ON THE ORIGINS OF THE SYNTHETIC MIND: WORKING MODELS, MECHANISMS, AND SIMULATIONS

BY

PETER M. ASARO

B.A., Illinois Wesleyan University, 1994
A.M., University of Illinois at Urbana-Champaign, 1996
M.C.S., University of Illinois at Urbana-Champaign, 2005

DISSERTATION

Submitted in partial fulfillment of the requirements
for the Doctor of Philosophy in Philosophy
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2006

Urbana, Illinois

© 2006 by Peter Mario Asaro. All rights reserved

Abstract

This dissertation reconsiders the nature of scientific models through an historical study of the development of electronic models of the brain by Cybernetics researchers in the 1940s. By examining how these unique models were used in the brain sciences, it develops the concept of a "working model" for the brain sciences. Working models differ from theoretical models in that they are subject to manipulation and interactive experimentation, *i.e.,* they are themselves objects of study and part of material culture. While these electronic brains are often disparaged by historians as toys and publicity stunts, I argue that they mediated between physiological theories of neurons and psychological theories of behavior so as to leverage their compelling material performances against the lack of observational data and sparse theoretical connections between neurology and psychology. I further argue that working models might be used by cognitive science to better understand how the brain develops performative representations of the world.

I dedicate this dissertation to my grandparents,
Thomas and Selma Sinks, and John and Marcie Asaro,
and to my parents Larry and Mary Lou Asaro.

Acknowledgments

I wish to thank all of my teachers over the years, who helped to guide me through the many steps of this journey. I am deeply grateful to the efforts of Mick and John Ashby for working with me for three long years to convince their mother and aunts to deposit the notebooks of W. Ross Ashby in the British Library, and to John for making digital scans while they still had them, and sharing these with me. Thanks to Herbert Brün for teaching me the cybernetics that cannot be communicated. And I would also like to thank my committee for never losing faith in me, and my mentor Andy Pickering for showing an unwavering interest in discussing all things cybernetic. Special thanks go to Diana Mincyte for her limitless supply of moral support.

Table of Contents

List of Figures

Chapter 1
On the Origins of the Synthetic Mind: Working Models, Mechanisms, and Simulations

Metaphors may create realities for us, especially social realities. A metaphor may thus be a guide for future action. Such actions will, of course, fit the metaphor. This will, in turn, reinforce the power of the metaphor to make experience coherent. In this sense metaphors can be self-fulfilling prophecies.

-George Lakoff & Mark Johnson

Prior to about 1940, mechanisms (and the "classical" physics that thought about them) were of a "cause-effect" type in which typically one cause led to one effect, then the process was complete; the wound watch ran for 24 hours, then stopped; the lathe, switched on, ran round endlessly; the typewriter, when a key is pressed, printed the letter and stopped. With these machines before him, the psychologist theorized similarly: stimulus elicits response, stop; a dog is subjected to a cycle of flashes and reinforcements and it develops a conditioned reflex, stop. More complex theories of behavior could not be developed because no one knew how to think about complex behavior.

-W. Ross Ashby

As soon as an Analytic Engine exists, it will necessarily guide the future course of science.

-Charles Babbage

1.1 Introduction

Every good philosophical inquiry begins with questions, and proceeds by what Martin Heidegger might call a questioning of those questions. "What is a good scientific model of the brain?" and "How does the brain model the world, if at all?" are two questions which I believe hold a powerful resonance with one another. Despite the fact that one is generally considered in the subdiscipline of the philosophy of science, and the other in the subdiscipline of the philosophy of mind, answers offered for these questions have had many interesting historical and theoretical connections to each other. Through the course of my research I have sensed that there are two general approaches to thinking about the role of models in science, and hence about what makes a good model. There are also two general approaches to thinking about how the brain might model the world, and an interesting analogy can be made between models in science and models in mind. I say "general" here because there are many approaches taken to these questions, as well as many definitions of "models." Specifically, I will be concerned with the general distinction between "theoretical models" and "working models." Within science, both types of models have long been used, but there has been a more recent development in the history

and philosophy of science toward understanding the role of working models, which had long been neglected and has yet to be fully articulated. There has also been a growing interest in alternative approaches to thinking about how the brain models the world within cognitive neuroscience. Yet, there have been only fragmentary accounts of how mental models might be more like working models than like theoretical models–as they are often construed. The aim of this dissertation is to re-examine the various aspects of scientific models and brain models and to organize them into an outline of how we might think about mental models as being working models. In doing this, I hope to illuminate the ways in which the production of knowledge in science and the ability of the individual brain to learn to cope with the world might be unified under a pragmatic account of working models, as they have been under a unified account of theoretical models.

Another central aim of this dissertation is to examine the relationship between these questions of modeling and the brain with a special concern as to their relation to a third question: "In what ways is a computer a model of the brain?" What is interesting about this third question is that the computer was designed after the brain, and is the central tool for building all sorts of models in science, and is central in the building of various forms of neuro-cognitive models. It would be difficult to write a complete history of the brain sciences in the 20th century that did not discuss the impact of ideas from cybernetics, information theory, and computation–areas of engineering that do not always have immediate and obvious connections to neuroscience or psychiatry. Similarly, a history of the computer that failed to mention the influence of the brain, psychology, and artificial intelligence on the development of the computer, especially in the first decades, would be woefully incomplete. The development of our scientific understanding of the brain, and our technological understanding of computation have been intimately bound up with each other over the course of the 20th century. This complex relationship has not gone unnoticed by some historians, and has even been a central premise of a handful of scientific and engineering disciplines which seek to understand the brain as a computer, and build computational models of the brain and intelligent behavior.

Indeed, whether the mind ought to be characterized as a computer or not remains one of the central debates in the philosophy of mind. This dissertation offers a new perspective on this

long and complex relationship. It simultaneously seeks to refine our understanding of modeling practices and their uses in cognitive science by coming to grips with the historical interplay of modeling practices between the brain sciences and computer technologies. Ultimately, I suggest that thinking about mental models as feedback-controlled regulators offers a way of seeing mental representations as being *performative*, rather than *symbolic*. This way of looking at the interaction of the brain and the world cuts across several contemporary debates in the philosophy of mind over mental representation and the place of the brain in the world, including Connectionism, Situated and Embodied Cognition, and Dynamicism, as well as theories of mental content. While the notion of performative representation that I develop is both compatible and amenable to recent cognitive movements such as dynamicism, I argue that getting beyond the computational approach will require rethinking the mind as being performatively engaged with its environment.

From its creation, the relationship between the computer and the human mind was recognized as a deeply significant one, even if not fully understood. The first computers were dubbed "giant brains"(Berkeley 1949) for reasons simultaneously affectionate, fearful, technical and speculative. Far from being merely the product of sensationalist journalists or self-promoting scientists and engineers, there were many significant theoretical and empirical connections being drawn between computation, information, and the mind/brain[1] at this time. This relationship was not always invoked unreflectively either, but was the frequent subject of philosophical reflection and theorizing.

In many ways, the computer was *the* technology of mind/brain. Computation was simultaneously a metaphor used to understand how the brain works, a mathematical theory itself derived from reflections upon how people perform certain algorithmic mental processes, a theory of how the neurons of the brain operate, and a technology by which artificial minds could be synthesized and experimented upon. Less often noted is the extent to which the ideas of

[1]The relationship between mind and brain can be quite contentious, and is in part what is supposed to be explained by appealing to the computer analogy. While I try to be careful to refer to mind or brain specifically, I will often use "mind" when discussing the abstract organization of the brain as opposed to the biological organ of the "brain." Nearly all of the historical and philosophical views that I will consider can be classified as materialist in the sense that they equate the mind and brain, making the use of "mind/brain" simpler, if uglier, in many cases.

psychiatrists and physiological theories of the operation of neurons influenced the design of early computers, and in particular memory systems and programmable logic. The development of computational technologies and their relation to the scientific understanding of the mind/brain grew up together and strongly influenced one another during the 1940s and 50s. This relationship continues today, through a complicated and ongoing series of interactions between technological advancements and theorizing about the nature and structure of the mind/brain and its operations.

Often cited as being the philosophical core of this history is Computationalism–the theory that the mind/brain is essentially a computer. Though Computationalism has been expressed in different ways, it has been frequently argued to be the foundation of the fields of Artificial Intelligence (AI) and cognitive science. Since the idea was first presented by Alan Turing in the pages of *Mind* in 1950, it has also been a central issue in the philosophy of mind. Over the past two decades, Computationalism has been the target of several well-known critiques and challenges, including those posed by Connectionism, Situated and Embodied cognition, and Dynamicism (Eliasmith 2003). These critiques each attack Computationalism from a slightly different angle, and each can be seen as challenging the conception of symbolic representation that Computationalism depends on. Connectionism stands as a direct challenge to the atomic nature of representation in Computationalism. Situated and Embodied cognition challenges the shortcomings of Computationalism in accounting for the locus of the representations and processes of computation and cognition. And finally, Dynamicism challenges Computationalism's fundamental ontological assumptions about time, discreteness, and symbolic representation, as well as its epistemic assumptions about what it means to model a system scientifically. It is now becoming widely accepted that Computationalism is flawed in various ways (Smith 2002), although it has inspired a great deal of significant research. Similarly, its challengers are aligned with new technologies that offer intriguing theoretical suggestions, but appear to fall short of comprehensive theories of mind. Thus recent interest in Computationalism has focused on the search for a viable successor theory–a next generation Computationalism (Scheutz 2002). I believe that any such successor theory of mind would benefit from a consideration of the nature of working models and performative representations, both

methodologically and theoretically.

I further believe that it is neither trivial nor coincidental that all three of the major critiques of Computationalism have benefitted in various ways from a reconsideration of the historical development of the computer, AI and cognitive science as they emerged from the cybernetic era of the 1940s. Recent attempts at a synthesis of Computationalism and its challengers (*e.g.,* Eliasmith 2003) have also relied upon some key ideas from cybernetic theory. These varied attempts in philosophical quarters to draw lessons from this historical era have so far stopped short of a systematic study of the early work on synthetic brains. Instead, these critiques take for granted current formulations of the issue, and seek either support for, or early suggestions of, specific critiques of current approaches from this history.[2] By going back to reconsider key aspects of the development of synthetic brains and the ideas of those who sought to build working models of the brain and mind in the 1940s, I hope to provide a new perspective on the central concepts, models, metaphors, and technologies of Computationalism and its challengers, and moreover to provide a re-examination of its scientific methodology and epistemology.

My own view of how we should think about models and the brain steers closer to the ideas originally pursued by cybernetics. The relationship between cybernetics and early computation is a complex and difficult one. As William Aspray poses the question in his history of von Neumann's work on computation:

> Given their contemporaneous appearance, what is the relationship between computing and cybernetics? The evidence suggests that the computer stimulated the growth of cybernetics. For example, Wiener credits his wartime design of computing equipment as a source of his cybernetic ideas; and the computer, a machine designed specifically to process information and control systems, is at the very least a powerful ideograph of the cybernetic program. The relationship also cuts the other way. Many of the powerful concepts of control, statistical information, and neural networks that Wiener appropriated

[2]For example, Connectionists have recently shown interest in a previous unpublished manuscript of Alan Turing which describes his own novel approach to "connectionist" networks (Teuscher 2002). The pioneer of situated robotics, Rodney Brooks, often cites the work of W. Grey Walter as a very early example of reactive autonomous robots, though they were largely forgotten by the research community in the intervening years (Brooks 1999). The leading theorist of Dynamicism, Timothy van Gelder, developed his theory around a core metaphor of Watt's steam engine governor, the icon of a feedback control mechanism exalted by Norbert Wiener in his *Cybernetics* (1948), and acknowledges the work of Wiener and W. Ross Ashby.

> for cybernetics were ones that von Neumann and others subsequently used for computer science. Von Neumann worked in the interdisciplinary spirit of cybernetics, though his final results fell more squarely in computer science and neurophysiology. (Aspray 1990, pp. 210-211).

This dissertation seeks to elaborate on several of the more interesting aspects of this complex relationship between computing and cybernetics, and in particular the ways in which they each sought to build synthetic minds. I pursue this understanding by looking to the writings of those who sought to build working models of the brain and mind in the 1940s. This analysis focuses much of its energy on the writings and machines of three researchers who worked at the intersection of brains, minds, computers, and cybernetics. The first of these is W. Ross Ashby–a British psychiatrist, a founding father of Cybernetics, and author of *Design for a Brain* (1952) and *Introduction to Cybernetics* (1956). This focus is unusual in that he is not often given the scholarly attention that the other subjects have received. I believe that Ashby's work amounts to a carefully considered, technologically grounded, and systematic theory of the brain based primarily in information theory and psychiatry that departs in significant ways from Computationalism. Ashby also made a careful consideration of the role of models in science. In addition to Ashby I consider the synthetic brains developed by Alan Turing and John von Neumann, two mathematicians influential in the development of the mathematical theory of computation as well as the modern electronic computer. As such, they are most often given credit for the theoretical and technological foundations of Computationalism. Despite their differing perspectives, all three men clearly expressed their interest in building a synthetic mind as a leading motive in the construction of their machines. What we gain by re-examining their work is a new sense of the connections between building synthetic minds, our theories of the relationship between mind and brain, and mind and world, and how their experimental and technological practices were bound up with the ways synthetic minds were treated as scientific models.

1.2 Different Kinds of Models

Something that quickly becomes apparent in trying to understand the development of brain models in the 1940s is the many different uses and senses of the word "model." One aim of

this dissertation is to begin the work of distinguishing the various types of models that scientists use. One type will prove to be particularly important, and that is the working model. In this section, I wish to review some of the more common types of models and modeling practices, and give a general formulation of the notion of a working model.

