p. 491-492).

The ‘what’-element plays its crucial role in premise two above, by showing what formal model describes the relation between the states of the system. The ‘why’-element plays its crucial role in premise three above, by showing what the relation is between the states of the world. A critical caveat is in order: the function \( f \) is rarely actually computed by a system; rather, it is only approximated by the system. Similarly, \( f \) only approximates the distal relation between the world states. As Shagrir puts it, “… a visual system, as a biological system, seldom computes the function \( f \) that is stated by the computational-level theory; it only approximates it. Likewise, the function \( f \) is, at best, an approximation of the distal relation. A computational-level theory thus relates the ideally computed function \( f \), with an (ideal) external \( f \)-relation” (Shagrir 2010a, p. 493).

So the response to the appropriateness problem relates an ideally computed function
with an ideal external relation. However, for the conclusion to follow, the approximated functions must still be the same function holding between the system states and the world states. In addition to the identity of the function, which was the first similarity discussed above, the second similarity must also be present.

Shagrir’s analysis is insightful and provides the grounds for inferring that the output of the causal processes in the system will correspond in their mathematical structure to the relations between states of the environment and the operations codified in the formal model. For some cognitive functions and processing challenges, such as Marr’s theory of vision, the schema works as designed. However, for other cognitive functions, the account is too restrictive. The main issue is that the identical mathematical function need not obtain between the system states and the world states, though it may. Premises two and three state that the very same mathematical relation $f$ holds between the world states and between the system states. But such a claim is far too strong: for example, the brain does not and cannot implement many of the mathematical functions purportedly holding between world states as specified by the formal model. Many of those functions are continuous, and the brain may not be. Furthermore, neuronal signaling processes must operate with limited ranges, whereas those natural functions that hold between world states probably are not so limited. Finally, if we look at case studies from cognitive neuroscience, this sort of identity of function is not invoked. For
example, consider the neural activation present in LIP during the RDMT. That activity is taken to execute the DDM. The integration present in the mathematical model is an integration of evidence for or against some particular hypothesis, a serial summation of log-odds ratios. But the integration in LIP is often described as a mathematical integration of the motion input signal (Mazurek et al. 2003), an exponentiation with decay in activity where motion input drives the rise in activity (‘leaky integration’; Usher and McClelland 2001), or an exponentiation with motion input driving the rise in activity as well as recurrent self-excitation and cross-inhibition (Wang 2002). All three of these descriptions are conceptually distinct from the serial summation of log-odds ratios. Thus, interpreting the second premise along the lines outlined above, as LIP cell activity and MT cell activity being related as a function of the amount of evidence in the display, is false.

However, this objection is easily remedied: all that really needs to hold is an input-output mapping within a restricted range. This sort of simple isomorphism is precisely the weak equivalence relation described above that is at the heart of my account of how cognitive functions are executed by cognitive mechanisms. So the first similarity that grounds the appropriateness of the formal models of cognition is absent on my view, to be replaced by an input-output similarity. The justification for the execution of the formal model by a series of mechanism states then becomes:
1. \( i(B_1) = W_1 \), where \( I = \text{encodes} \).
2. \( g(B_1) = B_2 \).
3. \( h(W_1) = W_2 \).
4. \( f(i(x)) = i(f(x)) \).

Therefore: \( i(B_2) = W_2 \).

Note that the function \( f \) has been replaced with two distinct functions \( g \) and \( h \); this is simply to indicate that the two functions are distinct. However, we no longer have the justification for concluding that the resulting system state encodes the corresponding world state. What is need is a statement of input-output equivalence over some interval \( R \subset \mathbb{R} \); that is, that \( g \) and \( h \) map onto each other at some grain for \( R \). This results in the following:

1. \( i(B_1) = W_1 \), where \( I = \text{encodes} \).
2. \( g(B_1) = B_2 \).
3. \( h(W_1) = W_2 \).
4. \( g(x) \approx h(x) \), for \( x \in R \).
5. \( g(i(x)) = i(g(x)) \).\(^{10}\)

Therefore: \( i(B_2) \approx W_2 \) for interval \( R \).\(^{11}\)

Premise four states that the output of \( g \), \( g(x) \), is approximately equal to the output of \( h \), \( h(x) \), over the domain of inputs \( R \). There is some flexibility in the equivalence requirements, hence the approximately equal condition; just how closely the outputs of \( g \) and \( h \) must be, which I call the grain of the equivalence, is relative to whatever standards

\(^{10}\) Note that I have changed the old premise four into the new premise five to reflect the shift in function.

\(^{11}\) The derivation runs as: by premise five, \( g(i(B_1)) = i(g(B_1)) \). But then by premise one, \( g(W_1) = i(g(B_1)) \). By premise four, \( g(W_1) = h(W_1) \), for some interval \( R \). So by premise three, \( g(W_1) = W_2 \), for some interval \( R \). So then \( i(g(B_1)) = W_2 \) for some interval \( R \). But since by premise two \( g(B_1) = B_2 \), \( i(B_2) = W_2 \) for some interval \( R \).

QED.
of the science are in place. This new argument incorporates weak equivalence, respecting the reality of the mechanisms of cognition as well as the empirical evidence for how research programs in cognitive neuroscience proceed.

Weakening the requirement to weak equivalence, however, comes at a price: there is no longer the connection between the ‘what’ and the ‘why’ components, and there are now two distinct formal characterizations present. In short, why should we grant premise (4)? The first formal model corresponds to the ‘why’-element and describes the mathematical relationship between environmental states that accomplishes the processing goals laid out by a theory of the cognitive function. The second formal model corresponds to the ‘what’-element and describes the mathematical relationship between the system states. For Shagrir, two similarities dictate the appropriateness of the formal model for the processing challenges faced by the cognitive system. The first similarity was between the functions executed over states of the system and the function relating the states of the environment. This similarity condition was weakened above to weak equivalence. With only weak equivalence between the set of states of the system

---

12 There may be ways of deciding between formal models in virtue of one model being approximately equal to the relations between the world states at a finer grain than another model, though I won’t be discussing such considerations herein.

13 In fact, there are three: recall that the formal model, the ‘why’-element, is defined in terms of environmental variables but is not identical to the mathematical description of the environment. I am eliding discussion of the mathematical description of the environment per se since the relevant model for the cognitive function is the formal model, defined partly in relation to the mathematical description of the environment.
and the set of states of the world, the similarity of the relations between the system’s states and between the world’s states is not guaranteed. This seems to make the appropriateness problem even more acute: why should a system state that follows some other state according to an internal sequence as determined by some mathematical relation be mapped on to an external state in the world if the mathematical relation is different? As applied to our sample case study of the RDMT, why should an integration (or exponentiation or…) over MT cell activity encoding motion direction and strength result in an encoding of log-odds ratio as defined by the DDM?

The resolution is to note that despite the dissimilarity at the level of the mathematical function, there is still a structural similarity between the encoding domain and the encoded domain. Removing the similarity of the function does not remove the second similarity, which on Shagrir’s account grounds that first similarity. The second similarity was between the mathematical structure of the internal components of the system and the mathematical structure of the world, such as the mathematical description of the sequence of LIP transforms of MT activity in the RDMT and the mathematical description of the sequential evidential sampling process in the DDM. The similarity of these structures can remain, however, even if the specific formulae that are utilized to relate the internal states is strictly speaking different from the specific formulae that relates the external states. As a consequence, we can retain premise four.
on the same grounds as were cited to justify the use of the same mathematical relation between the system states and the world states. Consider the DDM and the dynamic activity observed in LIP that results from motion encoding in MT. That activity clearly shares structure, even if the specific functions being executed are no longer the same. Thus, despite the presence of two distinct formalizations, one reflecting the mechanism and the other reflecting the formal model characterizing the cognitive function, the isomorphism between the two mathematical structures in which the local functions are embedded can remain, at some granularity determined by R. And so, premise four is justified, though now the encoding is only an approximate one, relative to some interval and grain. Once the derivation is in place, we are now more justified in ascribing to the processing in the system, the particular cognitive function that is mathematically characterized by the formal model.

D.3. Constructing a Cognitive Ontology

Thus far, I’ve defined a formal model of cognitive processing as a set of variables, parameters, and their interrelations, together with an account of why executing that model will respond to the processing demands placed on the system; I’ve defined what it means for a cognitive system to execute a formal model, in terms of weak equivalence; and after criticizing Shagrir’s presentation of Marr’s justification for utilizing some formal model, I presented a revised justification that is consistent with the requirement
for weak equivalence. The formal models of processing are mathematical functions.