The things scientists call "models" range from abstract concepts to sets of mathematical equations, to experimental laboratory setups, to scale-models, to computational simulations, to ideal case studies. For example, the "Bohr model of the atom" is an abstract idea that electrons orbit the nucleus of the atom in the way that planets orbit the sun[3] (itself a planetary model) combined with a notion of energy shells that hold the electrons at different distances from the nucleus. This model is often given a visual form, and it is a powerfully intuitive model for pedagogical purposes. Contrast the Bohr model to "animal models" in the medical sciences, *e.g.*, a "transgenic mouse model" in oncology. These models are genetically controlled strains of living mice which reliably reproduce cancerous tumors that are of scientific interest. While this might be the result of an isolated gene, the mouse model is not a theory or a set of equations, *per se*, but is really more like an experimental setup. Compare the transgenic mouse model to the scale model of an airplane wing in a windtunnel. Both are concrete physical models, which represent other concrete physical systems that are significantly similar in desirable ways, but different in others. The small scale wing exhibits many of the same aerodynamic characteristics, yet is cheaper and easier to build than the full size wing, thus making the experiments easier. The transgenic mice can be bred in scale and used to test new drugs that would be difficult to test on humans due to the risks, and would yield less consistent experimental data. In the social sciences we find economic models, social models, and political models as well. Some economic models can be highly formalized mathematical models–derived from theory or from historical empirical data or some combination of the two. Political models might be sets of principles, or refer to an historical exemplar which embodies an abstract set of principles.

Further, consider the complexity of an atmospheric model of global warming. It uses sets of equations, assumptions, and data to generate predictions of the future states of the atmosphere

[3]Of course the model of the atom draws an analogy to the planetary model of the solar system–there are multiple layers of models at work.

extrapolating from historical data using a computational simulation. The equations for such a model are derived from various sources such as accepted physical theories of thermal and fluid dynamics, socio-economic models of greenhouse gas emissions in the future, and projections based on historical data. They might also use as input data the output of other models. Models can be stand-ins for theories or for data, they can be interpretations of particular phenomena or situations, they can be physical devices or even living creatures, and "modeling" can also be an activity. How might we sort out this collection of models?

While there are numerous issues to consider and clarify in distinguishing the various kinds and uses of models, a few distinctions should be drawn early. One such distinction is between abstract models and concrete models. The difference here is the intuitive one of whether the model is something you can point to, touch, manipulate, break, etc. These are concrete models. Abstract models are conceptual, they might be mathematical, they might be theories, they might be interpretations of theories. There are two kinds of cases which seem potentially problematic. The first is when we have an experimental model that is a type rather than an instance, and in that sense is abstract though it consists of particular concrete instances. This is rather like an abstract symphony which can have concrete performances and recordings. The second problematic kind of case are sets of equations. These can be written down and manipulated in various ways that lend themselves to being considered concrete, especially when they can be written different ways, and instantiated, computed or simulated, in a computer. It will eventually be important to firmly distinguish these cases, but doing so requires drawing a slightly different distinction first.

There is another way to get at a crucial difference in types of models–by instead distinguishing "theoretical" from "working" models. At first look, these would seem to be equivalent to "abstract" and "concrete," yet the difference is that this distinction is based on their *use* in science, rather than their objective ontological status. A theoretical model is one which tries to capture or express a theory. A theoretical model could also be derived from a theory as an interpretation of some natural phenomena. In each case the model is something static. The working model is something concrete and dynamic that can be interacted with directly. Written equations can be theoretical models but not working models. Model types are also not working

models. Computational models and simulations are more complicated but can be working models, though they are not necessarily working models. Computational simulations will be considered and defined more carefully in Chapter 4.

In making the theoretical versus working distinction, I am pointing to the different ways that models are used. There are not just two ways to use models, however, there are many. And "model" can be an activity as well as an object. In the actual practice of science, we can see scientists modeling the design of an experimental system after some theoretical models, as well as some working models, drawing on exemplary models, and modeling test data to use in evaluating their system. Sometimes working models are the subject of highly controlled experimental evaluation, sometimes they are fiddled and tinkered on to discover what they are capable of doing. Sometimes a working model just has to work at all in order to prove its point. Sometimes a working model corresponds to a theoretical model, and is a concrete version of it. Sometimes a working model corresponds to another concrete system or phenomenon which is difficult to work on directly. Sometimes a working model is used as a starting point for developing a non-existent theoretical model. Sometimes a theoretical model is derived from the combination of other theoretical models, or theoretical models together with working models. More often than not several of these practices are going on concurrently or successively in the practices of scientific work and technological development.

I find working models to be especially interesting for several reasons. First, I believe that working models have not been given sufficient philosophical attention. A sizeable philosophical literature has developed around theoretical models, and there is a growing literature on simulations, but very little has been said about working models. More interesting are the wide array of roles that working models can play in science. While they can be used to gather evidence in support of theories, they can also be used to develop intuitions about how things work and to form new hypotheses. They can lead to technological understanding and breakthroughs, and they can communicate concepts and ideas in powerful ways. Working models are engaged with the world as much as the scientist might be engaged with them. They *do* things and they react to things–in short, they are *performative*.

The other principle reason I find working models interesting is that I think that the models

of mind pursued by cognitive science should be performative, working models, rather than static theoretical models. It is the performative aspect of working models that makes them interesting candidates for being models of how the brain works. Many of the difficulties facing computational models of the mind are due to the disembodied, semantically-neutral, abstract nature of symbolic computation–working models are none of these. Working models are physically embodied, directly causally engaged with the world, concrete mechanisms. They obtain semantic content through their performative engagement with the world. Just what this means for building cognitive models will be explored in depth in Chapter 5, but this will be the payoff of our re-examination of models, cybernetics, and computation in the intervening chapters.

There have been various attempts to justify the use of working models in Artificial Intelligence, for example, though the epistemology seems to have shifted over the years. It seems that working models were central to the epistemology of mechanistic psychology in the first half of the 20th century. I hope to show in Chapter 3 how working models had a very sophisticated scientific role through what Roberto Cordeschi (2002) calls the "synthetic method" that reached its height in the 1940s. While fields such as AI adopted the general strategy of the synthetic method, the years of unfulfilled technological promises took their toll, leaving the field and its methodology struggling in the 1980s and 1990s. More recently there has been a revival of the older epistemology in the fields of Neuromorphic Engineering and Biomorphic Robotics (Webb & Consi 2001). It is worth saying something more about the synthetic method before laying out the organization of this dissertation.

1.3 The Synthetic Method

I believe that the synthetic brain models of the 1940s exhibited various aspects of modeling, and ultimately were working models of a great and continuing importance to the methodology of cognitive science. Indeed they established a new epistemic framework for cognitive psychology which continues to this day. Roberto Cordeschi has called this concomitant rise of material technologies and a new epistemology the "Discovery of the Artificial." The name is meant to capture the idea of creating artificial systems for the study of

human psychology, tough it eventually expanded to many artificial realms such as Artificial Life and complex systems theory.

While the classic histories of cognitive science (Gardner 1987, Crevier 1993, McCorduck 1979), do a very nice job elaborating the themes of computationalism at the heart of the cognitive revolution, they do not fully explore the shifting epistemology that came with them and treat the cybernetics movement as a simple precursor to that revolution. And even the histories of cybernetics (Heims 1991, Dupuy 2000, Galison 1994) do little to clarify the complex relations between cybernetics and computation. Aspray's quote above regarding the complex interplay between cybernetics and computation in von Neumann's work is enough in itself to indicate that the story is more complicated than one of mere succession and improvement. Cordeschi (2002), however, has elaborated on the interrelated development of mechanistic psychology and the construction of synthetic brain models during this era. He focuses primarily on a set of phototropic robotic devices built by mechanistic psychologists in the 1910s, 1920s, and 1930s, which culminate in the cybernetic devices of W. Ross Ashby and W. Grey Walter in the 1940s. Cordeschi's history focuses on the how the scientific methodology was shifted towards a "synthetic method," which he defines as the construction of working models of complex behaviors and abstract mechanisms. This methodology also allowed the brain sciences to establish themselves as solidly grounded empirical sciences at a time when they were being challenged on their foundations.

While Cordeschi (2002) has elaborated the synthetic method as a tool for historical explanation, I wish to push his ideas in two ways. First, I believe that we can enhance our understanding of the role of models in science more generally, and the mind sciences more specifically, by examining the connections between the synthetic method and working models. Second, I want to examine how the specific needs and interests of the brain sciences in this historical period shaped the core ideas behind the synthetic method. It is towards this end that I examine the ideas of the cybernetician W. Ross Ashby regarding the various epistemic roles played by working models in science.

It was within the heart of the discovery of the artificial that the modern stored-program computer was formed. This is interesting first for its implications for the application of the

synthetic method to nearly every scientific and engineering discipline through the use of computers and simulations. This will draw us into a consideration of the nature of computational models and simulations, and their relation to theoretical and working models. There is a sense in which the universal computer is a universal model, and as such it provided the stimulus for taking the computer's architecture itself as a model for the brain and mind. It is this conception of the brain as a universal computer that will bring us back to a consideration of Computationalism, and what it means to "model the world" for brains in light of what can be gleaned from thinking about working models in science.

1.4 Summary of the Following Chapters

The dissertation is organized as follows. After this introduction, I examine the major views of models in the philosophy of science in Chapter 2. This examination is meant to contrast views of models that focus on scientific theories from those that focus on scientific practices and working models. This is followed in Chapter 3 by a consideration of the nature and role of working models in the brain sciences of the 1940s. Chapter 4 is an examination of the complex interplay of modeling involved in the construction and use of the first computers. In Chapter 5, I examine the idea of the feedback-controlled regulator as a working model of the mind, and consider its relevance to recent debates over the Dynamic approach to cognition and to outstanding problems in mental representation and content. Before concluding this introductory chapter, I give a short overview of the key ideas from each chapter.

Chapter 2 presents an overview of the leading philosophical views of models in the philosophy of science in an effort to stake out a place for working models among these views. Among the views considered are the classic deductive-nomological model as described by Wesley Salmon, the semantic models of Bas van Fraassen, the phenomenal models of Nancy Cartwright, and the models-as-mediators approach of Mary Morgan and Margaret Morrison. This overview highlights the shifting emphasis in views from explaining scientific theories to explaining scientific practice. It also develops a contrast between the view of theoretical models commonly held among philosophers and the notion of a "working model" more commonly held among working scientists and engineers. I argue that the working model has not been given the

philosophical attention it deserves, though the significance of working models has been noted in passing. I consider the nature of working models in relation to the treatment of images and scale models. This chapter also clarifies the notions of isomorphism and similarity relations between models and what they model.

Chapter 3 presents the early history of synthetic brains as an instance of the models-as-mediators view. This view emphasizes the *autonomous agency* of models, an aspect of models which fits well with the developing view of working models. The chapter also examines Roberto Cordeschi's notion of the "synthetic method." In this chapter I seek to combine these threads together into a richer conception of how the development of synthetic brains functioned in the development of the brain sciences by mediating between low-level theories of neurons and high-level theories of behavior, while simultaneously offering an autonomous agency that permitted experimentation. This chapter also examines Ashby's perspective on the roles of models and simulations in the brain sciences, which is highly compatible with the view of working models developed.

Chapter 4 considers the complexities of thinking about the computer as a model of the mind. It examines the computer as being a model of the brain in several very different senses of "model." On the one hand the basic architecture of the first modern stored-program computers was "modeled on" the brain by John von Neumann. Von Neumann also sought to build a mathematical model of the biological brain as a complex system. A similar but different approach to modeling the brain was taken by Alan Turing, who on the one hand believed that the mind simply was a universal computer, and who sought to show how brain-like networks could self-organize into Universal Turing Machines. And on the other hand, Turing saw the computer as the universal machine that could simulate any other machine, and thus any particular human skill and thereby could simulate human intelligence. This leads to a discussion of the nature of "simulation" and its relation to models and modeling. The chapter applies this analysis to a written correspondence between Ashby and Turing in which Turing urges Ashby to simulate his cybernetic Homeostat device on the ACE computer, rather than build a special machine.

Chapter 5 focuses on the feedback-controlled regulator as a model of mind. I believe it offers a better model of the mind than the computer because it is intrinsically performative. The

regulator is also the centerpiece of one of the more interesting challenges to computationalism that has drawn its critique from cybernetic ideas–Timothy van Gelder's Dynamicism. This challenge is interesting because it recognizes the complex relationships and processes of modeling, even if it is ultimately somewhat insensitive to important aspects of it. It is also extremely interesting insofar as it takes as its central metaphor the feedback-controlled regulator as a model of the mind–though van Gelder's formulation is problematic. The feedback regulator was the central metaphor of early cybernetics, and was eventually developed by Ashby into the paradigmatic notion of a model in a famous paper in 1970. I consider the regulator as a model for performative representation which might offer valuable insights into situated and embodied cognition by refocusing our attention on the importance of building working models and thinking about the brain as itself a working model rather than a repository for theoretical models.

I conclude the dissertation by drawing together the threads of working models, computer simulations, scientific explanation, technological development, and our understanding of the brain and mind. I argue that the synthetic method was a combination of practices, theories, agencies, and mediations which bootstrapped, cris-crossed, and intertwined the sciences of the brain with the technologies of the computer and irrevocably altered scientific practices in the process. As this method is extended to biology and genetics, it promises to do the same in the 21st century.

Chapter 2
On the Nature and Uses of Models I: The Role of Models in Knowledge Practices

Thus, in order to envisage a physiological system that cannot be entirely explained by reason or by experiment, we might construct a theoretical model which would permit us to investigate it by calculation of data with which we were empirically acquainted and even, when necessary, to evaluate certain constraints. When we have thus constructed a model which seems to copy reality, we may hope to extract from it by calculation certain implications that can be factually verified. If the verification proves satisfactory, we are entitled to claim to have got nearer to reality and perhaps sometimes to have explained it.

-Pierre de Latil

2.1 Introduction

What was the scientific role played by the synthetic brain models of the 1940s? To address this question, it will first be important to understand the role of models in science more generally. In the coming chapters we will consider in detail how various specific models functioned within specific scientific debates, supported various theories, and carried certain implications for establishing a scientific ontology of the mind. Without a clear view of what models are and how they are used, those detailed studies would likely prove difficult and frustrating. In seeking such a clear understanding, this chapter will consider recent work in the philosophy of science on models. From this survey I will outline a pragmatic theory of models that recognizes the role of models in scientific knowledge production. This involves taking a naturalist approach to understanding science, as opposed to a foundationalist or rationalist one. It also involves looking at the many kinds of practices that go into the production of scientific knowledge, beyond reasoning, theorizing, and explaining. The practical work of experimentation, innovation, and exploration are also crucial. Ultimately we must consider the fact that scientific knowledge is the product of a community, and the processes of articulation, communication, replication, and education are important to its success. In this chapter we will consider how models have been viewed by philosophers of science, and the progression of work in this area towards the kind of pragmatic view I shall endorse and in which working models figure more prominently.