Execution of these models occurs in virtue of a weak equivalence between some subset of the properties of the dynamical mechanisms of a cognitive system and the formal model. The justification for selecting some particular formal model over another is grounded in an analysis of the processing demands on the organism and an argument connecting the mathematical relations in the model to the properties of the environment, describing how transforming the encoded variable will result in the attainment of the processing goal.\(^\text{14}\)

---

\(^{14}\) While I won’t defend cognitive functions as being of one species of function or another, this justification seemingly provides grounds for thinking of the execution of formal models by dynamical mechanisms, that is, cognitive functions, as goal-oriented functions (Adams 1979; Enç and Adams 1992; Boorse 2002). The justification of a particular formal model in terms of the challenges facing an organism results in cognitive functions being a species of goal-directed functions. Goal-directedness is an important part of these processing models in two senses. First, the justification for a formal model identifies the goal in the analysis of the problem (e.g., determining what is where in the case of vision (or, if you prefer affordances, determining what I can do); gathering the largest amount of energy, information, etc. in a given period of time in the case of foraging or decision making; increasing the sensitivity of processing in the case of attention; and so on). Second, the system’s components have the goal of executing a particular processing model.

What does it mean for a function or model to be goal-directed? In the case of the formal models of cognition, goal-directedness could be defined in terms of a discrepancy function: some assumption that dictates a quantity that is to be optimized, and some measure that determines the discrepancy between the optimal quantity and the current value. The goal-directedness would then amount to the minimization of the discrepancy score. Suppose, for example, that we want to know how many apples are in a barrel. The optimal quantity in this case is the actual number of apples in the barrel. The discrepancy function is simply defined as the difference between our count and the true number of apples in the barrel. Defining discrepancy in this way requires that we know how many apples there are in the barrel. Often, however, we do not know the true quantity. Many different measures in statistics have been developed to handle these situations, providing a confidence bounds on our estimate for the number of apples in the barrel. The principal problem with this approach is that many formal models of processing define quantities that are either explicitly suboptimal, such as occurs in models of bounded rationality (Gigerenzer et al. 1999; Gigerenzer and Gaissmeier 2011), or for which no clear sense of optimal can be defined (e.g., choosing a mate, a job etc.). Without a notion of optimality, defining a discrepancy function is impossible.
This processing approach suggests a method for constructing a cognitive ontology, that is, the set of cognitive functions that are executed by cognitive systems. In particular, it suggests that a cognitive ontology should be driven by the environment-organism interaction to determine what the organism needs to do in order to behave flexibly and intelligently in its environment. This ethologically driven approach is also consonant with evolutionary considerations.

What is a cognitive ontology? Poldrack defines a cognitive ontology as “a theory about the structure of the mind that specifies the component operations that comprise mental function.... This ontology describes the “parts” of the mind....” (Poldrack 2010, p. 753; see also Bilder et al. 2009). For Poldrack, a cognitive ontology specifies the component operations that comprise what the mind does. Insofar as we conceive of component operations as themselves functions, then Poldrack’s cognitive ontology is a specification of the component functions of mind. Contrast this definition with Price and Friston’s focus on structure-function association as central to a cognitive ontology. They define a cognitive ontology as “a systematic definition of structure–function relations whereby structures predict functions and functions predict structures” (Price and Friston 2005, p. 263). They argue that though functions can be described at multiple levels, the level of description to be preferred is one that allows for cross-cognitive domain applicability.
“...the function of a neuron or neuronal population depends on its interactions with other neurons. At one level of description, a neuron can only do one thing—fire. The firing will stimulate activation in a fixed set of output regions. Therefore, there will be a limited range of functions that an area can perform. However, at another level of description, the consequence of the firing will depend on which neuronal populations (input regions) caused the firing, and which neuronal populations send concurrent signals to the output regions. In this sense, a neuron participates in multiple functions” (Price and Friston 2005, p. 268).

A more insular or sectarian approach prevents the kind of mapping that allows for this systematic connection. “[T]he most useful functional labels are those that explain and predict how an area responds in different contexts”, that is, in different tasks, and by implication the level of function that permits this sort of connection is preferable (Price and Friston 2005, p. 267). Importantly, this suggests that we can’t have a plurality of ontologies, for brain areas and for cognitive functions, for example, as one that unites will be preferable to one that does not.

What are the criteria for constructing a cognitive ontology? Poldrack has a selective association criterion for assessing cognitive ontologies. Poldrack says that the “...correctness of the ontology would be reflected in selective association between structures and functions. That is, if a specific structure or network is activated in association with only one putative cognitive process, then one could argue that the reality of this process has been established” (Poldrack 2010, p. 754). The selective association of a possible psychological factor with a particular brain region is the inclusion criterion for a cognitive ontology. Price and Friston offer two criteria for “good” ontologies:
“1. Have a hierarchical structure that predicts the coactivation of anatomical regions, where sets of coactivated regions should have demonstrable (effective) connections.
2. Enable cognitive processing to be predicted given any distribution of activations, based on which area, or set of areas, is necessary for that processing” (Price and Friston 2005, p. 272).

Both of these are predictive criterion. The first states that a ‘good’ ontology will have a hierarchical structure such that it predicts coactivation of regions when certain functions are co-executed (my terminology), and these coactivated regions have demonstrated connections between them to support the hypothesized functioning. The second states that a ‘good’ ontology will accurately predict which cognitive functions are being executed (again, my terminology) given a set of activated areas.

Both of Poldrack’s and Price and Friston’s criteria focus on structure-function mapping for constructing a cognitive ontology. Interpreting structure as indicating dynamical mechanisms, the previous discussion indicates two important issues with this approach. First, structure-function mapping may be more piecemeal than such a mapping implies. Second, an additional constraint on our cognitive ontology derives from considering the ethological demands placed on a cognitive system.

Structure-function mappings may be more piecemeal, with multiple structures described using different mathematical models jointly executing a particular cognitive function, than with the one:one mapping between structures and functions that Poldrack

---

15 Though neither Poldrack nor Price and Friston discuss dynamical mechanisms, I here discuss the viability of structure-function mappings, construed as dynamical mechanism-function mappings, for constructing a cognitive ontology. This adapts their approach to the componential dynamicist theory I’m developing.
and Price and Friston seem to imply. The integrate-to-bound can be described in a number of ways, such as integration, exponentiation, or squashing function as discussed previously, but none of these correspond to the entire DDM or MVT. The formal models that correspond to cognitive functions have their component variables and parameters encoded in different dynamical mechanisms, spread across the system.

Both Poldrack and Price and Friston may reply that the ideal level to assess cognitive functions is one that permits function to structure and structure to function predictions. This is explicitly stated in Price and Friston’s criteria for a cognitive ontology, and implied by Poldrack’s criterion of selective association. But the one:one mapping criterion is undermotivated: there is a distinction between the functions of the cognitive system and the functions of its parts, and assuming there is not begs the question against such views; and the predictability criterion that militates for the one:one criterion can be partly satisfied in the absence of a one:one mapping between structure and function.16

First, there is no reason to think that a function of the cognitive system will necessarily map on to a function of the parts of the system. Specifically, there are two

16 There’s another reason to be skeptical of such a strict criterion, which space limitations prevent me from exploring fully. Prima facie, if cognitive functions are multiple realizable, capably of being realized in many different sorts of systems, then a one:one structure:function mapping will only hold for particular types of cognitive systems (Polger 2004; Figdor 2010). I hedge this claim because the multiple realization debate is very mature and very deep, and the very possibility of multiple realizability is disputed (e.g. Shapiro 2000, 2004).
types of function at work in cognitive systems: cognitive functions, in the sense just
explicated, and mechanistic role functions of the components of cognitive systems.
Recall that cognitive functions are formal models of processing that are executed by
cognitive systems in response to processing challenges posed by the environment. In
virtue of a weak equivalence between the dynamics of cognitive systems and the formal
model of processing, cognitive functions are executed by cognitive mechanisms. These
are not the only functions present in cognitive systems. In addition to executing these
cognitive functions, component cognitive mechanisms also have their own functions,
mechanistic role functions. Part of analyzing mechanisms involves the attribution of
mechanistic role functions, describing “an item in terms of the properties or activities by
virtue of which it contributes to the working of a containing mechanism” (Craver 2001,
p. 61). Specifying mechanistic role functions, those activities that are played by the
components of the mechanism, requires the functional characteristics of the mechanism
to be provided, that is, the way the parts of the mechanism change either
spatiotemporally or in respect to each other.17 Those functional characteristics are
constrained by the weak equivalence relation outlined above; so long as the parts of the
mechanism have the right sort of properties for mapping onto the inputs and outputs of

17 Craver characterizes mechanistic role functions as a species of causal role function (Craver 2001; Cummins
1975). I happen to agree, but nothing in this current argument turns on how mechanistic role functions are
assimilated into the typology of functions.
the formal model, then the mechanism has the right sort of functional properties for executing the cognitive function. In the above account of cognitive functions, these mechanistic role functions are described mathematically as relations between system states, the ‘what’-element in a theory of a cognitive function.

Similar distinctions appear in both the cognitive and biological literature. For example, Bock and von Wahlert distinguish form, the description of the material composition and arrangement of some biological feature, from function, the description of the causal properties of some feature arising from its form, and define the faculty as the combination of the form and function. Form is “… the class of predicates of material composition and the arrangement, shape or appearance of these materials…” (Bock and von Wahlert 1965, p. 272-273). Functions are “… that class of predicates which include all physical and chemical properties arising from its form (i.e., its material composition and arrangement thereof) including all properties arising from increasing levels of organization…” (Bock and von Wahlert 1965, p. 274). Faculties results from combinations of forms and features; they are “defined as the combination of a form and a function of a feature. It may be defined formally as… that class of predicates each of which includes a combination of a form (material composition and arrangement) and a function (physical and chemical properties) of the feature…” (Bock and von Wahlert 1965, p. 276). Different faculties can result from the same form possessing different
functions (Bock and von Wahlert 1965). Translating between the accounts, Bock and von Wahlert’s functions are roughly mechanistic role functions, and their faculties are roughly cognitive functions.

Anderson and Bergeron draw a similar distinction (Bergeron 2007; Anderson 2010). Bergeron argues that one mode of functional specification regards the specification of a component’s cognitive role, “the function of a particular component… specified relative to a cognitive process, or group of such processes, in which that component is thought to participate” (Bergeron 2007, p. 181, italics in original). Another mode regards the specification of a component’s cognitive working, “the component’s function… specified relative to a cognitive process, or group of such processes, that it is thought to perform” (Bergeron 2007, p. 181, italics in original). Anderson similarly notes a distinction between the functions of local circuits, or circumscribed neural areas, and cognitive functions (Anderson 2010). Anderson notes that “… the “workings” of local neural circuits are put to many different higher-level “uses”, and that the flexibility and variety of our cognitive repertoire results in part from the ability to put together the same parts in different configurations to achieve different behavioral outcomes” (Anderson 2010, p. 295). For Anderson, a working is “whatever specific computational contribution local anatomical circuits make to overall function” (Anderson 2010, p. 252) or “whatever single, relatively simple thing a local neural circuit does for or offers to all of
the functional complexes of which the circuit is a part” (Anderson 2010b, p. 295, italics in original), and a use is “the cognitive purpose to which the working is put in any individual case” (Anderson 2010, p. 252). Translating again between accounts, Bergeron’s cognitive roles or Anderson’s workings are the mechanistic role functions, and Bergeron’s cognitive workings or Anderson’s uses are the particular aspects of the formal model to which the mechanistic role is weakly equivalent.

A similar distinction is drawn by Young and colleagues between a global function and the functions of the processors in the system. Young et al. define five distinct notions of function operative in cognitive neuropsychology. A particular global function $f_g$, that is, some aspect of “the behavior of the whole animal” (Young et al. 2000, p. 155), can be considered “…to be delegated among the processors in the brain in such a way that some set of processors’ functions ($f_c$) are sufficient to generate the global mapping observed. Each processor’s function $f_c$ could also be captured formally as a mapping between its inputs and outputs in the context of the connectivity and dynamics of the system” (Young et al. 2000, p. 156). Young et al.’s usage correlates very closely with mine: the global functions are my cognitive functions, and the processor functions are my mechanistic role functions.

The distinction between cognitive functions, considered at the systemic level, and mechanistic functions, considered at the level of the components of the system, is
widespread in the literature. By assuming that the ideal level for constructing a cognitive ontology is the level of the components, Poldrack and Price and Friston beg the question against these alternatives. In reply, however, Poldrack and Price and Friston could appeal to the predictive power associated with constructing a cognitive ontology. This is explicitly present in Price and Friston’s two criteria for a good cognitive ontology, which is meant to permit inferences about function based on structure and inferences about structure based on function. On the predictive power response, the individuation of cognitive functions should occur at the level at which a one:one mapping between structure and function is possible. Such a mapping will have the greatest predictive power, and insofar as predictive power is important to our scientific inferences and experiments, a cognitive ontology that maximizes predictive power relative to another such ontology is eo ipso preferred.

This brings us to the second reason that the one:one, structure:function cognitive ontology approach is undermotivated: the predictability criteria can be partly satisfied even in the absence of a one:one approach. Poldrack and Price and Friston desire a cognitive ontology that permits inferences between the structural properties of neural areas and the cognitive functions of those areas. If we observe some set of structural properties, including activity during some cognitive task, then we can predict the

---

If we remain at the level of mechanistic role functions, this one:one mapping appears much more plausible.
execution of some cognitive function during that task. If instead we analyze a particular task such that the execution of some cognitive function is required, then we can predict activity in some particular structure in the system.

While this one:one mapping is not plausible on my view, what is permitted is a weaker form of predictability. Insofar as there must be weak equivalence between the formal model that specifies the relevant variables and their interrelations and the mechanisms that compose cognitive systems, there will be some constraints provided in either direction by supposing some function or some mechanism is relevant for a behavior during a task. Granted a theory that picks out the formal model to be executed by a mechanism, if a task analysis determines that such a formal model should be executed during a task, then there are constraints on the functional organization of the mechanisms that jointly execute that function. Similarly, uncovering a particular mechanism will serve to provide constraints on the possible formal models implemented by the system.

Undoubtedly, Poldrack and Price and Friston would agree with these claims. What is desired, on their view, is a more robust mapping, such that granted knowledge of one side of the relation, we will be able to pick out a unique mechanism or function on the other side of the relation. Otherwise, they would contend, our cognitive ontology is predictively less fruitful than one that does provide this inferential power. The
problem with this reply is no longer about the predictability claim per se, as some predictive power is granted. Rather, it ignores the way that cognitive functions are analyzed: cognitive functions are analyzed as formal models that are responsive to the processing demands placed on organisms by their environments. A theory of the processing required for adaptive behavior, including a specification of the formal model and why executing that formal model will result in adaptive behavior, occurs at the level of the organism. Perhaps, ceteris paribus, the presence of a mapping from this organism-level analysis of the cognitive function to the level of the component mechanisms is preferred. But as demonstrated by the case studies of the RDMT and the patch-leaving task, often times the individual mechanisms will only map on to parts of the relevant formal models. Ultimately, we want our criteria for a good cognitive ontology to respect not just inferential power, codified in the predictability criteria, but also to respect the actual functional organization of cognitive systems and our theoretical analysis of cognitive functions.

I conclude that a cognitive ontology should provide predictive power, but not just predictive power, and not necessarily in the form of the one:one structure:function mapping touted by Poldrack and Price and Friston. In addition to this predictive power, and in line with the analysis of cognitive functions above, a theory of each purported cognitive function is also required. That theory will motivate the formal model to which
the functioning of the system is weakly equivalent by considering the environment of the organism and the ethological demands placed on the organism’s processing. A sound cognitive ontology is not only predictive, allowing inferences from cognitive functions to sets of possible mechanistic functions and inferences from sets of mechanistic functions to the formal structure of the cognitive functions therein executed, but also ethologically responsive, reflecting the environmental challenges facing the organism.  

**D.4. Getting Our Cognitive Ontology Right**

The importance of getting our cognitive ontology right can be illustrated by considering the case of so-called executive functions and the neuropsychological tasks used to probe this function in both control and patient populations. This discussion will also suggest that getting our cognitive ontology right outlines a way for the emergence of more sophisticated cognitive function from the execution of simpler functions. Not only this, but getting our cognitive ontology right has practical implications, for how we diagnose and treat mental dysfunction.

A fundamental executive function in cognitive systems is the ability to ‘follow rules’. Two prominent clinical assessments of rule-following are the Iowa Gambling

---

Note that these environmental challenges need not be the historical environmental challenges present in the ancestors of the organism; they may also reflect current challenges from the contemporary environment (Keeley 2000).
Task (IGT) and the Wisconsin Card Sort Task (WCST). Both tasks have been proposed to assess executive function, and in particular the ability of patients to appropriately incorporate information about rewards in the environment to augment future behavior. This is often cast in terms of rule-following, that is, in terms of the subject inferring or otherwise learning about a particular local generalization, one that holds for the current task environment, and then following that generalization in order to accrue reward. I contend that these tasks probe many cognitive functions at once, thus failing to probe any individual cognitive function, due to each task enlisting multiple cognitive functions for successful performance. Understanding the formal framework of processing that a patient’s brain must implement in order to perform optimally on a task provides new explanatory and predictive resources, potentially guiding new clinical and therapeutic insights.

The Iowa Gambling Task (Bechara et al. 1994; Bechara 2007) is a risk-based gambling task. Subjects are given an initial sum of money and attempt to maximize their profit by selecting cards from one of four decks. The different decks pay out randomly, with some decks advantageous over the long term, resulting in a net gain, and others not, resulting in a net loss. The IGT is utilized to assess cognitive function in patients with focal brain lesions, as well as a number of psychiatric illnesses (Bechara 2007). In particular, patients with ventromedial prefrontal cortex (vmPFC) lesions exhibit more
disadvantageous deck choices (Anderson et al. 1999) and show no anticipatory affective autonomic response prior to selecting from the risky decks (Bechara et al. 1996). Whether other prefrontal regions are implicated in deficient choices in the IGT is controversial.