As we will see, careful thinking about models in the philosophy of science has been

closely connected to a notion of *isomorphism*–a correspondence between some aspects of reality and some aspects of a model. This view of models grew out of logical empiricism and its attempts to rigorously formalize scientific explanation, though it was the anti-realist critics of this project who have done the most explicit work on models. The rigorously analytic approach to scientific reasoning within the philosophy of science was eventually challenged by historical, sociological, and anthropological studies for failing to account for scientific practice. The shift to looking at what scientists actually do has brought the philosophy of science into the field of science studies. I wish to affirm and continue this trend toward a fuller integration of the philosophy of science and science studies. This integration depends on reconsidering traditional issues in the philosophy of science in terms of *practice* rather than just *reasoning* or *theories*, and a careful consideration of science as it is actually performed now and was performed historically. It is in this spirit that I will examine the synthetic brain models for new insights into the roles and functions of models in the next chapter. For this chapter, the goal will be to review recent work on models, and outline the conception of models that has emerged from this work.

Before we begin it is important to note that I do not wish to propose a rigorous theory or definition of models. In fact, it will be the conclusion of this survey that we should not attempt a rigorous definition of what a model is apart from how it is used by scientists. There is nothing about models in themselves which pick them out as models–it is only scientists, through their construction and use of models that make them count as models. This is part of what is implied by looking at scientific practices, rather than theorizing. This is also what makes models essentially representational in nature–in ways that theories need not be. But this view of models can also be sensitive to the robust nature of representation–models are not just symbolic representations, not just images, not just perceptions or observations, not just logico-linguistic structures. Models can be representational in all of these ways and more.

My ultimate aim is to understand how the synthetic brains functioned as models in early cybernetics and AI. I believe that these models operated on many different levels. Some of these levels are the traditional ones of theory, hypothesis, and experiment. These are the epistemic uses of models that have been considered by the philosophy of science. There is another set of levels at which these models operate in the social domain of science. This includes the "models

as mediators" view that has emerged from the science studies discourse. But there are other roles being played by models that have not been previously considered. On the epistemic side, models serve to transmit knowledge within the scientific community, as wel as for purposes of education, pedagogy, and the like. There are also roles for models as mediators that go beyond what the science studies discourse has so far considered, such as mediation between theories and disciplines. But the most significant shortcoming of both the philosophy of science and science studies discourses is to overlook the significantly different ways that working physical models and theoretical models operate. Nearly all of the philosophical literature on models has focused on the latter. While digital computers and computational simulations have done much to blur the distinction between theoretical models and working models, there remain differences worth noting. These will be especially important for our consideration of the differences between the synthetic brains of cybernetics and the intelligent programs of AI. I believe that the debates over the computational theory of mind and the alternatives proposed by its various critics ultimately come down to the nature of scientific models and representation, and what types of models are best.

What, then, can we learn from the philosophy of science about models to help us straighten these issues out? The first objective of this chapter will be to dissolve the traditional conception of science based on the deductive-nomological model developed by logical empiricism. This dissolution is not new, but has been performed and refined by the anti-realist critics of the deductive-nomological model for the last several decades. It is worth reviewing, however, so as to avoid any tendencies to return to more formal definitions of models. Even the most recent work on models in the philosophy of science concerns itself primarily with theoretical models, rather than with working models. It will thus be necessary to consider what aspects of working models are not covered by current views of theoretical models.

2.2 Logical Models in the Philosophy of Science

For most of the first half of the 20th Century, models received very little special attention in the philosophy of science. Discussion of scientific methodology was largely focused on issues of hypothesis formation, empirical tests, falsification, and confirmation. For the most part, this

work was also dominated by the linguistic approaches of analytic Anglo-American philosophy. The main approach was pushed by logical positivism, and later logical empiricism, and is evidenced by the great importance placed upon the formulation of hypotheses and theories in terms of sentences and propositions, and their deductive relations. While models occasionally appeared in these studies, they were not explicitly theorized initially.

Clearly defining the things philosophers of science have referred to as "models" is somewhat daunting. This is because the word is not always used to refer to the same things that scientists refer to when they use the word. The influence of modern logic on analytic philosophy has been enormous, and this is reflected in the tendency to use the word "model" as it is used in logic. A logical model is any assignment of variables in a system to particulars. There is, in fact, a whole branch of modern logic dedicated to "model theory" but which has little to do with the kinds of models used in science that we are interested in. There is, however, one use of logical models which is of particular interest. This was the leading theory of scientific explanation for much of the 20th century called the "deductive-nomological model" (D-N model). This is the formal theory of the logical structure of scientific explanation that stands at the center of logical empiricism. It is a theoretical model meant to capture the logical structure of scientific reasoning itself.

The D-N model holds that a phenomenon is "explained" when propositions corresponding to it can be deduced from fundamental axioms in a logical system, if those axioms correspond to laws of nature (Salmon 1984, 1989). The problem, as we will see shortly, lies in establishing the correspondences. These correspondences are an assignment of features in the natural phenomena to the variables in the logical system. The D-N model thus captures an element of modeling that is important, namely the relating of material reality to theoretical terms, but the logical apparatus which it uses to do this is impractical and fails to capture how scientists actually draw these relationships.

There has been a tradition in the philosophy of science which has developed various criticisms of the D-N model. Beginning with Bas van Fraassen (1980), those seeking to challenge logical empiricist interpretations of science have turned to an examination of the role of models. While van Fraassen's approach placed a heavy reliance on what he called "partial

isomorphisms," similar approaches were soon taken by Ian Hacking (1983) and Nancy Cartwright (1983) that placed less significance in isomorphism as an epistemic property, instead placing it in scientific practices more generally. For Hacking these practices are representing and intervening, while for Cartwright they involve capturing the phenomenal qualities of natural events, and the causal structure of those events when they can be found. Ronald Giere (1988) adopts a very similar account of models, but in the context of a somewhat different project–that of offering a cognitive understanding of science. Finally, the notion of an ideal or complete isomorphism, or what Paul Teller (2001) has called the "Perfect Model Model," has been identified as underlying much of the philosophical theorizing of models. But the perfect model model is untenable and frustrates our attempts to get a clear picture of the role of models in science. My review of these critiques will be partial, selective, and far from exhaustive. It is meant not as a comprehensive review, but merely to delineate a single thread leading towards, and helping to motivate, the pragmatic approach to models currently being developed in science studies.

2.3 Semantic Models

According to traditional philosophy of science, the knowledge produced by science is captured in *theories*. Depending on the particular philosophy, establishing just what a theory is, or how it is supposed to relate to the world, or the laws of nature, is the objective. The general tendency is to present theories as meta-linguistic structures called *propositions*–the semantic contents of sentences, independent of certain properties of sentences, such as a sentence having to be in a particular language. Just what the propositional contents of theories are supposed to be ranges from abstractions meant to cover wide and varied natural phenomena, to more direct summaries of empirical data and predictions. Cartwright (1983) goes so far as to distinguish these as "fundamental laws" and "phenomenological laws" respectively. The differences are relevant insofar as philosophers wish to give some account of how the propositional content of theories comes to have the epistemic property of *truth*.

It is traditional in logic and linguistics to separate the formal structure of propositions from the content. "Syntax" is the formal–logical or grammatical–structure, while "semantics" is

the content–this includes things like meaning, reference, truth, denotation, connotation, etc. It is in this framework that accounts of scientific theories were first presented:

> Impressed by the achievements of logic and foundational studies in mathematics at the beginning of this century, philosophers began to think of scientific theories in a language-oriented way. To present a theory, you specified an exact language, some set of axioms, and a partial dictionary that related the theoretical jargon to the observed phenomena which are reported. Everyone knew that this was not a very faithful picture of how scientists do present theories, but held that it was a 'logical snapshot,' idealized in just the way that point-masses and frictionless planes idealize mechanical phenomena. (van Fraassen 1980, p. 64)

And so, it is in the context of the logico-linguistic interpretation of scientific theory that models are introduced:

> The use of the word 'model' in this discussion derives from logic and meta-mathematics. Scientists too speak of models, and even of models of a theory, and their usage is somewhat different. 'The Bohr model of the atom,' for example does not refer to a simple structure. It refers rather to a type of structure, or class of structures, all sharing certain general characteristics. For in that usage, the Bohr model was intended to fit hydrogen atoms, helium atoms, and so forth. Thus in the scientists' use, 'model' denotes what I would call a model-type. (van Fraassen 1980, p. 44).

Two things are clear from this passage. First, that van Fraassen believes that the philosophical tradition did not always feel compelled to pay attention to what scientists actually do, at least insofar as trying to account for the "models" that scientists actually use. Instead, philosophers tended to be concerned with their own logical models of science. Second, we can see in van Fraassen a realization that it might be productive to change this, and to begin considering the models scientists construct and use in a more careful and naturalistic way.

It is important to note that the move towards a naturalist understanding of scientific practice is not necessarily an abandonment of normativity for description. While it is a descriptive project, it must also account for why certain practices are successful, and others not. Some of these practices are explicitly epistemic, and seek to provide justification and ensure the accuracy of the knowledge produced with them. Models can thus play a role in a normative account, though the aim of the current project will be largely a descriptive one which seeks to explain how models function in their various roles.

Before we look at how scientists use models, let us consider how the philosophical

idealization is supposed to work. Just what is the role of models in the logico-linguistic interpretation? The tradition coming down through logical empiricism views theories as primarily syntactic:

> The syntactic picture of a theory identifies it with a body of theorems, stated in one particular language chosen for the expression of that theory. This should be contrasted to the alternative of presenting a theory in the first instance by identifying a class of structures as its models. In this second, semantic, approach the language used to express the theory is neither basic nor unique; the same class of structures could well be described in radically different ways, each with its own limitations. The models occupy centre stage. (van Fraassen 1980, p. 44).

Thus, the introduction of models is, for van Fraassen, a shift toward a semantic theory of scientific theories. But what does this mean? If theories are axiomatic systems, then their axioms are stated. They are ideal formal entities, like the points and lines of Euclidian geometry. The empirical problem is then how to connect those theoretical axioms to the real world so as to evaluate their truth. According to the D-N model, we can deduce observational statements from theories using logical rules and a special set of axioms called "bridge laws" or "bridge equations" which are supposed to express correspondences or "bridges" between theoretical terms and real entities.

For van Fraassen, scientific models are supposed to replace bridge laws. They do this by offering structures which instantiate theories:

> To present a theory is to specify a family of structures, its *models*; and secondly, to specify certain parts of those models (the *empirical substructures*) as candidates for the direct representation of observable phenomena. The structures which can be described in experimental and measurement reports we can call *appearances*: the theory is empirically adequate if it has some model such that all appearances are isomorphic to empirical substructures of that model. (van Fraassen 1980, p. 64).

We will consider in some detail what is meant by "isomorphism" as referred to in this passage. First, however, let us note that certain logical structures of models (empirical substructures) are acting as a bridge between abstract theories and concrete experiments and measurements (appearances and observations). There are also presumably hidden substructures which explain the relationships between the observable features, and may or may not correspond to theoretical entities. The role of the models is thus to preserve the appearances, while organizing them in a

logical structure. It is the models which do the work of spanning between the theory and the empirical substructures, and it is the representational function of the correspondence between the empirical substructures and the phenomenal appearances which is referred to as "isomorphism." The epistemic quality of the theory is then one of "empirical adequacy," conceived in terms of the ability to find or actually instantiate the models which match the theory to reality.

2.4 Phenomenal Models

Van Fraassen's semantic theory employs models to get beyond the limitations of the syntactic D-N model. However, the semantic theory's insistence on "truth" as a semantic property was also to be challenged. In particular, it is not clear that the laws science uses are necessarily concerned with truth in the sense of "empirical adequacy," and that there are other ways to view the aims of scientific knowledge production. The approach taken by Cartwright (1983) distinguishes the search for truth from the search for explanations of phenomena as being different, and sometimes conflicting, scientific goals.

Cartwright's approach is leveraged on a distinction between "phenomenological laws" and "fundamental laws." The phenomenological laws are meant to capture observed phenomena, while the fundamental laws are meant to capture the underlying or hidden laws which govern observed phenomena. The fundamental laws are supposed to be more powerful because they apply to a broader range of phenomena, but this broadness of range comes at a price, namely that there is a great deal of hidden work which goes into making the fundamental laws fit a broad range of phenomena. In other words, fundamental laws lie:

> The appearance of truth comes from a bad model of explanation, a model that ties laws directly to reality. As an alternative to the conventional picture I propose a *simulacrum* account of explanation. The route from theory to reality is from theory to model, and then from model to phenomenological law. The phenomenological laws are indeed true of the objects of reality–or might be; but the fundamental laws are true only of objects in the model. (Cartwright 1983, pp. 3-4).

Phenomenological laws do not lie, however, and this is because they have smaller ambitions. What interests our current study, however, is the role played by models, which mediate between theories and phenomenological laws.

In van Fraassen"s notion of models as "sets of structures" there is already the idea that

there are different models for matching the same theory to different phenomena. However, we can also have different models of the same phenomena:

> In physics it is usual to give alternative theoretical treatments of the same phenomenon. We construct different models for different purposes, with different equations to describe them. Which is the right model, which the 'true' set of equations? The question is a mistake. One model brings out some aspects of the phenomenon; a different model brings out others. Some equations give a rougher estimate for a quantity of interest, but are easier to solve. No single model serves all purposes best. (Cartwright 1983, p. 11).

It is important to note in this passage that Cartwright observes that not only do scientists often use different models of the same phenomena, but more importantly that the choice in model is not necessarily aiming at truth. What often matters are more practical issues regarding how accurate the inputs to a given model need to be, or how accurate its outputs are. It also matters which equations are easier to solve, and so what calculations are easier to perform. This is a point we will return to when we consider working models and simulations and the question of performing calculations that simulate minds and brains.