Some studies find that dorsolateral prefrontal cortex (dIPFC) and dorsomedial prefrontal cortex (dmPFC) damage does result in deficient choices compared to controls (Manes et al. 2002; Fellows and Farah 2005), while others fail to find such deficits in dIPFC lesioned patients (Bechara et al. 1998, 2000b; Bechara and Damasio 2002; Bechara 2003; Fellows 2004). Patients with amygdalar lesions show a deficit in IGT performance (Bechara et al. 1999; Brand et al. 2007a; Brand et al. 2007b), though such patients exhibit general physiological deficits. vmPFC patients, in contrast, exhibit appropriate physiological responses to outcomes on the task, only lacking the anticipatory autonomic response (Bechara et al. 1999). There is also a laterality effect, where right vmPFC damage results in impairment and left vmPFC damage does not, hypothesized to be connected to the association of right hemispheric activity with affective processing (Clark et al. 2003; Tranel et al. 2002; Buelow and Suhr 2009). Adding to this confusion, neuroimaging studies of healthy controls implicate medial orbitofrontal cortex (mOFC) in mediating IGT performance [double check this](Adinoff et al. 2003; Bolla et al. 2003; Ernst et al. 2002; Grant et al. 1999; Tucker et al. 2004; Windmann et al. 2006; for review, see Buelow and Suhr 2009).
Although there has been criticism of the IGT in the past (Buelow and Suhr 2009; Gansler et al. 2011a; Gansler et al. 2011b), the IGT is a perfect illustration of the shortcomings of focusing on gross anatomical and behavioral deficit. What formal processing must the patient’s brain execute in order to behave optimally on the task? Or what processing challenges is the system responding to as presented by the task environment? Certainly keeping track of risk, defined as the coefficient of variation (McCoy and Platt 2005; Weber et al. 2004), is an important component of the task. Risk influences expected value computations, and risk computations potentially explain the mOFC activation present in neuroimaging during the IGT, as the mOFC encodes the value of options in the environment and mediates reward-guided decision making (Padoa-Schioppa and Assad 2006; Noonan et al. 2010; Walton et al. 2010; Watson et al. 2012). However, the IGT is more complicated than simple tracking of risk. In particular, optimal performance on the task requires tracking reward rates over time, a fundamental capacity of organisms that is aptly captured by optimal foraging theory (Stephens and Krebs 1986). Dopaminergic signals originating in the basal ganglia play a fundamental role in reward signaling (Schultz; Others), and subjects with dopaminergic disorders such as Parkinson’s patients show deficits in foraging behavior (Rutledge et al. 2009). Disorders evident in IGT behavior may result from a failure to implement foraging models, particularly resulting from deficits in the targets of dopaminergic
projections to prefrontal areas. Critically, however, risk assessment, expected value computations, reward intake rates, model-based learning of the rewarding characteristics of the different decks, and presumably other formal models of cognitive functions are implicated in successful IGT performance. Thus, the IGT assesses behavior in a fashion that combines multiple cognitive functions.

Analyzing evidence from the Wisconsin Card Sort Task (WCST) suggests a similar conclusion. Utilized in the assessment of frontal lobe dysfunction, the WCST requires the subject to match a sample card with one of four key cards along one of three dimensions: color, number and shape (Grant and Berg 1948; Milner 1963; Heaton 1981; Heaton et al. 1993). The subject is not informed of the matching rule, but receives feedback after each sample card about whether the categorization was correct. After ten consecutive correct matches, an unsignaled change in the active matching rule occurs, and the subject has to explore to determine the new rule. The subject’s performance can be analyzed along a number of dimensions, including the number of completed rules, number of perseverative errors (sticking with an old rule after a switch), and number of non-perseverative errors (switching from a correct rule; Nyhus & Barcelo 2009). A number of brain regions have been implicated in successful WCST performance. Milner’s original 1963 study found more perseverative errors for dIPFC lesioned patients than those with OFC, temporal or parietal lesions (Milner 1963). The relative
importance of frontal cortex for WCST has since been corroborated by a large number of studies (see Nyhus and Barcelo 2009 for recent review). However, damage to temporal (Corcoran and Upton 1993; Giovagnoli 2001; Hermann, Wyler, and Richey 1988; Horner, Flashman, Freides, Epstein and Bakay 1996; Strauss, Hunter and Wada 1993), subcortical (Mukopadhyay et al. 2008), hippocampal (Corcoran and Upton 1993; Giovagnoli 2001; Igarashi 2008), and cerebellar (Mukopadhyay et al. 2008) regions impairs WCST performance. Neuroimaging studies of patients or normal controls have revealed increased activation of the dlPFC (Berman et al. 1995; Gonzalez-Hernandez et al. 2002; Kawasaki et al. 1993; Lie et al. 2006; Lombardi et al. 1999; Marenco et al. 1993; Mentzel et al. 1998; Monchi et al. 2001; Nagahama et al. 1996; Nagahama et al. 1997; Nagahama et al. 1998; Ragland et al. 1998; Rogers et al. 2000; Volz et al. 1997; Wang et al. 2001), and vlPFC (Lie et al. 2006; Monchi et al. 2001), among other, non-frontal areas (Nyhus and Barcelo 2009).

The breadth and number of regions activated suggests that the WCST requires a number of different cognitive functions for successful performance. For example, ‘set-shifting’, the ability to switch between active rules, is often invoked as one of the main functions probed by the WCST and disturbed in patient populations that exhibit WCST deficits (Nyhus and Barcelo 2009; Barcelo et al. 1997; Rubinstein et al. 2001; Shallice et al. 2008; Braver et al. 2003; Monsell 2005). However, once again, optimal performance on
the task requires keeping track of formal variables such as reward rates, or constructing and updating models of the environment such as in model-based reinforcement learning. In particular, upon a decrease in the local reward rate, the system may be forced into an exploratory regime where it attempts to determine the correct model. The notion of set-shifting as switching between encoded rules may be an outdated legacy of the classical approach to theorizing about how rules are implemented by cognitive organisms (Haugeland 1985), and instead performance on the task may be driven by cognitive foraging mechanisms as subjects search through an abstract space of possible patterns of behavior to determine the adaptive response (Hills and company).

Individuals with nonspecific brain injury have been shown to exhibit deficits on the WCST and a foraging task, exhibiting a preference for local reward rates (Schlund 2002), indicating the potential benefits of viewing aberrant WCST behavior through the lens of formal foraging theory. Similar lessons may be gleaned by reconceptualizing how errors on the WCST are classified and analyzed. Non-perseverative errors have been criticized in the past as conflating two types of error, called efficient (appropriately exploring for a new rule) and random (inappropriately switching rules) errors (Barcelo & Knight 2002). Perhaps, however, the neural mechanisms underlying WCST performance evolved to support a simulated annealing process, whereby changes in behavior permit exploration for and potential exploitation of new resources. If so, both types of error are forms of
exploratory behavior owing to the implementation of models for foraging through possible behavioral patterns in order to optimize reward rates. Likewise, perseverative errors, associated with dIPFC lesions (Rogers et al. 1998; Shallice et al. 2008)[double check‼], may not result from failures to disengage from previously activated rules, but failures to appropriately assess reward rates or failures to integrate local reward rates with information about the environment, both resulting in failures to forage in the space of possible actions to maximize reward. Much like the IGT, the WCST assesses cognitive function in a fashion that may in fact cut orthogonally across multiple different cognitive functions needed for successful performance on the task.

Reconceptualizing what the system is doing by focusing on the evolved cognitive functions, ones that respond to the natural ecology of the organism, can provide insight into the more complex processes evoked by tasks like the WCST or the IGT. As Anderson comments, “brains evolved to control action…, to manage the values of agent-environment relationships” (Anderson 2014, p. 134), and

“the organism perceives the values of salient organism-environment relationships and, in light of some goal, acts so as to perceive the right changes in those relationships. The brain that manages this process is structured in such a way that its various parts have different dispositions to manage the values of these perceived relationships, and the summed cooperation and competition between the active dispositions of these regions in a situation both determines the current goal and structures the control loop that facilitates the required behavior” (Anderson 2014, p. 138). Focusing on the basic suite of such relationships results in a set of cognitive functions, related to ethologically relevant aspects of the environment, such as food, friends, and
foes. Characterizing how organisms respond to these basic processing problems—when to leave a depleting food source to travel to a new one? which environmental stimuli are rewarding and which punishing? is that a predator approaching through the bushes?—can result in a different set of basic cognitive functions. These cognitive functions can then operate simultaneously to produce more complex cognitive activity. And, since many such component cognitive functions are operating simultaneously, the activation of many different brain regions to support more complex behavior no longer appears incomprehensible.