Cartwright's own "simulacrum account" of science focuses on explanation as the primary goal of scientific work, rather than truth. Indeed, she notes that truth and explanation are different goals and can often be at odds with one another. One of the most significant ways in which explanation differs from truth is that it can be multiple, where truth is presumably singular:

> On the simulacrum account, to explain a phenomenon is to construct a model which fits the phenomenon into a theory. . . . The success of an explanatory model depends on how well the derived laws approximate the phenomenological laws and the specific causal principles which are true of the objects modelled. There are always more phenomenological laws to be had, and they can be approximated in better and different ways. There is no single explanation which is the right one, even in the limit, or relative to the information at hand. Theoretical explanation is, by its very nature, redundant. This is one of the endemic features of explanation in physics which the deductive-nomological (D-N) account misses, albeit with the plea that this annoying feature will no longer be present when the end of physics is reached (Cartwright 1983, p. 17).

In other words, the D-N model is simply too idealized to deal with what real science does. Scientists often have to deal with the real world, a world that is very messy compared to the logical ideals sought by philosophers:

> This is generally the situation when we have to bring theory to bear on a real physical

> system like an amplifier. For different purposes, different models with different incompatible laws are best, and there is no single model which just suits the circumstances. The facts of the situation do not pick out one right model to use. (Cartwright 1983, p. 104).

The factors that go into choosing the right model are not the pure epistemic criteria of truth, justification, and confirmation. Again, we can sense an urge to account for what scientists actually do, a shift from "science as reason" to "science as practice."

As we make the shift to studying scientific practice, we need to consider the kinds of factors that weigh on model choice and construction:

> It is important that the models we construct allow us to draw the right conclusions about the behaviour of the phenomena and their causes. But it is not essential that the models accurately describe everything that actually happens; and in general it will not be possible for them to do so, and for much the same reasons. The requirements of the theory constrain what can be literally represented. This does not mean the right lessons cannot be drawn. Adjustments are made where literal correctness does not matter very much in order to get the correct effects where we want them; and very often, as in the staging example, one distortion is put right by another. That is why it often seems misleading to say that a particular aspect of a model is false to reality; given the other constraints that is just the way to restore the representation. (Cartwright 1983, p. 140).

Thus, even as models are sought to represent reality, they may purposely be made less accurate and more approximate in some respects, or even literally false, in order to achieve the accuracy desired in another respect. This is one sense in which context and values can play a significant role in the construction and use of models.

There are also factors outside of any single context which influence the use of models across contexts, but still have nothing to do with "truth" *per se*. These factors have to do with the social coordination of scientific practice:

> The phenomena to be described are endlessly complex. In order to pursue any collective research, a group must be able to delimit the kinds of models that are even contenders. If there were endlessly many possible ways for a particular research community to hook up phenomena with intellectual constructions, model building would be entirely chaotic, and there would be no consensus of shared problems on which to work. The limitation to bridge principles provides a consensus within which to formulate theoretical explanations and allows for relatively few free parameters in the construction of models. (Cartwright 1983, p. 143-4).

And so here again we can see an anticipation of the shift to more social studies of science. Note,

however, that these are not simple social categories, social forces, cultural biases, or economic and political interests. Rather, there are specific aspects of the scientific community as a social enterprise that impose certain constraints on scientific practices–a social epistemology if you will. While many practices and models are possible in principle, or even actually, having shared constraints, models, and practices leads to greater coordination of the scientific community. These are not absolute values, however, as progress also requires a steady stream of novelty. But how far could geometry have gotten if no one ever committed to Euclid's axioms? It seems appropriate at this point to remember van Fraassen's comments on the nature of scientific commitments:

> To accept a theory is to make a commitment, a commitment to the further confrontation of new phenomena within the framework of that theory, a commitment to a research programme, and a wager that all relevant phenomena can be accounted for without giving up that theory. . . .Commitments are not true or false; they are vindicated or not vindicated in the course of human history. (van Fraassen 1980, p. 88).

In other words, it can be useful for progress and coordination to adopt various constraints on model-building, but the measure of that usefulness will be a matter of history. I wish to turn now towards developing a view of models as practical tools, essential in various ways to scientific knowledge production. The next major step in this direction has been taken by Ronald Giere, and followed by Paul Teller, and takes models in a cognitive direction.

2.5 The Cognitive Approach

In the preceding accounts of scientific models, there was an acknowledgment of the differences between what many philosophers had called models and what scientists called models. However, the concern was primarily with refining the philosophical conception of these based on observations of actual scientific practices. The next step in our progression towards a pragmatic view of models is to shift our goal. Rather than try to account for an ideal of scientific success conceived as universal truths, we can instead seek an account of scientific as the practice of knowledge production, application, and dissemination. This is a naturalist project that seeks to explain science as phenomenon, rather than a rationalist project trying to justify science conceptually.

In the consideration of models, a major step in this naturalist direction was taken by Giere (1988, 1999). Giere rejects the idea of starting with a conception of what scientific theories or models ought to be:

> Philosophers pursuing what they call foundational studies *begin* with a conception of what a theory ought to look like and seek to *reconstruct* theories in that mold. The question of whether their conception matches that actually employed in science is begged from the start. (Giere 1988, p. 63).

Many of the innovations in thinking about science have come from those who looked to scientific practice rather than idealized notions of scientific reasoning alone, including the semantic view of models. Giere's project is to extend the semantic view of models (which he calls the "model-based approach") into a cognitive account of science:

> My goal . . . is to explore a new reason for preferring a model-based approach. It is, in a nutshell, that adopting a model-based framework makes it possible to employ resources in cognitive psychology to understand the structure of scientific theories in ways that may illuminate the role of theories in the ongoing pursuit of scientific knowledge. This is part of a more general project of developing a comprehensive, naturalistic account of science as a human activity. (Giere 1999, p. 99)

Merely adopting a naturalist approach does not automatically imply any particular sort of account, it could be psychological, or sociological, or anthropological, or evolutionary or even economic or political.[4] While I believe this is an improvement on the earlier semantic view of models, I believe that there are serious limitations to a purely cognitive view of models. Primarily, cognitive approaches tend to focus on a single mind and its understanding of the world. As a result, such approaches have difficulty accounting for interactions between multiple minds, *i.e.,* social interactions, as well as interactions with a complicated dynamic material world. This is precisely what a pragmatic view of models does, as it moves toward seeing them as mediators between the cognitive and the material, as well as between multiple social agents in the scientific community and the larger community, as we will see in the next section.

That said, the cognitive approach has several advantages over previous approaches. An important aspect of this shift is to move away from explaining the foundations of science, and the need to establish an ultimate justification of its truths:

---

[4] The move to a cognitive approach will, however, bring us into an approach strongly influenced by the history of synthetic brains we are considering.

> I [argue] that a concern with the cognitive structure of scientific theories is largely orthogonal to the sorts of foundational concerns with which the phrase "the structure of scientific theories" has typically been associated. (Giere 1999, p. 99).

But the principle advantage is that it opens up the study of knowledge production to the vast conceptual resources of cognitive psychology:

> [In the cognitive sciences] one finds models of cognitive agents who develop representations of the world and make judgements about both the world and their representations of it. It took no great leap of imagination to think of scientists as cognitive agents and of scientific models as a special type of representation. Likewise, scientists' decision making about models can easily be seen as an exercise of ordinary human judgement. (Giere 1988, pp. xvi-xvii).

Moreover, even though I shall argue that there is more to models than their cognitive aspects, a very important aspect of them will always be constituted by their cognitive, conceptual roles.

Giere's naturalistic shift also looks to forms of understanding beyond formal linguistic or logical structures. This begins by redefining the mediation role of models between theory and reality as being one of *similarity* rather than one of *isomorphism*:

> The appropriate relationship, I suggest, is *similarity*. Hypotheses, then, claim a *similarity* between models and real systems. But since anything is similar to anything else in some respects and to some degree, claims of similarity are vacuous without at least an implicit specification of relevant *respects* and *degrees*. The general form of a theoretical hypothesis is thus: Such-and-such identifiable real system is similar to a designated model in indicated respects and degrees. (Giere 1988, p. 81).

The difference between similarity and isomorphism is subtle but profound. While isomorphism is meant to be a rigidly definable relationship, similarity is a looser notion largely dependent on the context of use, intentions of users, and the ability to articulate a similarity in any particular use, rather than define it for all possible uses. Paul Teller (2001), who largely follows Giere's conception of models, explains the naturalistic usefulness of the similarity notion:

> I take the stand that, in principle, anything can be a model, and that what makes a thing a model is the fact that it is regarded or used as a representation of something by the model users. Thus in saying what a model is the weight is shifted to the problem of understanding the nature of representation. I do not begin to have a workable account of representation, so what is accomplished by this move? The point is that when people demand a general account of models, an account which will tell us when something is a model, their demand can be heard as a demand for those intrinsic features of an object which make it a model. But there are no such features. WE make something into a model by determining to use it to represent. Once this is fully appreciated it becomes clear that

> we can get on with the project on the strength of a good supply of clear cases of things which are used to represent. These will adequately support study[ing] the variety of such uses, the way they function in the scientific enterprise, their interrelations, and so on. (Teller 2001, p. 3)

We will consider the nature of representation in models at various places in the pages to come. The point to be made here is simply that a notion of similarity is sufficient to cover everything we want to count as models, without requiring a universal definition of what counts as a model, or the rigid limitations imposed by isomorphism. The result is a view of models that is contextual, local and particular, rather than abstract, general and universal–models are tools and as such are only understandable in relation to their use.

One downside of the cognitive approach is that it privileges the conceptual role of models, and stays too much "in the head." Giere ends up presenting a very conceptual and theoretical account of scientific theories and models:

> I suggest calling the idealized systems discussed in mechanics texts "theoretical models," or, if the context is clear, simply "models." This suggestion fits well with the way scientists themselves use this (perhaps overused) term. . . . As the ordinary meaning of the word "model" suggests, theoretical models are intended to be models *of* something, and not merely exemplars to be used in the construction of other theoretical models. I suggest that they function as "representations" in one of the more general senses now current in cognitive psychology. Theoretical models are the means by which scientists represent the world–both to themselves and for others. They are used to represent the diverse systems found in the real world: springs and pendulums, projectiles and planets, violin strings and drum heads. (Giere 1988, p. 79-80).

While I do not wish to argue that there are no theoretical models, and understand the need to account for these, I do believe that these are not the only kinds of models a thorough-going philosophy of science should concern itself with. The philosophy of science must also account for diagrams and images, physical models, working models, and other forms of models. This can be done by looking at scientific practice, and is a significant aspect of a naturalistic understanding of science.

2.6 Images and Scale Models

Once we see models as representations, a reasonable question to ask is what kind of representations they are, or can be. If we take Teller's particularism about models seriously, then

we cannot say definitively what can or cannot count as a model outside of any particular case of trying to model something. However, there are some pretty commonly employed categories of models that are worth considering at this point. We have already considered to some extent models as linguistic representations. This includes models as sentences or logical propositions, as well as mathematical models, considering mathematics as a formal language. Two other categories of model frequently referred to are images and scale models. Visual images have been considered in great length, while scale models are usually only mentioned in passing. I wish to briefly summarize some of the major features of each insofar as they might suggest an approach to pragmatic models.[5]

In his later work, Giere seems to have recognized that the representational aspects of models extend beyond the linguistic and conceptual. In particular, he considers in great detail the role of images in science. Giere (1999, Chapter 7) gives a lengthy account of how visual images can function in scientific reasoning and debate. In doing this, he makes it clear that the relevant aspects of visual models are non-propositional representations. He also continues his cognitive project by limiting his analysis to how images function as specific instances of reasoning, argumentation and debate–always with a view toward justification and agreement. The basic strategy is to account for how visual aspects of these models are brought into similarity relations with natural phenomena and data on the one hand, and aspects of theory on the other. Again, I believe that this is a highly illuminating example of the use of images, but retains the traditional goal of leading directly to the conclusion of a logical argument. One aspect of models that turns out to be very important to their usefulness is their open-endedness–their ability to support multiple views, to be extended in multiple directions, and to be experimented upon with unexpected results. These all follow indirect routes to the traditional goals of justifying a theory or reaching agreement in the scientific community and usually never reach such conclusions, but are no less important to the productivity of science because of this. Trial-and-error still needs errors in order to be successful. It is this openness which allows for science to extend and

[5]One goal of this dissertation, of course, is a fuller appreciation of the representational character of physical models as a form of distributed cognition, but this will have to await a description of situated cognition and wide computationalism in Chapter 5.

progress.

The broad class of physical models, and the subset of these called scale-models, are usually mentioned only in passing by philosophers of science. This is because the use of physical and scale models is considered to be much simpler and more straightforward than visual images:

> What are models? The worry is not so pressing with physical, and especially, scale models. One understands, at least in outline, how science makes good use of engineers' aircraft models in wind tunnels, analog computing devices, and the like, all of which do good service in both prediction and explanation, especially in technological explanations. When we specify 'scale models,' the problem of similarity also dissolves: To specify the scale is just to specify the operative similarity. But in the theoretical sciences, for the most part, one does not have physical models in mind. (Teller 2001, p. 3).

As my concern in the coming chapters will be with physical models, that these should be philosophically unproblematic should be comforting. Yet I contend that in fact little has been said about such models by philosophers, and that this is a good indication that much is being taken for granted about them. For instance, are scale models supposed to be the ideal form of physical model? How are physical models supposed to be the prototype of mathematical and theoretical models? Are computational simulations essentially a form of scale model? How do physical models stand in similarity relations to other bits of physical reality which they are meant to model? And most importantly, what sorts of scientific practices do physical models support that theoretical models cannot? Ultimately, establishing something as a model by pointing out its similarity relation is just the first step in model use, and the more interesting processes are what follows on after that. It is not just the nature of the similarity relation between a particular physical model, the digital computer, and a bit of physical reality called the brain that has presented the most challenging questions for the philosophy of mind for the past 50 years? What has made the computer model of the mind so compelling is a vast array of conceptual alignments, technological products, and material performances. This being the case, we should perhaps not be too quick to accept that these sorts of models present simple cases of representation.