D.5. Conclusion

In this appendix, I have provided an account of cognitive functions in terms of the formal models of processing executed by cognitive mechanisms. These formal models exhibit a weak equivalence with the cognitive mechanisms. Furthermore, I’ve argued that the justification for selecting a particular formal model as the one executed is provided by a consideration of the ethological challenges facing organisms in their environment. The set of cognitive functions considered together comprises our cognitive ontology. I distinguished between these cognitive functions, which are executed by the whole cognitive system, and the particular mechanistic role functions of the individual mechanistic components of such systems. Finally, I illustrated the importance of getting our cognitive ontology right: consideration of the ethological demands placed on
organisms potentially provides insight into the otherwise bewildering empirical
evidence obtained from tasks meant to probe classically defined cognitive functions.
Appendix E. Theories of Cognition, Use, and Reuse

The issue of defining use and reuse hounds extant theories of cognition that utilize the notion. The failure to define use results in skepticism about whether or not a cognitive theory of reuse, that is, a theory of cognition that features reuse, states that the target systems are actually reused. Many theories of cognition posit various kinds of reuse, as I review shortly. However, many of these proposals are unclear about whether reuses, multi-uses, or banal different instances of use are occurring, though to be fair, their purpose is not analysis of the concept of reuse but rather a proposal for how cognition works; that is, these theories are properly subject to the first set of constraints on an analysis of reuse, but not the second. (The lone exception is Anderson’s analysis of reuse, which is in part philosophical.) Furthermore, some of these theories are specifically neural or neuronal, and others are specifically computational, so they present a broad class of types of reuse. I contend that none of these theories describe the reuse of dynamical mechanisms as posited in my theory, and also that many of the theories equivocate between reuse and multi-use. An exhaustive review of theories of cognition is beyond the scope of this appendix, but the sampled theories will give a flavor for how use and reuse play a role in cognitive theories, as well as illustrating how a concern for dynamical mechanisms and their reuse departs from such theories.
Consider, first, Dehaene’s neuronal recycling hypothesis (Dehaene 2005; Dehaene and Cohen 2007). Dehaene is interested in determining how certain cognitive functions, like reading and arithmetic, can have specific neural substrates given that these cognitive functions are culturally transmitted and far too recently acquired to have evolved dedicated machinery. He suggests that “...the human ability to acquire new cultural objects relies on a neuronal “reconversion” or “recycling” process whereby those novel objects invade cortical territories initially devoted to similar or sufficiently close functions” (Dehaene 2005, p. 148). This recycling occurs during an individual’s lifetime, not at evolutionary timescales, and “each cultural acquisition must find its ecological niche in the human brain, a circuit whose initial role is close enough and whose flexibility is sufficient to be reconverted to this new role” (Dehaene 2005, p. 148).

Dehaene provides three postulates of the neuronal recycling hypothesis. First, “[h]uman brain organization is subject to strong anatomical and connectional constraints inherited from evolution”; second, “[c]ultural acquisitions... must find their “neuronal niche,” a set of circuits that are sufficiently close to the required function and sufficiently plastic” to accommodate the novel use; and third, “[a]s cortical territories dedicated to evolutionarily older functions are invaded by novel cultural objects, their prior organization is never erased”, thus constraining cultural acquisition and mature neural organization (Dehaene and Cohen 2007, p. 385-386). Ultimately, culturally acquired
cognitive functions are bounded by preexisting neuronal constraints in the form of the
plasticity of such circuits and the preexisting function associated with those areas.

Dehaene’s neuronal recycling framework is very clear and interesting, but it is
tied distinctly to cortical/neuronal mechanisms. The notion of reuse that my theory
advocates invokes reuse of dynamical mechanisms, though there may be, in addition,
reuse of cortical mechanisms. Since his hypothesis focuses on the physical mechanism, it
fails to capture the possibility that different physical mechanisms may in fact exhibit the
same dynamics executing some formal model, and thus constitute an instance of reuse.

Setting aside the issue of dynamical mechanisms, Dehaene’s hypothesis regards
the neural mechanisms of culturally acquired cognitive functions. Consider the cognitive
ability to read. Dehaene cites the example of the visual word form area (VWFA), an area
in the left lateral occipito-temporal sulcus, as an example of neuronal recycling (Cohen
and Dehaene 2011). The VWFA is activated most by written words, and less by a wide
range of spurious, word-like stimuli, relative to control stimuli, suggesting that this area
functions as a word detector for reading (Cohen and Dehaene 2004; for recent review,
see Cohen and Dehaene 2011; for dissent, see Price and Devlin 2011). Dehaene and his
group have proposed the Linear Combination Detector (LCD) of VWFA function
(Dehaene et al. 2005; Cohen and Dehaene 2011). The LCD model describes a parallel

\[ \text{Though, as with many of the localized functional brain regions, the location of the VWFA is plastic and can appear on the right side if the left neural region is lesioned (Cohen et al. 2004).} \]
process whereby neurons in the ventral visual stream, starting in primary visual cortex (V1) and ascending up the processing pathway through higher visual cortex, are hierarchically organized and have larger and larger receptive fields (RFs), and these neurons are tuned, or selectively respond, for successively larger stimuli, eventuating in responses to words (Dehaene et al. 2005).

Multi-use threatens. The neural substrate for the capacity to read appears to include this cortical area. But in order to conclude the area is reused, a definition of use needs to be provided. Going on the analysis provided above, a specific biological function contained in that cortical area needs to be specified. Dehaene wants to claim that the current function of the area is the recognition of visual words, a cognitive function that results from a hierarchically organized response to word-like stimuli, the computational model. This computational model is implemented by neurons that have larger and larger RFs, and so the biological function would be the response to visual word-like stimuli by cells in the VWFA.² Cells in the VWFA are taken to respond to word-like features of the environment and then developmentally tuned to respond to words (Dehaene and Cohen 2007, 2011). So, while the same biological function may be used for both objects and words, it is not clear that the function is contributing to

² Though there is tantalizing evidence that the area is multi-modal, or “meta-modal” (Dehaene and Cohen 2011, p. 260), as the VWFA responds in Braille readers relative to a control task (Reich et al. 2011). Note that this use of implementation is distinct from the notion used in my account, and remains unanalyzed.
different computational processes. Furthermore, responses in the VWFA to non-word, though word-like, stimuli are observed. As they note, the “recycling view predicts that reading acquisition should always occur at a reproducible localization in the visual cortex and with a functional specialization for reading-specific processes, although not necessarily with full regional specificity because both word and object recognition may still be intermixed at the same cortical site” (Cohen and Dehaene 2011, p. 256). Unless the prescribed biological function is experimentally determined to reoccur for different computational functions, the area may simply exhibit different instances of being used, and unless the same neurons are active for those different computational functions, the area may just exhibit multi-use.

Next, consider Gallese’s neural exploitation hypothesis (Gallese and Lakoff 2005; Gallese 2008). Gallese and Lakoff construct a general theory of concepts for cognitive neuroscience. They maintain that “conceptual knowledge is embodied, that is, it is mapped within our sensory-motor system” (Gallese and Lakoff 2005, p. 456). They advocate the neural exploitation hypothesis: “the adaptation of sensory-motor brain mechanisms to serve new roles in reason and language, while retaining their original functions as well” (Gallese and Lakoff 2005, p. 456). As a matter of empirical fact, “imagining and doing use a shared neural substrate”: imagining something in a particular sensory modality activates the neural machinery that processes that type of
sensory information, and imagining moving likewise activates the same neural machinery used in actual movements (Gallese and Lakoff 2005, italics removed, p. 456). Furthermore, they claim that “the same neural substrate used in imagining is used in understanding”: the neural areas used in imagining are used in understanding, in that the meaning of a sentence in a context is understood in virtue of imagining that meaning, resulting in an “interactionist theory of meaning” (Gallese and Lakoff 2005, p. 456, italics removed). Gallese expands this reuse to include social cognition, arguing that “key aspects of human social cognition are underpinned by neural exploitation, that is, by the adaptation of neural mechanisms originally evolved for sensorimotor integration, later on also employed to contribute to the neurofunctional architecture of thought and language, while retaining their original functions as well” (Gallese 2008, p. 327). For example, he posits that there are two modes of operation of the premotor system. First, “the circuit structures action execution and action perception, imitation, and imagination, with neural connections to motor effectors and/or other sensory cortical areas” (Gallese 2008, p. 327). With real action, the motor system is activated; with imagination and simulation, it is inhibited. Second, “the same system is decoupled from its action execution/perception functions and can offer its structuring output to non-sensorimotor parts of the brain…. contribut[ing] to the mastering of the hierarchical structure of language and thought” (Gallese 2008, p. 328). So neural reuse clearly
appears in the theory, as the same area—e.g. the premotor system or sensory processing areas—are utilized for multiple different cognitive functions. Thus, their theory is a theory of neural reuse: reuse of the same neural mechanisms for different cognitive functions.