Take, for instance, the nature of numerical representation in computers. Von Neumann discusses this at length in his early lectures on computation and the stored program computer. It is here we find the distinction between analog and digital representation, and arguments for the technical superiority of the latter. It is analog representation which Teller is referring to in the

passage above, both in the form of analog computation and in the principle relation of scalar quantities in scale models. And it is discrete, digital representation which we shall see challenged as metaphysically unsound by some of the proponents of the Dynamicist critique of Computationalism as being metaphysically unfit to represent the mind. There is clearly a story to be told about how these forms of representation came into being, how they were used, and how they functioned in theoretical debates over the nature of mind and scientific models of mind.

I believe the lesson we should take from the fading twilight of the perfect model model is that while models cannot hope to be "perfect" or "exact" representations, they will always be some sort of representation. Moreover, the diversity of ways in which they can represent is better thought of as a crucial element of the usefulness of models in helping to extend knowledge into previously unknown realms. This view of models is now being expounded in the science studies discourse around the notion of models as mediators. Before the view of models as mediators, I wish to follow up on some of the themes of model use in the philosophy of science by looking at the nature of pragmatic epistemic virtues of models.

2.7 Pragmatic Models and Epistemic Values

> *The final test for models, however, is in the range of phenomena that they actually contribute to explaining.* (Cordeschi, 2002, p. xix).

I would like to end this chapter by arguing to extend the list of epistemic values based on a pragmatic understanding of the function of models and modeling-making practices in the sciences. The current philosophical views of models, in my opinion, still hold close to a correspondence theory of truth, in which models ultimately aim to confirm or disconfirm theories and hypotheses, and it is these theories and hypotheses and their Truth–or their proximity to the Truth–which are the ultimate source of epistemic values. In these views, Truth is a purely semantic quality, a relation between symbol and object. Even the model-as-mediator could be seen as bridge built to convey the empirical truth mined from observational data to the bank vaults of support for theories–the truth-value of a theory is supposed to be determined by the accumulated confirmation of its hypotheses and predictions, built up through controlled experiment and observation. There are, of course secondary epistemic values in this view. These secondary epistemic values include: the features of theory such as simplicity, elegance, scope and

range of applicability; features of models such as structural isomorphisms between theory and model, accuracy of predictions, and extensibility; features of experiments such as accuracy of measurements and the reliability/reproducibility of results; and features of the scientific process such as confirmation/disconfirmation, falsification, and the relevance or centrality of a given experiment to the theory in question. Of course, models mediate in all these ways in debates over the truth of scientific theories and in the interpretation of experimental results. Rather than rejecting this view, I wish to augment it by acknowledging the role of models in the pragmatics of scientific knowledge production. And thus, what were traditionally held as the ultimate objectives, and ultimate virtues, of scientific models, namely their universality, accuracy, and "truth," are all reducible to pragmatic manipulations of effective representations. The pay-off of such a view will be a more holistic view of mind.

There are several senses of "pragmatic" that are relevant to the use of models in science, all of which revolve around a notion of "usefulness." First is the linguistic sense, in which pragmatics stand apart from semantics and syntax as indicating the context in which a symbolic structure is interpreted. This gets at the idea that the usefulness of a model is conceptually independent of its semantic truth-relations, but also appears to divorce usefulness from epistemology–barring a round-about argument that concludes usefulness is a good indicator of truth or something of that sort. Another sense is that of the founder of pragmatism, C. S. Peirce, that knowledge is a triadic relationship between the symbol, object and the mind doing the symbolizing. This adds agency to the equation, knowledge is not just the relation between belief and object, but between the believer-as-an-agent-in-the-world and the object, through the use of a representation. However, it is the third sense of "doing something in the world" which completes the notion of agency. Knowledge is in this sense performative. The performative view of knowledge is also integral to Andy Pickering's mangle of practice (1995). We shall consider these in detail in the next chapter. For the remainder of this chapter, I would like to characterize just what is meant by a "pragmatic epistemic virtue," and show that the cybernetician W. Ross Ashby endorsed a pragmatic view of models.

Even within the semantic view of models, there has been explicit consideration of their pragmatic value. The issue has been taken up by van Fraassen (1980), and I would now like to

clarify and extend his ideas on the matter. We should begin by noting that van Fraassen follows Charles Morris's definition of pragmatics as a branch of linguistics, alongside syntax and semantics (van Fraassen 1980, p. 89). As a consequence of this, truth is counted among the semantic properties of a theory, and thus the pragmatics have the appearance of being theoretically distinct from epistemology. The introduction of pragmatics is thereby the introduction of context to the interpretation of theory statements.

I wish to follow van Fraassen's suggestion of looking at the pragmatic virtues operative in science. However, I wish to move beyond the restrictions of linguistic and semantic interpretations. Theories are not sentences or propositions or, according to van Fraassen, sets of models. Instead, I want to include physical working models among the models which constitute science. I also wish to broaden the notion of pragmatics from the linguistic sense of context to the sense more aligned with the pragmatism of C. S. Peirce and William James, and identify pragmatics with situated and embodied scientific practice. Thus the content of theory is not its syntactic content (as van Fraassen has already argued) nor is it its semantic content in the sense that includes experimental instantiations of theory. Rather, it is the pragmatic value of theory that is primary–scientific practice is at the center of scientific knowledge production, and it lays at the intersection of knowledge and technique, representing and intervening.

We can see van Fraassen's commitment to the semantic framework in the way he frames the significance of "usefulness" as a value apart from truth, and thus competing with epistemology:

> There are specifically human concerns, a function of our interests and pleasures, which make some theories more valuable or appealing to us than others. Values of this sort, however, provide reasons for using a theory, or contemplating it, whether or not we think it true, and cannot rationally guide our epistemic attitudes and decisions. For example, if it matters more to us to have one sort of question answered rather than another, that is no reason to think that a theory which answers more of the first sort of questions is more likely to be true (not even with the proviso 'everything else being equal'). It is merely a reason to prefer that theory in another respect. (van Fraassen 1980, p. 87).

This is to say that epistemology under the semantic view is concerned primarily with the truthfulness of knowledge. Here van Fraassen is following the linguistic definition of pragmatics, and placing human concerns, values and interests (apart from truth) in the class of

pragmatics. While this allows him to claim that model choice can be based on such human values, it simultaneously puts them outside the realm of the epistemic. Other passages confirm this interpretation:

> Nevertheless, in the analysis of the appraisal of scientific theories, it would be a mistake to overlook the ways in which that appraisal is colored by contextual factors. These factors are brought to the situation by the scientist from his own personal, social, and cultural situation. It is a mistake to think that the terms in which a scientific theory is appraised are purely hygienic, and have nothing to do with any other sort of appraisal, or with the person or circumstances involved. (van Fraassen 1980, 87-88).

> [T]he answer is that the other virtues claimed for a theory are *pragmatic* virtues. In so far as they go beyond consistency, empirical adequacy, and empirical strength, they do not concern the relation between the theory and the world, but rather the use and usefulness of the theory; they provide reasons to prefer the theory independently of questions of truth. (van Fraassen 1980, p. 88).

I believe that there are epistemic values in scientific knowledge production that operate independently of truth. Epistemology is fundamentally about knowledge, not truth, and it is possible to have an epistemology that instead seeks its foundations of knowledge in the success of interactions between an agent and a dynamic world.

There are several immediate consequences to be drawn for the virtues of scientific models from such a pragmatic epistemology. There may be models which are known to be literally false (those laws of physics which Cartwright demonstrates as lies, for instance) but which are used in the production of good science. They can do this because there can still be aspects of such models which are particularly compelling and influential in explanation and prediction. There can still be experimental demonstrations of it, it can be taught to students. In short, such a theory can be *productive* even if it is not literally true. In fact, theories can be productive even if no one believes them to be literally true. Notice that "being productive" is different than "being good to believe" which is the standard interpretation of the pragmatic theory of truth. We do not have to believe in the truth of a model in order to use it, we merely need to believe in the effectiveness of using it. Once we view science as social knowledge production, we add important temporal and social elements that are missing from traditional epistemologies of "justified true belief." Temporally, sometimes a theory known to be false can provide a needed element in a

developmental process that arrives at a theory that is true or closer to the truth.[6]

Socially, the production of knowledge does not stop with the understanding of a single scientist working alone in a lab. She must share her knowledge with colleagues, and if it is accepted it must be taught to future generations of scientists. Sometimes, these theories must be explained to the society at large in order to secure funding, or to secure acceptance of scientific practices or its applications. Though the practices which enact these processes are often far from the lab and mind of the individual scientist, the ultimate success of a theory in terms of its wide-spread acceptance and influence on society depends on these processes. In this broad view of knowledge, epistemology must take into account these processes.

The features of ideas which enable their communication and dissemination are practical issues, again independent of truth *per se* but deeply connected to the practical issues of scientific justification. Thus such values of models as being easily comprehended or visualized, or being easily transmitted from one lab to another, of offering consistent, reliable, and accurate results, and the like–these are all significant values in the establishment and adoption of theories, models, and technics. They are all values that are simultaneously practical and epistemic. Traditional epistemic notions like "reliability" take on new meanings. It is not simply a property of beliefs in an individual cognitive system. Now we must worry about the reliability of the communication of beliefs and theories, the proliferation of interpretations, the equivocation of key theoretical terms, and so forth. It thus seems natural to look to models in understanding these practical epistemic virtues. In the competition of ideas, those which are able to spread more easily, and more widely have much greater chances of success.

2.8 Modeling and Models as Mediators

Now that we have reviewed how the philosophy of science has treated models, we should return to consider how the broader field of science studies now sees models. There are two important views that, though closely related, view models and modeling somewhat differently.

---

[6] I will leave aside consideration of whether such a theory can, or must, abandon truth altogether. I am confident that it is possible to defend such a theory with a concept of truth that is real but always inaccessible, or a concept of truth that is always partial and approximate. All that is required of a theory of truth is that it provide a reliable means of determining whether a belief is closer to the truth than another for some purpose.

One view, that proposed by Morgan and Morrison (1999), stays close to the philosophical views above, and sees models as bridges or mediators between theories and data, or theories and natural phenomena. As we shall see, they do make a significant theoretical move beyond the semantic and cognitive approaches, but do not fully realize the implications of this move in much of their work. The other view is far greater in scope, and takes *modeling* as a process of cultural extension. This view, proposed by Pickering (1995), is thus much larger than scientific *models*, and is even larger than science itself, subsuming all forms of cultural extension such as art, architecture, economics, music, politics, sports, and technology, and indeed any human endeavor which develops and grows over time.

In the introduction to their edited collection on the subject, Morgan and Morrison (1999) articulated the notion of models as mediators by invoking a notion of autonomous agency. The idea that models are autonomous agents is crucial to a pragmatic understanding of scientific practice. Too often representation is taken to be a purely conceptual or purely symbolic phenomena. What this view of models suggests is either that models are *more than* merely representations, or it suggests that there is something to being a representation which involves agency. Let us consider each view in turn.

According to the first interpretation it is likely to be the physical and working models which have agency, as it is harder to see how symbolic and theoretical models can achieve this. This might explain in part why the philosophers see the physical models as less problematic. But there is a good account of how theories and concepts can express agency. Pickering (1995) presents a wonderful discussion of disciplinary agency in the practice of mathematics, through a case study of Hamilton's development of "quaternions," which does offer us a view of symbolic and theoretical structures as having an agency of their own in virtue of the disciplinary requirements imposed on their use by the researcher.

If we can understand the representational aspects of models independently of their agency, as the first interpretation suggests, what would be the value of considering them as integral? Apart from any evidence for or against such a view, the value of such a view seems quite clear. I believe that a better understanding of the representational agency of models will go a long way towards helping us to understand representation more generally. Indeed, the very

kinds of cognitive representation which will become crucial to the story of the synthetic brains will benefit greatly through an understanding of the agency of representations. I will wait until the next chapter to flesh out this idea, however.

The approach to studying scientific knowledge production which is most sensitive to the dynamic and open-ended processes of practice is Pickering's (1995) mangle of practice. While there are several aspects of this approach operative in my project, the aspect I wish to draw special attention to at this point is Pickering's division of scientific practice into three fundamental realms: the conceptual, material, and social. His analysis of scientific practice thus proceeds by examining the interactions between the elements of these realms in the unfolding of time, *e.g.,* how the material resistances of the laboratory shape theory, how theory suggests experiments, how social networks transmit material technologies and laboratory technics, how social politics shape theory, as well as practices which sometimes extend knowledge through "purely" conceptual, material or social means in isolation. The point is that no single realm determines the course of scientific progress, and any particular history is an open-ended unfolding of these dynamic interactions in time. The elements of these realms are best seen not as objects but as agencies, they are not vague abstract forces, but specific actions, events, phenomena, concepts, etc. which all do things in their own realm and influence one another to varying degrees. For instance, cellular telephones do something in the material world by sending radio wave signals through networks of towers to other cellular phones, but they also change our social practices rather dramatically, and in turn have changed our conceptions of space–their emergence involves complex interactions between all these realms.

Pickering's mangle is a broad view of science and culture, but one which I believe is compatible with the view of models as mediators. The notion of models as autonomous agents in the models as mediators view draws on the same concept of agency operative in the mangle of practice. As such, we can connect these views together, and in so doing investigate how the agency of models extends into the epistemic processes of knowledge production. What is attractive about such an approach is that it allows us to consider epistemic processes as practices performed in laboratories by scientists, rather than as logical processes in the heads of scientists. In this way they are observable, and we avoid privileging any specific epistemic theory before

seeing what it is that scientists do.

A crucial aspect of models that enables them to serve their mediator role is the process of their construction. Models, especially material models like the synthetic minds, travel through social space and develop over time. The practice of constructing models builds upon and extends both theory and technical knowledge–the disciplining of material agency–and is also intimately connected to economic, social, and political agencies, but is not determined by any of these alone. This process is not arbitrary or without constraint, but faces resistance from conceptual, material, and social agencies. By shifting our focus to the ways in which scientific *practices* derive their epistemic power we come to see models as the principal aim of the bulk of scientific practice. As such, a consideration of the activity of modeling offers a view into the production of scientific knowledge. In particular, we can ask what features of models contribute to their success in the establishment and communication of scientific knowledge; these will be epistemic values in the production of public knowledge. In other words, an explanation of the development of synthetic brain models ought to consider which models had favorable advantages in virtue of their exploiting or accommodating the conceptual, material, and social agencies they confronted.