Central to the argument for their view is the role played by mirror neurons and functional clusters. Mirror neurons and “other classes of premotor and parietal neurons are inherently “multimodal” in that they respond to more than one modality” (Gallese and Lakoff 2005, p. 457-458), where as a first pass, mirror neurons are neurons that respond to intentional actions regardless of the actor, be it the organism’s own actions or the actions of another. The multimodality of these responses refers to their responsiveness to signals from across sensory and motor domains. Multimodality “is realised in the brain through functional clusters” such as “parallel parietal-premotor networks”, forming “high-level units” that characterize “the discreteness, high-level structure, and internal relational structure required by concepts” (Gallese and Lakoff 2005, p. 458, italics removed). These functional clusters are used for both mental simulation and for acting and perceiving. The conceptual machinery of the mind in terrestrial organisms like humans is the neural machinery of movement and perception, and in particular, the functional clusters that dictate and govern perception, action, and simulation. These functional clusters are reused for conceptual understanding as well as
for sensorimotor processes. As they say, “rational thought is not entirely separate from what animals can do, because it directly uses sensory-motor bodily mechanisms—the same ones used by nonhuman primates to function in their everyday environments.... Rational thought is an exploitation of the normal operations of our bodies” (Gallese and Lakoff 2005, p. 473).

Setting aside the issue of dynamical mechanisms, which Gallese and Lakoff do not address, the issue of use, reuse, or multi-use arises again. Without a definition of use, it is hard to say whether the sensorimotor circuits that purportedly are reused are in fact reused or merely exhibit multi-use. Like Dehaene, in order to substantiate the claim for reuse, Gallese and Lakoff must specify what functions are being used, and hence reused, for different domains. The neural exploitation hypothesis may entail that areas have multi-use, and that may be the most that they are claiming. For the functional clusters of disparate areas at the heart of their theory, which provide the high-level, rich internal structure of concepts, the use of the same functional clusters for different cognitive functions may be no more than multi-use of a set of neural areas for different cognitive functions. Closer to true reuse are the mirror neurons that partly constitute those functional clusters. However, given the role that they propose for mirror neurons, as responding to more than one modality, it is unclear if the mirror neurons are actually put to different uses in their respective computational structures. The functional clusters
in which those mirror neurons appear are given a specific computational role in underwriting concepts, but the nature of that role varies across different concepts. The contributions that the mirror neurons make to these clusters may not go beyond their multimodal responses—that is, their contribution may be identical across different functional clusters, constituting an instance of use or multi-use and not reuse. (See Goldman 2012 for a defense of the reuse of mirror neurons.)

Perhaps the closest extant view to my view comes from the neuroscientific and biological literature on network motifs (Sporns and Kotter 2004; Sporns 2011). For example, consider Sporns and Kotter’s approach to network motifs in the brain (Sporns and Kotter 2004). Sporns and Kotter take a data-driven, graph-theoretic approach to analyzing the reoccurrence of different patterns of network connectivity across different cortical connectivity datasets. A motif is “a connected graph or network consisting of M vertices and a set of edges… forming a subgraph of a larger network” (Sporns and Kotter 2004, p. 1916). Each set of motifs with M vertices (“M set”) has a set of motif classes, corresponding to the different ways the vertices can be connected. A structural motif is “composed of a specific set of M vertices that are linked by edges” and larger networks can be “structurally assembled from a finite set of such motifs” (Sporns and Kotter 2004, p. 1916). Functional motifs are constructed from parts of structural motifs and “consist of the original M vertices of the structural motif, but contain only a subset
of its edges” (Sporns and Kotter 2004, p. 1916). All of these definitions map on to brain areas in specific ways, and reflect the organizational properties of the mechanisms that support cognition. A motif frequency spectrum results from “[s]orting all possible structural motifs within a network as a function of motif class”, recording the number of distinct motifs in each possible structural motif class. For example, a 3 node motif has 13 motif classes, constructed from the different possible vertices that interconnect all 3 nodes. Given this motif frequency spectrum, the motif number is “the total number of distinct occurrences of any motif of size M, and the motif diversity… [is] the number of classes that are represented within the network by at least one example” of that class (Sporns and Kotter 2004, p. 1911). Using these tools, they are able to assess the motif properties of large networks. To assess the presence of various motifs and their properties, they examined datasets consisting of macaque visual cortex, macaque cortex, and cat cortex. They found that the large-scale networks tended to have low structural motif number and high functional motif number. All networks showed maximal functional motif diversity and submaximal structural diversity.

This approach is a theory that supposes reuse of the mechanisms for cognition and, as the connectivity matrices analyzed are neural datasets, an instance of a neural reuse theory. As they say, “as networks become more complex, already existing simpler

---

3 They also looked at a neuronal connectivity matrix for C. Elegans, but only to contrast the results with the cortical connectivity matrices.
networks are largely preserved, extended, and combined…. Complex and highly evolved networks arise from the addition of network elements in positions where they maximize the overall processing power of the neural architecture...” (Sporns and Kotter 2004, p. 1910). These motifs are the basic abstract components that get reused for different functions. Thus, their theory posits mechanistic reuse at a level more abstract than the neural, as the same motifs can reoccur for different neural datasets. As they wish to analyze the cognitive information processing architecture of the system, the motif approach is also a cognitive reuse theory.

Though spiritually akin to the dynamical mechanism approach, their view suffers from the same problems as those surveyed: overlooking the importance of dynamical mechanisms and failing to justify claims of reuse as opposed to multi-use. First, they do not define the use of a particular motif. The way they define functional motifs—as the set of structural motifs included in the presence of some other structural motifs—is not truly functional, specifying what function a motif performs in a system. The activation of a functional motif may occur for different information processing operations without entailing that the same function is being executed in each such instance. Their account does not say how information processing occurs: what signals are getting transformed or how they are getting transformed. In short, they don’t say
what the function of these functional motifs are, and so can’t assess claims of reuse (though, like the other theories discussed, claims of reuse per se are not their concern).

The breadth and variety of theories of cognition that prima facie feature reuse is overwhelming and beyond the scope of any brief review. However, I hope I have motivated the need to take philosophical care in defining use. The science of cognition can proceed largely independently of this philosophical analysis, especially as regards the physical mechanisms that result in cognitive phenomena. But, insofar as the reuse of cognitive mechanisms may be an indicia of regularities in the architecture of thought, regularities that a science of cognition ignores only at the peril of being incomplete, defining use, in order to identify reuse, becomes a central scientific and philosophical concern.

Against the list of criteria for a philosophical analysis of reuse, how do extant analyses fare? The principal—and as far as I can determine, only explicit—philosophical analysis of reuse is Anderson’s analysis of the concept (Anderson 2007a, 2007b, 2007c, 2008, 2010, 2014). I will argue that, as with the theories of cognition just assessed, Anderson’s analysis of reuse ignores the role of dynamical mechanisms in cognition and fails to provide an adequate definition of use.

In multiple publications, Anderson develops a particular theory of reuse, the massive redeployment hypothesis (MRH), as well as analyzing the concept of reuse as it
occurs across different theories of cognition. Brain areas work together to execute
cognitive functions, and are “therefore not generally deployed in support of only a
single function, but are instead redeployed in many different functional complexes, which
do many different... things” (Anderson 2007a, p. 330-331). Anderson notes “cognitive
functions typically have several necessary participants” but that also “individual brain
areas can be participants in several cognitive functions” (Anderson 2007b, p. 148). He
elaborates: “(i) a typical cognitive function requires the participation of more than one
brain area, and (ii) each brain area may be a participant—may be redeployed—in
support of other cognitive functions” (Anderson 2007b, p. 148). This many:many
mapping between cognitive functions and brain areas forms the basis for Anderson’s
claim of reuse.

To get clearer about redeployment, Anderson notes two different interpretations
of (ii). First, “the redeployed brain area does the same thing... in each instance of
redeployment, and differences in function are the result of differences in the structure
and dynamics of the functional complex as a whole” (Anderson 2007b, p. 148).
Individual brain areas have a particular function, and differences in cognitive function
result from differences in the way that individual neural functions are formed into a
complex. Anderson elaborates on this idea of reuse. Suppose that “component c
computes function f, and that it does this because it and its participants compose a
circuit of a particular description…. [W]e have an easy sort of story to tell about how component d, which shares some of c’s participants, can compute a different function, g, just so long as each participant, although doing the same thing, is put into such relations with other participants so as to produce a different outcome” (Anderson 2007b, p. 151).4 Second, ”the area does something different in each case of redeployment…” (Anderson 2007b, p. 148). Individual brain areas have different functions, and the areas contribute these different functions to the complex. Anderson views cognitive functions as resulting from the execution of component functions, and different components can contribute to different such cognitive functional complexes, resulting in reuse. Furthermore, different neural areas work in concert to form those components. Reuse manifests throughout the whole system, as the view permits many-many relations between cognitive functions and component functions, and between component functions and areas.

Anderson highlights four commitments and predictions of MRH, however the relevant prediction for the present discussion is the redeployment thesis that “redeployed areas play the same “role” in each of the functional complexes they

---

4 This is the technique of design reuse. Anderson argues that a cardinal rule of re-use is “never to break prior functionality” and to preserve backwards compatibility. “For suppose that one can get function g to work by changing the role of participant a, shared by component c. By hypothesis, this means changing its physical/functional properties, thereby altering the functional characteristics of component c, which introduces the possibility that it will no longer compute function f…” (Anderson 2007b, p. 151-152).
support” (Anderson 2007a, p 332). In order to construct arguments for the redeployment thesis, he assumes first that “the functional properties of a neural circuit are determined by its configural properties, such as the number, strength and topology of its connections. Then it follows that a given neural circuit, in a given configuration, does some specific functional thing… when it is activated…. It is only when the configuration of these things changes that they can be said to be doing something different” (Anderson 2007a, p. 336). In particular, “[s]o long as the configuration of a neural circuit remains fixed, we should say… that it is doing the same specific thing whenever activated” (Anderson 2007a, p. 336). Call this the configuration assumption.

The issue of dynamical properties confounds some of Anderson’s claims and in particular the configuration assumption. The configuration assumption maintains that neural functions contributing to cognitive ones only change when the physical configuration of the neural area does. The issue of dynamical properties of neural areas challenges the configuration assumption. On my view, nothing about the system per se needs to change to adapt the dynamics to the processing problem. Functional differences, understood as changes in a neural area’s function, can arise from more ways than Anderson imagines, once the dynamical properties of a neural area are considered.

---

5 Anderson provides several arguments for (2), but certain foundational issues remain unresolved that preempt these arguments. Since the purpose herein is not to criticize Anderson’s arguments, but rather to lay out the basic issues prerequisite to a theory of reuse, I will not be addressing the arguments individually.
Functional differences can arise from different subsets of dynamical properties being relevant for executing some formal model. Though Anderson does not address the role of dynamical mechanisms in cognition, Anderson notes that “it could be that the dynamic response properties of local circuits are fixed, and that cognitive function is a matter of tying together circuits with the right (relative) dynamic response properties” (Anderson 2010, p. 265, italics in original). However, there is no reason to suppose that the dynamic responses are fixed. Regardless, exploring the flexibility afforded by the dynamical approach will helpfully illustrate the various possible types of reuse on that approach.

A second set of objections relates to defining use and cognitive function. Anderson does not justify the set of cognitive functions, what exactly is a neural use or function, and fails to distinguish between neural, cognitive and other notions of function. Hence, the arguments he constructs and the evidence he cites for neural reuse are ambiguous. First, he assumes that cognitive functions are individuated classically, that is, according to something like a derived or refined faculty psychology, common in cognitive psychological treatments of the mind and neuroimaging work on the brain. Anderson is sensitive to this problem, noting that the “issue of how to classify and decompose cognitive functioning is indeed thorny…” (Anderson 2008, p. 246). This thorny issue, however, is central to the claim that neural areas are reused for different
cognitive functions. He uses behavioral tasks as a probe for cognitive function. But these tasks require an analysis of the processing problem underlying the task. More tasks activating more regions do not entail that the region has more diverse cognitive functions with which it is associated, if the processing analysis of the problem facing the organism results in some aspect of that processing that is the same across tasks.

Experimental evidence of neural reuse may in fact indicate that the cognitive function probed by a set of experiments is in reality a functional complex, and the prima facie indication of reuse in fact reflects a singular contribution of a neural area to a cognitive function.

Let a cognitive ontology be the set of basic cognitive functions, and a task analysis be the division of a behavioral experiment into the component cognitive operations drawn from the cognitive ontology (Price and Friston 2005; Bilder et al. 2009). In reviewing the activation of a given area in many different tasks, he addresses the problem of defining tasks, noting that “[i]f cognitive scientists are very bad at categorizing their experiments…”—that is, at giving a task analysis—“that could explain the simple finding that regions are activated by multiple tasks, because some experiments that belonged in one category would have instead been place in another” (Anderson 2010, p. 265). Anderson is sensitive here to the misclassification of tasks

---

6 See Appendix D for much more discussion of this issue.
according to our cognitive ontology. However, more radically, the entire cognitive ontology may have to be thrown out, as incapable of providing a task analysis that captures the natural division of cognitive functions. Either objection would threaten the evidence suggesting reuse. In the first case, results that appear to indicate reuse in fact reflect misclassification, and in the second case, results that appear to indicate reuse in fact reflect the failure of the utilized tasks to probe isolated cognitive functions.

Against this, Anderson levels two points. First, even if this were true, and he doesn’t doubt that designing effective experiments that capture the underlying functions is difficult, that wouldn’t cast doubt on other predictions made by the MRH (Anderson 2010, p. 265). This is false, and to illustrate why, suppose that what is required an entirely novel cognitive ontology. The reply fails to respect how the objection threatens the very idea of reuse, which is by claiming that evidence of reuse is in fact an indication of the failure of the task to probe a single cognitive function.

The other predictions made by the MRH, that more recently evolved cognitive functions will recruit a more spatially diverse neural coalition and that evolutionarily older neural areas will participate in more cognitive functions (see Anderson 2007a, 2007b, or 2010), both require psychological tasks that carve cognition at its creases. If they don’t—that is, if psychological tasks typically used in cognitive neuroscience fail to

---

7 A mix of misclassification and reconceptualization may also occur.
probe isolated cognitive functions—then the possibility remains of some task analysis decomposing the tasks used into a suite of cognitive functions that result in some bland one structure—one function mapping. This threatens both predictions. On such a novel cognitive ontology, more recently evolved cognitive functions may in fact be more complicated functional complexes, composed of many basic cognitive functions; determining neural activity related to such behavior would result in many areas being activated even though each area contributes to a sole cognitive function. And, on such a novel cognitive ontology, evolutionarily older neural areas may participate in more cognitive functions simply because they execute more easily evolved singular cognitive functions that are also the ones more often used in putting together functional complexes. Furthermore, a processing analysis of the tasks may result in many such processing models needing simultaneous execution. Widespread activation of an area across many tasks could reflect the coordinative requirements of such widespread execution of many different models. Thus, the possible need for a novel cognitive ontology threatens both predictions.

Second, he and Chemero have run a clustering analysis to see if there is a way of dividing experiments into different groups so that neural activations don’t overlap (Anderson 2010, p. 265). This re-analysis of the data failed. Of course, on the second form of the objection, this second reply is moot: the idea is not that the experiments can
be reclassified so that neural activations don’t overlap, but rather that the tasks
themselves are designed such that they force the system to invoke multiple functions
distributed throughout the brain to accomplish the goal of the task, that is, to produce
adaptive behavior. Without a motivated cognitive ontology that justifies the analytical
decomposition of tasks into a set of cognitive functions, the possibility of
misclassification or reclassification threatens all of the evidence for reuse.

At the heart of a cognitive ontology is the concept of a cognitive function. For
reuse to be true of the brain, cognitive functions must be distinguished from brain
functions, and the use of brain functions for cognitive functions defined. Anderson
defines a cognitive function as “...a process of cognitive or psychological interest... that
can be specified in terms of inputs, outputs, and the... transformation of the former into
the latter” (Anderson 2007b, p. 144). He calls this a pragmatic definition, and is meant to
appeal to whatever fills the boxes in a black-box diagram of a cognitive function. He
continues, “a brain area participates in a function if activity or processing in that area
supports the transformation of inputs to outputs that define the function. A given
function may... have more than one participant....” (Anderson 2007b, p. 144). A
collection of such participants is a functional complex. He defines a brain function as “a
specific, identifiable mechanism implemented in the brain for transforming signals
and/or information from one form to another”, including such processes as “filtering,
smoothing, integrating, enhancing, shifting waveform or periodicity, interpolating, and the like. Information-processing functions might be described in mathematical terms (adding, subtracting) or in terms of data processing (store, search, sort)” (Anderson 2007b, p. 161). He argues that the relationship between cognitive function and brain function is likely to be many-many. The functions of various areas often interact, such as possessing inhibitory interrelations, such that “a brain function is in fact often the product of cooperation between pieces of brain matter… playing distinctly different roles” (Anderson 2007b, p. 162). Thus, a brain function is a “component function” and the “term role, or area role” is used to denote “whatever it is determined that a particular, contiguous bit of brain matter actually does” (Anderson 2007b, p. 162). For anatomical parallels, he defines a functional complex as “the entirety of participants in a cognitive function” and a component is “the sum of the participants that implement a component function” (Anderson 2007b, p. 162).

In describing the distinction between the functions of local circuits and cognitive functions, Anderson invokes the language of “workings” and “uses” from Bergeron (Bergeron 2007). On neural reuse theories, “most of the interesting cognitive work is done at higher levels of organization, but they also emphasize that local circuits have specific and identifiable functional biases. In general, these models make a strong distinction between a “working”—whatever specific computational contribution local
anatomical circuits make to overall function—and a “use”, the cognitive purpose to which the working is put in any individual case. For neural reuse theories, anatomical sites have a fixed working, but many different uses” (Anderson 2010, p. 252). Anderson emphasizes that “workings” and “uses” are not the same; in short, a working is “whatever single, relatively simple thing a local neural circuit does for or offers to all of the functional complexes of which the circuit is a part” (Anderson 2010b, p. 295, italics in original).