By understanding scientific knowledge production as a process of open-ended cultural extension, as suggested by Pickering (1995), we come to see the role of models as mediating not only between theory and data, but also between previously disparate theories, between old and new material technologies, and also mediating between all of these elements and social, political, and economic forces. As in the models as mediators view, models are autonomous agents, and their agency can range from the disciplinary agency of abstract formal models, to the material agency of experimental and physical models. The ability of a model to translate itself from one scientist and one laboratory, to other scientists and laboratories depends upon all of these forms of mediation, and not simply the mediation between theory and data. For example, a model which bridges two otherwise unrelated theories may become widespread despite a lack of any new data being accounted for or novel empirical observations being produced.

The example of the McCulloch and Pitts (1943) neuron model is a case in point. Their model of the neuron was highly idealized, in spite of some detailed knowledge of the physiological processes occurring in real neurons which was available at the time. Yet they

sought to build a theoretical bridge between mathematical logic, and hence computation, and the behavior of the brain at a neuronal level. The move achieved two bridges really, which made it a very compelling metaphor for the computational brain. The first was a bridge between logic, as the language for expressing the contents of thought, and the brain. Here the goal was to show how the philosopher's logical mind might arise from the physiology of neurons. The other bridge was between the specific physical structure of the brain and the effective logical structure of computation. The goal of this bridge was to show how universal computations might be possible, and further how that possibility could be realized by networks of neurons. Both bridges would turn out to be important to the development of models of the brain in the 1950s. The computation-by-network bridge in particular was crucial to the conceptualization of the stored-program computer by both Alan Turing and John von Neumann, while the logical-structure-of-thought bridge formed the foundation of Artificial Intelligence research. I mention the McCulloch and Pitts example to demonstrate that models can mediate between different theories, as well as between theory and reality. In this example, like the electronic brain models I will discuss shortly, there is no real attempt made to "fit" any particular set of observational data, though they are often presented as illuminating or even explaining the gross behavior of intelligent organisms.

In these cases, the mediation is really being done between different levels of description, and there will be much more said on this in the next chapter. Another role is as a substitute or simulation of a natural system which was not directly or easily accessible to researchers (real brains, hypothetical brains). How does this differ from philosophical accounts of models we have considered? Typically, scientists start from sets of observations at a single level of analysis and attempt to construct models of natural systems to correspond to those observed phenomena. In the case of the brain sciences, models were constructed to bridge across levels of description, despite the fact that there were no observable phenomena which linked these levels of description. These brain models mediated between known pieces of data at different levels of analysis and interpolated what natural systems might lie between these. This then drove empirical research, which went looking for the postulated entities. The next chapter will elaborate the ways in which the synthetic brain models served as mediators, and how this shaped

the unfolding sciences of the mind. Before we consider the synthetic brains themselves, I want to first give an overview of Ashby's ideas on models and simulations, and show that they are highly compatible with the pragmatic view I have just outlined.

2.9 Ashby's Pragmatic Models

It seems to me that if we are willing to abandon the notion that the only epistemic value of scientific knowledge is exclusively derived from the truth of theory, then we are free to consider the pragmatics of knowledge, derived through the construction and use of these models. W. Ross Ashby seems to have been very much aware of this feature of models:

> It seems to me that this purely pragmatic reason for using a model is fundamental, even if it is less pretentious than some of the more "philosophical" reasons. Take for instance, the idea that the good model has a "deeper" truth–to what does this idea lead us? No electronic model of a cat's brain can possibly be as true as that provided by the brain of another cat; yet of what *use* is the latter as a model? Its very closeness means that it also presents all the technical features that make the first so difficult. From here on, then, I shall take as a basis the thesis that the first virtue of a model is to be useful. (Ashby 1972, p. 96).

In this passage, Ashby is not only endorsing a pragmatic view of models, but pointing to one of the more powerful advantages of this view–namely that models are useful. Ashby had much more to say about the specific ways in which models can be useful in scientific practice, and the kinds of virtues that enable this usefulness.

The cyberneticians built their models not to "account for specific data" as we might expect from traditional philosophy of science accounts. Indeed, at the time there was very little specific data about the inner workings of the brain available, at least in terms of gross organization, cognition and behavior. Rather, they attempted to triangulate where one might look for such data by constructing models that bridged disparate elements of theory, and instantiated these as material models which shared some of the phenomenal properties of interest. These models could then be themselves experimented upon in lieu of real brains, and thus provide a basis for the development of data and theory, which could then be translated or bridged back to understandings of the brain. Moreover, Ashby recognized that much of the value of these models lay not in their instrumental precision or formal structure, but in their principles of construction

and ability to convey difficult concepts in an accessible way. It is in this regard that he recognized the epistemic pragmatics of the communication of ideas through models as often being as important as, if not more important than, the literal truth of the underlying theories, the accuracy of predictions, degrees of confirmation, or the extent of isomorphisms of structure between theories and models.

The essence of a pragmatic theory of knowledge lies in the usefulness of knowledge for acting in, and getting along with the world. The same principle that applies to the individual knower also applies to knowledge communities. With communities there are several challenges facing the production, distribution, and retention of knowledge that are either simpler or remain hidden in the case of the individual knower. For our purposes it is best to think of science as an example of community knowledge *par excellance*. Even when knowledge is produced by individuals, it must be adopted by the scientific community, which imposes restrictive criteria on what counts as "good" science. But the sharing of data and theories and the community coming to accept these as knowledge are not the only aspects of knowledge which are shared. It has been noted by philosophers of science that background assumptions, motives, and interests must also be shared in order to achieve shared acceptance of theories (Feyerabend 1975). The communication of theories is also tied to the communication of laboratory techniques, mathematical techniques, working models, symbolic systems, and representations, experimental setups, instruments, and the like (Kuhn 1962). The communication and sharing of these cannot be simply taken for granted, but involves a great deal of work and shapes the knowledge that emerges. That is to say, these elements of knowledge production have *agency*–they do work and in virtue of this they influence the shape of emerging knowledge.

Ashby was certainly well aware of the pragmatics of knowledge in which the cybernetic brain models were embedded. Indeed, he notes that even the development of theoretical models in the science of mechanics was strongly shaped by the mathematical techniques available:

> We are in danger, perhaps, of being led astray by the outstanding merits of certain well-known particular models. We rightly admire Newton's system of equations and laws and, after its great success, are apt to think that he discovered *the* model. I suggest that his model is widely used largely because pencils and paper are widely available, and *his* type of mathematics widely known. Had our circumstances been very different we might well have preferred a different model: had we lived, for instance, in a world where algebra was

> not a practicable process, but where many point-sources of light and many conical bodies made the geometric development of ellipses instantly available, we might well have found that wholly geometric methods were preferable to the algebraic. (Ashby 1972, p. 97-8).

The "availability of pencils and paper" is clearly a pragmatic concern. Moreover, Ashby clearly ties the practicality of Newton's equations not to their discovery, but to their preference, which we might just as well read as acceptance.

For Ashby, it is the models themselves that constitute scientific knowledge. What are considered to be the great "laws" discovered by science are really just models with similarities that hold over a broad range of natural phenomena:

> Sometimes, as Newton found, a comparatively simple model can be isomorphic with an extremely broad range of phenomena; then we speak of his discovering a great "law," (not universal, however, for his "law" fails at the cosmic and nucleonic extremes). When this happens, the event is so striking and worthwhile that whole generations of later scientists take as their aim to finding another such isomorphism. I do not for a moment wish to suggest that no more such laws remain to be found, but it is true that most scientists cannot expect to be as lucky: much of their work, especially in the behavioral sciences, will have to be on the construction of isomorphisms that are not only of narrower range but are also much more complex in their structure. (Ashby 1972, p. 111).

Thus, scientists need not feel obligated to seek out only "universal laws" or to seek to establish the universality of the models they find useful. The value of a model can rest in explaining a limited range of phenomena, for a limited set of purposes. As we saw in the discussion of Cartwright's (1983) critique of scientific laws above, universality (or even just generality) usually comes at the price of accuracy or truthfulness. Newton's "universal" law of gravitation requires all sorts of caveats and margins of error, and outright ignorance of known factors in order to get them to work. But the deeper epistemological point here is that similarities must be drawn, and isomorphisms are *constructions*.[7]

Isomorphisms are not merely symbolic relations achieved by linguistic pointing, but must rather be constructed through the practices of experimentation, demonstration, and communication. This is important because the success of the isomorphism is not completely

[7] I will continue to use "isomorphism" in the ensuing discussion because this is Ashby's term. I believe he has the right idea in mind with it, but would prefer that these be understood as similarity relations, rather that strict logical isomorphisms.

contained within the isomorphism relation itself, but depends on the surrounding practices to achieve the isomorphism. In the absence of those practices, an isomorphic relationship is of little interest. Consider two arbitrary symbolic systems and sets of relations within them, and the assertion that there is some partial isomorphism between them. This isomorphism by itself tells us little about either system. It is the processes of construction that grounds the elements and relations in material reality, and serves to guarantee the reliability of the asserted isomorphism. These processes are hybrids of pragmatic technique, epistemic verification, and metaphysical action.

Another pragmatic aspect of models that is often overlooked is their complexity. While the status of "simplicity" as an epistemic virtue has long been debated, this debate has largely been framed as whether simplicity is in itself an argument for truth. Missing from this debate is a clear notion of just what is meant by "simplicity." Ashby explains formally what he means by simplicity and complexity by employing the theoretical apparatus of information theory. As a pragmatic virtue, simplicity has several things in its favor. First, there is the cognitive advantage that simpler theories are generally easier to understand, and so all else being equal a simpler theory is more likely an easier theory to explain and communicate. The practicality of using various theories, equations, data, and models depends on the availability of tools to manipulate these with–both their existence and their communication within a knowledge community. These practices themselves evolve to suit the needs of the knowledge community. It is in this regard that the computer is a striking example of a scientific apparatus that automates the labor of science. The notion of a computer simulation is, after all, the automation of the mathematical manipulations of numerical representations. In later chapters we will consider whether this constitutes a qualitative shift in the nature of these practices, but for now it is sufficient to recognize computational simulation as a powerful extension of these traditional scientific practices.

While there is a natural tendency to think that the goal of scientific knowledge is a complete understanding of some natural phenomena or system, this is misleading. Much of the power of knowledge, both in terms of general applicability and in ease of application, lies in managing the complexity of the world by simplifying it. The brain remains one of the most

complex objects of scientific study, for several reasons. First there are various levels of analysis from which one might begin studying it, folk psychology, behavior, physiology, etc. As a biological system, the types of cells, the vast number of neurons and the number and variety of their interconnections, as well as the electro-chemical processes that connect them taken together represent an astronomical complexity unknown in other sciences, even astronomy. The truly complex nature of mental phenomena implies that models will necessarily need to correspond to only limited aspects of any given mental phenomena. This is not a failure of the model, but a virtue:

> From this point of view we transfer from system to model to *lose* information. When the quantity of information is small, we usually try to conserve it; but when faced with the excessively large quantities so readily offered by complex systems, we have to learn how to be skillful in shedding it. Here, of course, model-makers are only following in the footsteps of the statisticians, who developed their techniques precisely to make comprehensible the vast quantities of information that might be provided by, say, a national census. "The object of statistical methods," said R. A. Fisher, "is the reduction of data." (Ashby 1972, p. 100).

The reduction of informational complexity is one of the most useful aspects of models. A book of tide-levels may be able to provide us with all of the information about tide levels that we may want, but it is not in a useful form. Instead we might prefer to compress this information into a set of representative equations from which we can compute the tide-levels at will. Moreover, those equations may do more than predict the levels, they are also likely to summarize the relationships between the key factors which determine tides. In the case of the brain, very little is understood about what factors were relevant to particular behaviors, diseases or phenomena, much less how they were interrelated, and thus any reduction of data to reliable correlations might provide insights into the brain's organization.

Chapter 3
On The Nature and Uses of Models II:
Models as Mediators Between Neurons and Behavior

The making of Images may have a variety of purposes, and the intention of the maker is not always clear to those who look on them. This chapter will be about mimicry of life, so it may be well to explain at the outset why a scientist resorts to methods which might seem more appropriate to the entertainer, the artist, or the priest. The suspicion that the scientist is not quite sincere in professing that his purpose is purely mechanical and illustrative goes a long way back. The notion of magic is deep-rooted. A term, fairly applicable to our subject, for instance, 'Electrobiology,' is found in Roget's *Thesaurus* (1946) listed under "acts of Religion," in the sub-section 'Sorcery.'

- W. Grey Walter.

3.1 Introduction

In the first half of the 20th century a small scientific movement emerged and took an unexpected turn, beginning a strange journey towards a mechanistic understanding of the mind. What made this journey strange was that much of it was not focused on humans nor animals nor brain tissues, but on machines–electronic models of the brain. This was a huge departure from traditional philosophical approaches to mind, as well as the empirical approaches of neurophysiology and psychopathology. Yet, it was these strange electronic brains which eventually proved to be central in forging fundamental connections between those different traditional approaches to the mind, and provided a sense that a comprehensive understanding of the mind, brain, and behavior was within the grasp of science. The ensuing cognitive revolution in the brain sciences depended upon these electronic brain models in an essential way–indeed the digital computer was in many ways itself conceived and constructed as such an electronic brain model and eventually served as the central metaphor for the brain. This chapter aims to understand the nature and role of these machines as working models. It was because these electronic brains were able to function as working models that they were so successful in advancing the mechanistic view of the mind.

While the Cybernetics movement of the 1940s and 1950s is often cited as a precursor to the cognitive revolution in the late 1950s brain sciences, its role is often seen as limited to providing general inspiration and some rudimentary theory. Rather, I believe that its role was far more significant, despite the fact that the particular details of its theories were largely discarded

by mainstream cognitive science in the following decades. In particular, little attention has been paid until recently to the devices that the cyberneticians actually built, and how they devised a novel scientific methodology around the construction of their synthetic brains. Roberto Cordeschi (2002) has written a history of the development of these devices beginning in the first years of the 20th century, and has described this new methodology as the "synthetic method." In his history, two devices stand out as exemplary cases of this methodology: the Homeostat and the Tortoise, designed by the British Cybernetics pioneers W. Ross Ashby and W. Grey Walter, respectively, in 1948. While Cordeschi (2002) presents much of the historical background for these devices, and recognizes the new methodology at work, he does not go so far as to provide an analysis of how the new synthetic method actually functioned, or why it succeeded. This is precisely what I aim to do in this article, and I believe the key to understanding the synthetic method lies in the epistemology of working models and their role as mediators between disciplines and theories.

In recent years researchers in science studies have begun to pay a great deal of attention to scientific models. As we saw in the previous chapter, there are many views of what models are, but they are traditionally associated with abstractions–mathematical or cognitive models–and are often treated as the metaphysical glue which binds theory to empirical data. In this chapter, I will not seek to supplant this view of models, which I will call "theoretical models," but rather will seek to add to it the concept of a different kind of model which also appears in science–the "working model." Working models differ from theoretical models in that they have a material realization, and are thus subject to manipulation and interactive experimentation. It is this dynamic material agency which sets working models apart from other closely related kinds of scientific objects such as theoretical models, simulations, and experimental instruments.

This is not to say that they lack any theoretical basis or significance, indeed their relevance to the broader field of scientific knowledge depends upon having some form of theoretical implication. However, unlike theoretical models they are not solely abstract constructs, nor merely the practical formulation of a theory for some specific application. By having a material face, working models participate in the world of material culture. In virtue of this, working models function in significantly different ways than theoretical models function.

Most importantly, working models exhibit what Andy Pickering (1995) calls material agency–in virtue of their materiality they behave in ways that are unintentional, undesired or unexpected. This is not always a negative feature, and it can often be essential to the productive contribution of such models–that they become oracles of new insights, novel phenomena, and serendipitous knowledge.

While I wish to single out working models for more careful scrutiny, I do not necessarily wish to argue for a strict distinction between them and other closely related scientific objects–such as simulations, instruments, and experimental apparatus that often share many of the same properties and participate in the same material culture of science. In the universe of scientific models, working models fall somewhere between simulations and instruments, or might even be seen as a sort of hybrid of the two. Clearly defining the boundaries between these, if such a project is possible, is well beyond the scope of the present chapter. It is useful, however, to note the similarities to other kinds of models described in the science studies literature, in particular discussions of simulations, scale models, and models as mediators can all help inform an understanding of working models. I will review some of these briefly before examining the case of working models in the brain sciences of the 1940s.

3.2 Models in Science Studies

> *From this point of view, there is no such thing as the true model of such a complex system as a cat's brain. Consider, for instance, the following four possible models, each justifiable in its own context:*
> *1. An exact anatomical model in wax.*
> *2. A suitably shaped jelly that vibrates, when concussed, with just the same waves as occur in the real brain.*
> *3. A biochemical soup that reacts biochemically just as does the cat's brain when drugs are added.*
> *4. A programmed computer that gives just the same responses to auditory stimuli as does the cat's brain.*
> *Clearly, complex systems are capable of providing a great variety of models, with no one able to claim absolute authority.* (Ashby 1972, p. 97)

Recently there has been a growing interest in the nature of computational simulations and their use in science. In one of the more carefully thought out analysis, Winsberg (2003)

examines the use of simulated experiments in physics. He identifies these simulations as a scientific practice that lies somewhere between traditional theorizing and traditional experimentation. In these simulations, theory is applied to virtual systems in an effort to test and extend the theory. Like experiments, simulations have what Hacking (1992) called a "life of their own." Hacking originally applied this notion to thought experiments, but it implies that there is an autonomous agency to the simulated experiments–they are not simply an expression or extension of the experimenters. For Winsberg, this autonomous agency expresses itself through the life-cycle of a simulated experiment–the recycling and retooling that goes into developing a useful simulation responds to the autonomous agencies expressed in the practices of simulation over time.

Working models might easily be construed as simulated experiments in this sense. That is, they too lie somewhere between theory and experiment in the traditional senses, and have a "life of their own." However, it is important to acknowledge the material basis of these simulations. While it may be tempting to conceive of computational simulations as virtual systems devoid of materiality, this is not really true. Computers are very sophisticated material artifacts, and their calculations are not ethereal or disembodied. A vast number of circuits and electrical currents are involved in realizing a computational simulation. But the nature of computational simulations, which are really an outgrowth of mathematical models, is to realize a mathematical formalism as precisely as possible. While certain mathematical techniques can never be realized by computers, the general aim of these simulations is to at least approximate the proper formalism. As such, the materiality of these computational simulations is seen as irrelevant to the theory being examined, even if it is necessary to consider when designing the simulation. And thus, the materiality can only be a potential cause of errors, rather than as a potential source of insight. Working models, by contrast, are more likely to embrace their materiality, rather than hide it. While they may start by realizing a mathematical theory, they often aim to demonstrate phenomena that may not be easily expressed by mathematical formalisms. And the "life" of the working model dwells heavily in the material realm, though it also gets expressed in the iterative redesign of models.

Perhaps the most insightful perspective on models in the recent science studies literature has been on their role as mediators. In the introduction to their edited collection on the subject, Morgan and Morrison (1999) articulated the notion of models as mediators:

> [W]e want to outline . . . an account of models as *autonomous agents*, and to show how they function as *instruments* of investigation. . . . It is precisely because models are partially independent of both theories and the world that they have this autonomous component and so can be used as instruments in the exploration in both domains. (Morgan & Morrison 1999, p. 10)

The idea that models are autonomous agents is crucial to a pragmatic understanding of scientific practice and material culture. In this statement we can also see that the notion of models they have in mind is one which stands in between theory and the world–whether the world is construed as phenomena, data or experiment. And we see in the various cases described in the volume that models are seen as offering a space of play in a conceptual realm where theory and world are not strictly related, but their inter-relation is in play through the mediation of the models. While this is a nice image, and captures important aspects of the presented cases well, it is not the only notion of autonomous agency operative in the working models of the brain sciences. There, the mediation is being done between multiple theories–the electronic brains mediate between theories of neurons and theories of behavior, not between theory and data–and the agency lies also within the material world.

Moreover, Morgan and Morrison's notion of autonomy is more concerned with the fact that models are *not determined* by theory or by data, that they have freedom from both. But all this comes to in practice is that scientists are able to manipulate models freely, where theories are constrained formally and data is constrained empirically. Models then become a plastic medium (at least to the extent that it is not constrained formally or materially) and is subject to willful manipulation by human agents. It is difficult to find much agency in the models themselves in these accounts, though the models do become the focus of much scientific work.

I am not the first to recognize the significance of synthetic models in the brain sciences. Not only did the scientists who devised those electronic brains engage in a critical reflection on the nature and use of their own models, but Cordeschi (2002) has written a wonderful history of the development of robots and mechanistic psychology up to and including Ashby and Walter. A

crucial element of that history is something he calls the "synthetic method." The synthetic method is supposed to stand against an analytic method–according to which understanding is obtained by isolating and manipulating various factors in a phenomenon until one can determine the control variables of the phenomenon and their causal effects. According to the synthetic method, scientific understanding can also be derived through constructing a complex model, and examining its properties and behaviors. There is, of course, an element of analysis involved in the determination of which aspects of the phenomena to synthesize in one's model, but the scientific practices of experimentation are instead focused on the synthetic model, rather than the original phenomenon. It is only when scientific practice is taken seriously and its pragmatic epistemic virtues are considered, that we can begin to consider what role the synthetic method plays in science.

I believe that the "synthetic method" and "models as mediators" approaches can both inform our conception of working models. It would seem that a complete explanation of the epistemic basis of the synthetic method requires an explanation of the scientific context in which such models are constructed. At least, this is the case if we wish to explain the construction of the synthetic brain models as contributions to the understanding of cognitive neuroscience in the 1940s. And in this case, we can only understand the role of these models by examining the mediations they achieved between the disjoint and incomplete theories being developed in the various disciplines studying different aspects of the mind, brain, and behavior.

It is important to recognize that there are multiple ways in which a model can serve as a mediator. While the models as mediators literature focuses primarily on the role of models as mediating between theory and data, there are at least two other sense of mediation which I want to focus on in this chapter. The first is the notion that a model can mediate between two theories. The theories involved could be very similar, or could refer to different entities, different types of entities, or the same phenomena at different levels of analysis. But in each case it is sometimes possible to build a model which includes key aspects of those theories, and shows how they might work in conjunction, reinforce one another, or even suggest that a theoretical synthesis of the two is possible. The second is that models, including the simulations and scale models just described and working models in particular, have the ability to "stand in" for natural systems and

phenomena during experimental investigations. This ability is an integral part of the synthetic method that Cordeschi describes. By building the synthetic brains according to certain *principles of construction*, these models were convincingly argued to exhibit certain theories of mind and mechanisms of behavior.

3.3 The Tortoise and the Homeostat

> *In order to study this* [feedback] *abstraction more easily, models have been built containing only two elements connected with two receptors, one for light and one for touch, and two effectors giving progress and rotation, with various possibilities of interconnection. This device is in the nature of a toy rather than a tool and reminds one of the speculations of Craik and the homeostat of Ashby*. (Walter 1953, p. 3).

W. Ross Ashby began designing the Homeostat in 1946, and gave his first public demonstration at a meeting of the EEG Society in 1948 (see Figure 1). His design actually began with a set of mathematical functions which exhibited a property Ashby called *ultrastability* (similar to what is now called *convergence*), meaning that iterated calculations would eventually result in stable, unchanging values. He then set about to devise a mechanical or electronic device that would exhibit this behavior. After various attempts he arrived at an electro-mechanical design that included four Homeostat units, each receiving an input current from each of the other three units and sending an output current to each of the other three units (Ashby 1952). The units themselves were built out of war surplus parts, mainly old radar units. Each unit consisted of a black box with four rows of switches, and water trough on top with a movable needle resting in the water. The state of the machine is displayed by this needle, and the machine is "stable" when the needle is in the center of the trough, and "unstable" when the needle moves to one end or the other, or moves back and forth. The currents from the other units are routed through a resistor, selected from a fixed set of resistors by a dial, to the water trough, where the needle responds by moving in relation to the current gradient created in the trough.

The behavior of each unit depends upon its specific configuration of switches and dials. In the mundane configuration, each unit merely displays the summation of currents by the position of its needle, and the needle is only stable, *i.e.,* in the middle of the trough, by

coincidence. However, when the unit is switched so that, instead of a simple resistor, its Uniselector circuit is engaged, it becomes ultrastable. In this configuration, when the needle is not in the center of the trough, a circuit is closed which charges up a coil (capacitor) that, when it passes a certain threshold, discharges to the Uniselector causing it to change to a new resistance (chosen from a pre-arranged randomized list of resistance values). It is thus a sort of random resistor and instantiates a trial-and-error search to find a resistance that stabilizes the needle (see Figure 2). The result is that the unit will continue changing its internal organization of resistance values until it finds an equilibrium with the needle in the middle of the trough.

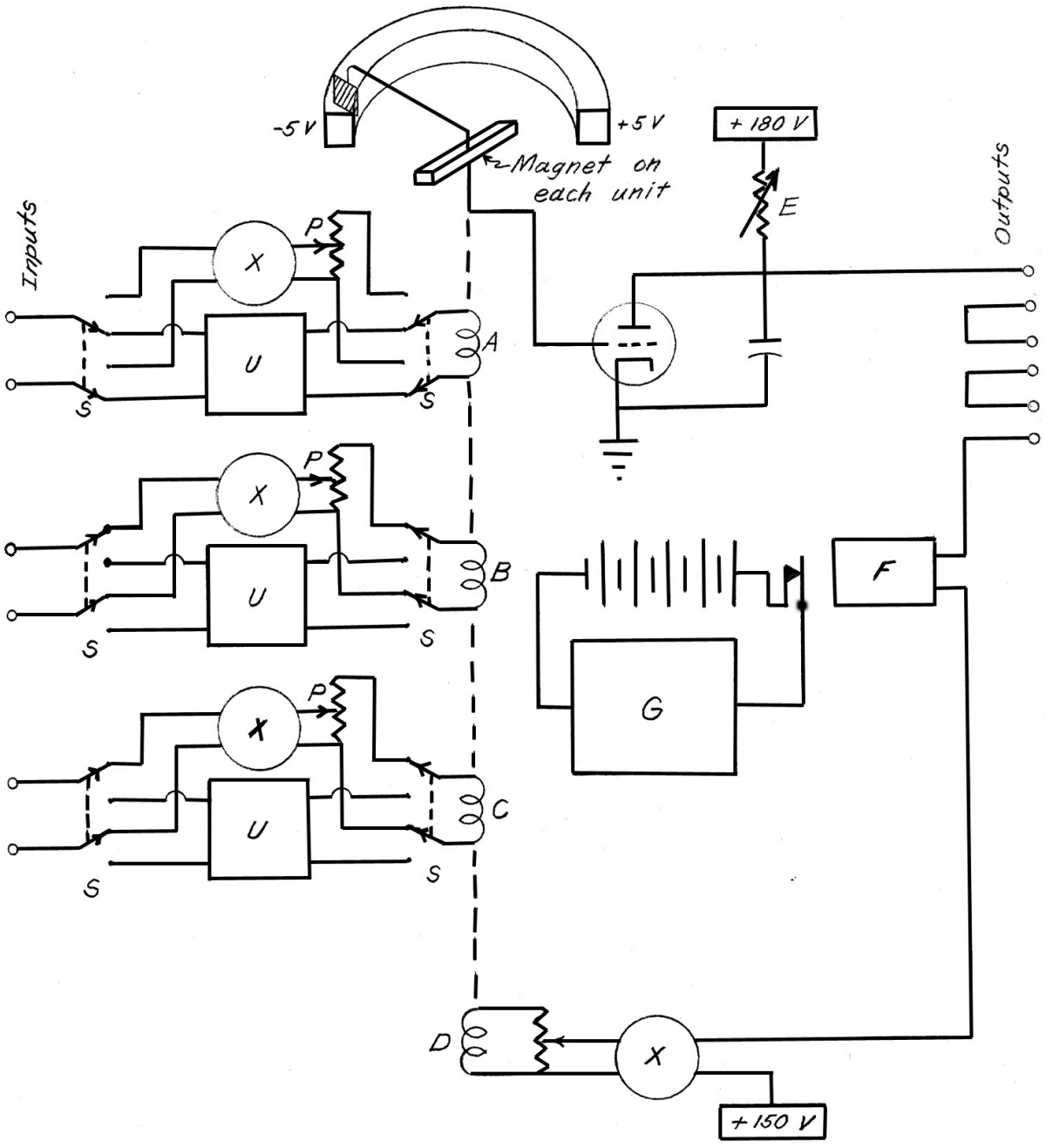


Figure 1. Schematic diagram of the circuit of one unit of the Homeostat (Babcock 1960, p. 40).