However, there are at least four distinct notions of function at work: behavioral, computational, mechanistic and neural. He distinguishes behavioral and neural; those are his ‘workings’ and ‘uses’. But computational function is also critically important to the system. Computational functions can be taken to be the behavioral functions, albeit more well-specified, describing not just the output behavior but also the processing operations that result in the behavior. Computational functions can also be taken to execute the qualitative functions cited, for example, in classic faculty psychology (Reid 1785) or cognitive psychology (Miller 2003). But there are also mechanistic functions, the abstract functions of the physical device that are properly the result of the device’s organization. In my theory, these are the functions of the dynamical mechanisms. The physical device will have particular physical functions, such as neural functions, that

---

8 I’m not here attempting to provide a typology of types of function. Rather, I’m just arguing that the ambiguity in the notion causes trouble for Anderson’s argument.
result in the mechanistic ones. But different neural areas can have the same mechanistic function. So even if a neural area is not necessary for a behavioral function, a mechanistic or computational function may be. This issue of function is crucial to assessing the empirical evidence Anderson cites in support of reuse. For example, he justifies looking at studies that utilize the cognitive functions licensed by cognitive psychology by reasoning that if “one were to compile a number of these studies in various task domains, one could ask, for each region of the brain, whether it supported functions in multiple domains, and whether such reuse was typically limited to regions of the brain implicated in supporting” certain tasks (in the relevant case he discusses, sensorimotor tasks) (Anderson 2010, p. 258). But the critical slide is from studies in various task domains, which are classes of behavioral function, to functions in multiple domains, which is ambiguous between computational, mechanistic and neural function.

Finally and most fundamentally, Anderson never defines the use of a neural area. He does offer several arguments for seeing reuse as the appropriate concept, and not multi-use, including that more evolutionarily recent uses are subserved by more spatially spread out neural areas than older uses; cognitive interference and cross-domain effects suggest reuse and not multi-use; and functional or semantic inheritance again suggests reuse and not multi-use. Constancy in local workings, though, is the only sure way to establish reuse, as Anderson notes (Anderson 2010b, p. 298). But to infer
that the same local working is being utilized over and over requires a definition for what it is for a working to be utilized in the first place.

In sum, though Anderson presents a very careful case, well-buttressed by the neural evidence, for the reuse of neural areas, his analysis is beset by multiple problems. Anderson fails to motivate the cognitive ontology used in investigations in cognitive neuroscience. He also fails to define and distinguish neural functions and cognitive functions. Finally, Anderson does not define the use of a neural area, and hence reuse remains unanalyzed. Furthermore, though he acknowledges the importance of the dynamics of neural circuits and how these circuits can reconfigure into different functional complexes to support different cognitive functions, he neither recognizes the distinctness of dynamical mechanisms nor realizes how the dynamics of various areas provide more ways for neural areas to be reused than mere consideration of the physical mechanism permits.

Before moving on, most recently Anderson has developed a much more nuanced position that closely reflects my own (Anderson 2014). On the issue of cognitive function, he notes “we in the cognitive neurosciences have been... captured and possibly limited by a specific taxonomy of mental function... inherited from cognitive psychology, which... is... likely motivated by concerns, desiderata, and... even metaphysical assumptions quite different from—possibly even incompatible with—
those that are most illuminating for an investigation into the functional organization of the brain” (Anderson 2014, p. xii). However, he advocates trying “explicitly to give the brain its scientific voice—to let it show use what aspects of its world it is in fact attuned to” which involves “significant revision to the vocabulary of cognition, the way we categorize and label experiments and mental activities” (Anderson 2014, p. xvi). So, though sensitive to this concern, his approach would fundamentally shift the cognitive ontology away from the behavior of the organism, allowing the brain to contribute to this classificatory program. But the problem is not that the brain has no voice in constructing this ontology; rather, the objection is that cognitive function itself remains an undefined concept.

He is similarly sensitive to the issues surround neural functions. “[I]n my initial formulation of the massive redeployment hypothesis (Anderson 2007a; 2010b) I was also trying to preserve the notion of low-level functional components in the form of “workings”, while also trying to do justice to the complex functional profiles of local regions….In considering these issues, I have come around to the idea that we need to give up on the notion of componentiality in the brain…. It appears that in the brain we have Transiently Assembled Local Neural Sub-systems (TALoNS). TALoNS—which are the constituents of large-scale functional networks—exhibit temporary, reproducible functional selectivity, but do not have the normal functional characteristics of
components” (Anderson 2014, p. 70-71). Or: “But given the picture painted above—of temporally dynamic networks with multiple physical connection states and physical and functional boundaries made fuzzy by the diffusion of neurotransmitters to synaptically unconnected neurons...—I myself have become pessimistic that it will generally be possible to identify the specific neural mechanisms (the “working parts” (Craver 2007)) to which we can assign cognitive operations” (Anderson 2014, p. 81). Anderson divines radical implications of such a view, arguing that it is “more scientifically prudent to devise alternate models for understanding structure-function relationships in the brain—models that do justice to the cross-domain activity profile of a typical region, but don’t require the assignment of specific, individual cognitive operation” (Anderson 2014, p. xv).

Despite these concerns, Anderson still adheres to a restated reuse hypothesis, characterized by two principles of reuse: “(1) functional differentiation leading to local functional biases; and (2) reuse of regions in multiple cognitive contexts” (Anderson 2014, p.69). These functional biases are “…a set of dispositional tendencies that capture the set of inputs to which the circuit will respond and govern the form of the resulting output”, crucially allowing for neural areas to exhibit different dynamics depending on not only their inputs but also their effective connectivity (Anderson 2014, p. 12). The context within which circuits operate can change, referred to in the literature as neural
modulation, and “while some of the observed functional diversity is due to the fact that a given circuit in a given configuration is often useful in multiple contexts, some of the observed diversity is likely also due to the fact that the local network can be in multiple different states” (Anderson 2014, p. 12, italics in original). Specifically, Anderson cites a range of examples of neuromodulation from across the animal kingdom (Anderson 2014, p. 24ff), agreeing with Bargmann (Bargmann 2012) that “[a]ny given circuit will have a number of possible uses, only some of which are available at any given moment depending on the neuromodulatory state of the organism. The other uses will be “latent”; part of the set of functional possibilities afforded by the physical structure of the circuit, but not currently expressed in its functional state” (Anderson 2014, p. 27). For Anderson, this militates against the assignment of particular neural functions, in the sense that “it would be worth our while to see how much science we can do with the weaker notion of functional differentiation in the brain, before engaging in the kind of abduction to specific, local computational operations…” (Anderson 2014, p. 12, italics in original). Thus, Anderson’s position has evolved. Whereas before neural areas could be reused, whatever that ultimately might amount to, but the circuits themselves have something like a fixed function, now neural areas could be reused but only in virtue of neuromodulatory influences creating conditions that allow the expression of different latent functional states of the circuits or areas.
Despite this evolution, Anderson’s view is still beset by unresolved issues. These neuromodulatory influences are central to the set of dynamical properties that potentially contribute to a dynamical mechanism on my theory of cognition. As such, Anderson has moved conceptually much closer to the view I have laid out. The dynamics of neural areas are now central place in his proposal for the brain’s functional architecture. Nonetheless, he does not analyze dynamical mechanisms, retaining the ‘brain’s eye’ view throughout. Furthermore, my view doesn’t require the expression of latent functional states of circuits, as the same functional state of the physical mechanism may contribute different subsets of its dynamics to the dynamical mechanisms that execute the cognitive function. Note though that Anderson still has not defined use, regardless of the object of analysis, and so an ultimate determination on reuse or multi-use remains out of reach. Thus, though much closer in spirit to my theory, Anderson neither defines use and reuse, nor explores the consequences of the dynamical turn in the same way that I do, either for a general theory of cognition or for understanding reuse.

While it would be impossible to survey all of the extant theories that posit some degree of some type of reuse, I hope I have made at least an initial case both that many theories of cognition equivocate on reuse and multi-use and that the central role of the reuse of dynamical mechanisms has been ignored. Furthermore, of the philosophical
approaches to reuse, focusing on what precisely constitutes use has been neglected. Only once use has been well-defined can reuse be reliably attributed to a particular cognitive mechanism. And on my theory, use is well-defined, determined by some component mechanism performing a function that aids in the execution of a formal model of processing.
References


362


364


Biography

David Leo Barack was born on December 26, 1981 CE, in Highland Park, Illinois in the United States of America. He attended Pitzer College, graduating with honors with a BA in Consciousness Studies in May 2004. He graduated from the University of Wisconsin - Milwaukee with an MA in Philosophy in May 2007. After studying for a year at Northwestern University, completing only the first year of a two-year post-baccalaureate pre-medical certification program, David matriculated to Duke University in August 2008. In addition to his philosophical work, David maintains a separate research program in the neural mechanisms of decision making in primates.