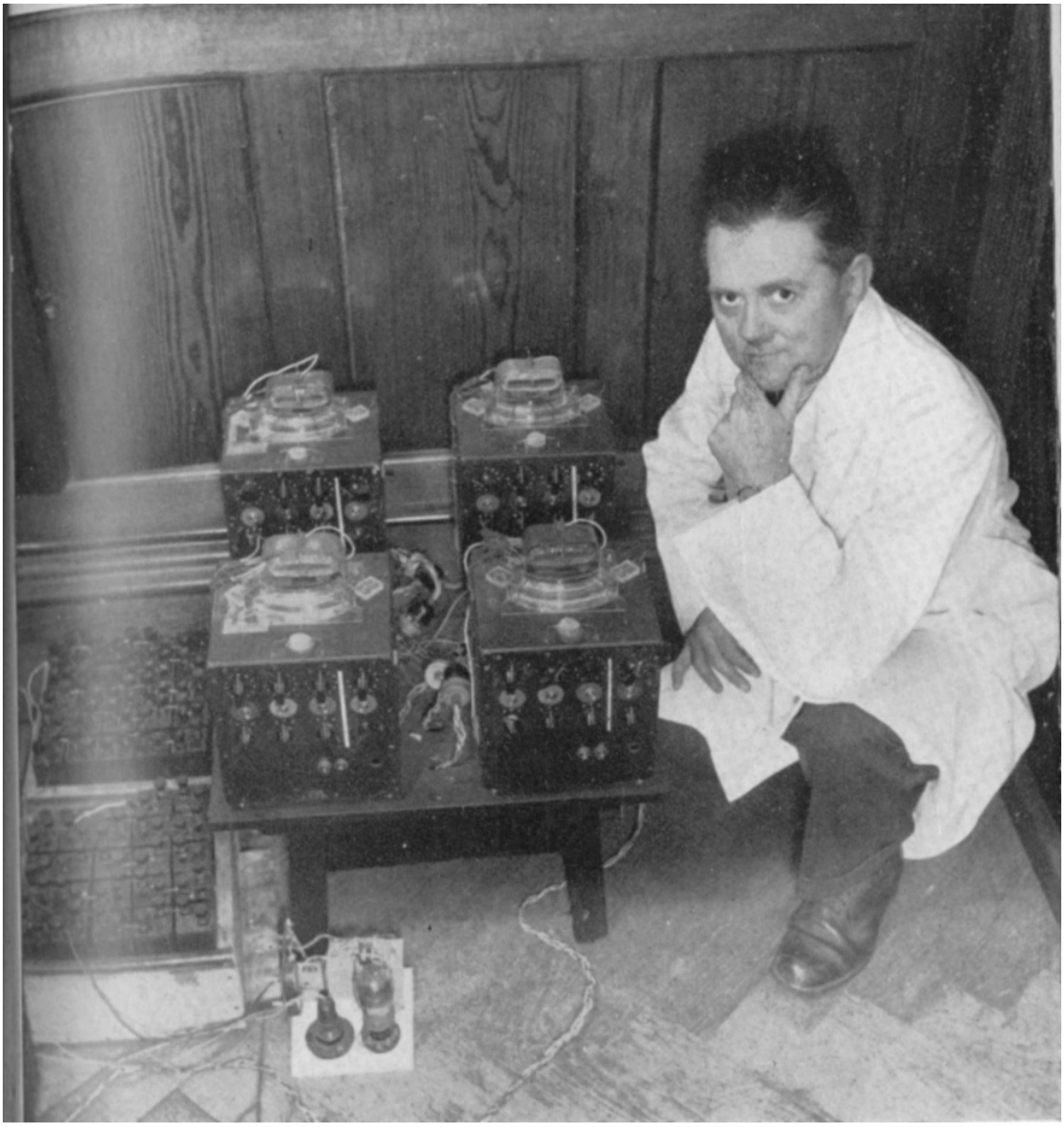
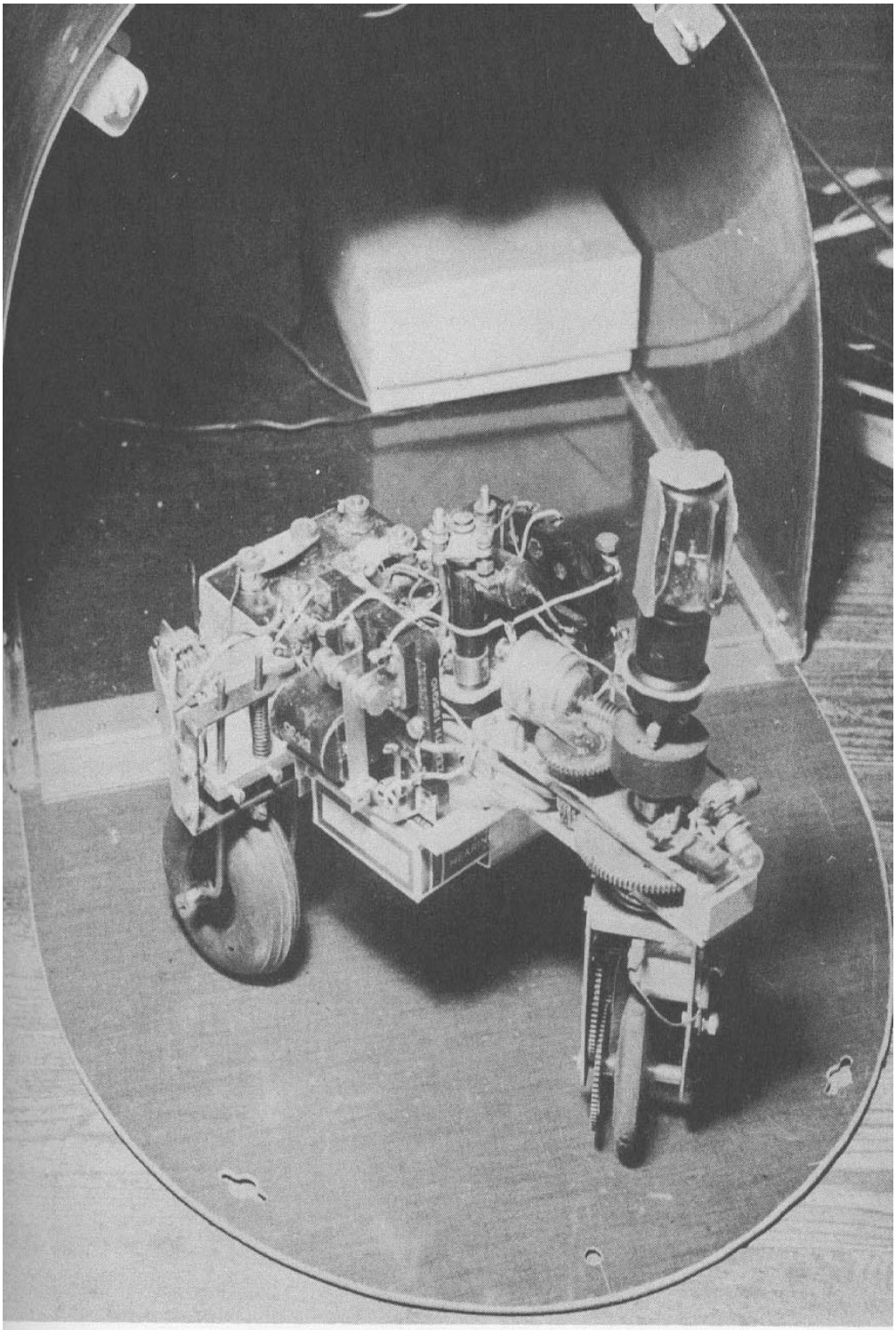


Figure 2. W. Ross Ashby with four-unit Homeostat setup (Eames & Eames 1973, p. 148).

Because the four units are interconnected, they must each find an equilibrium in the environment of inputs supplied by the other units, in other words they must all reach an equilibrium at the same time. If a particular state is unstable, a slight disturbance–like pushing a needle out of place–will throw the needles out of equilibrium and the unstable units will continue searching until they find another equilibrium state. Any number of additional disturbances can

be introduced, including switching a resistance value or an input's polarity, holding a needle in place, or even tying two needles together or to a rod that forces them to move in unison. These sort of interventions provide a basis for systematic experimentation with the device that we will consider shortly. Thus, by a mechanism of searching through possible reorganizations by random trial and error, the Homeostat will eventually find its desired equilibrium.

W. Grey Walter's Tortoise was a different animal than the Homeostat. In fact, Walter was fond of describing his Tortoise as an electronic animal, while the Homeostat was described as an electronic vegetable (Walter 1950). Still the devices were built in the same year, 1948, and exhibited a similar regard for the importance of feedback. The Tortoise was really a simple autonomous robot, based on a 3-wheeled chassis built from war surplus radar parts and the mechanical gears from an old gas meter (see Figure 3). The two rear wheels spun freely, while the front wheel was driven by an electric motor, and could rotate back and forth or all the way around a 360^{o} rotation by a second motor. Normally, both motors moved at a constant speed and direction, and the resulting motion of the robot was to travel in a spiraling sort of path, at least until it was disturbed. The two motors were subject to the combined control of two sensors. The first sensor was a simple contact switch between the tortoise-like shell of the robot and its base, such that any collision or contact with an object or obstacle would push the shell against the base and close the circuit. The Tortoise would react to this as a collision by reversing its drive motor. The other sensor was a photocell set atop the front wheel assembly, such that it always pointed in the same direction as the front wheel as it rotated about. As such, it acted as a scanning device, and reacted to bright lights. When the photocell received enough light, it would temporarily stop the motor that caused the turning (but not the drive motor) resulting in the robot driving in a straight path towards the bright light.



A photograph of the internal anatomy of Elsie, who belongs to the species *Machina speculatrix* created by Grey Walter.

Figure 3. Inside the shell of a Tortoise (de Latil 1957, p. 51).

3.4 Working Models and the Synthetic Method

Roberto Cordeschi's (2002) brilliant history of what he calls the *discovery of the artificial* tells the story of the building of behavioral models before, during and after the cybernetic era of the 1940s. Central to this story is the *synthetic method* and how it aided the restoration of psychology as a legitimate science through the construction of electronic, robotic, and computer simulations of the mind and brain. But how were these synthetic brain models able to do this? Cordeschi's account of the synthetic method draws upon the concept of a working model, but does little to define or explain it. He does provide one key to understanding working models with the notion of a model's *principles of construction*. While he has described the historical context and development of these models, I wish to consider what those developments can tell us about the epistemic nature and scientific use of models. I thus want to start where Cordeschi leaves off.

Cordeschi frequently stresses the importance of working models over other types of models. Unfortunately, he offers only brief comments on just what makes them so desirable or effective. He does, however, contrast working models with analogies:

> Mechanical analogies for nervous functions, however, occupy a secondary position in the present book, compared to the, albeit naive, *working models* of such functions which were designed or physically realized. Those analogies are discussed in some sections of the present book because examining them allows one to clarify the context in which attempts to build those working models were made (Cordeschi 2002, p. xvi).

Cordeschi thus finds the physicality of working models to be an empirically powerful force. Implicit in this is that actually realizing a model that "works" imposes some significant constraints on free-wheeling theorization. An important aspect of the constraints on building a model is that it must be designed, and the technical achievement of getting a physical model of this kind to work implies that the design is successful, along with the principles that underlie that design. We should also note in the above passage and throughout Cordeschi's book that he does not consider any hybrids which might lie somewhere in between working physical models and descriptive analogies. However, this is precisely where we might wish to place many computer programs and simulations–part material and part formal.

Cordeschi also sees that working models, as simulated experiments, can be used to test hypotheses:

> Working models or functioning artifacts, rather than scanty analogies, are the core of the discovery of the artificial, and in two important ways. First only such models, and not the analogies, can be viewed as tools for testing hypotheses on organism behavior. The way to establish the possibility that complex forms of behavior are not necessarily peculiar to living organisms is "to realize [this possibility] by actual trial," as Hull wrote in 1931 about his own models of learning. . . . The second way in which working models, rather than analogies, are at the heart of the discovery of the artificial [is that] . . .behavioral models can be *tested* (Cordeschi 2002, p. xvi).

Though it is not clear to me what the distinction between the two forms of testing is meant to be, there do seem to be two different kinds of demonstration going on. The first point that he makes is really that working models were a demonstrative "sufficiency" argument in the debates over mechanistic approaches to biology going on at the time. This kind of argument for models seeks to demonstrate that mechanisms of certain types are sufficient for certain complex behaviors. By demonstrating that an inorganic mechanism is capable of something assumed to be a strictly organic or biological function, one can falsify an argument to the contrary. This is a fairly weak form of argument, but was crucial at the time to defend the development of an empirical psychology and to take it in a mechanistic direction.

The second point in Cordeschi's discussion of working models is about scientific methodology–that they can be used to actually test theories. This is different from the ways in which theoretical models have been argued to serve in the confirmation of theories, however. Theoretical models are typically argued to serve as bridges between theoretical entities and empirical elements (data or observations). They thus offer explanations of experimental results by relating theoretical explanations to real phenomena. This is the traditional form of mediation performed by models–that between theory and data–and it obscures the work done in configuring material reality and setting up a controlled experiment in order to make the demonstration effective. Working models exercise their agency from the material domain and generate real phenomena which are also in some sense *simulations*. The metaphysics of simulations is rather complicated, especially insofar as one is tempted to argue that there is an easy distinction to be

made between "the real" and "the simulation" which has metaphysical import.[8] Of course, a working model (unlike theoretical models) actually *does something* and in virtue of that is part of a concrete dynamic material reality (where theoretical models remain in the realm of abstract static concepts). So our synthetic brains may not be real brains, but they are real electronic devices. In virtue of this they generate real phenomena which demand their own explanation. That is to say, the machines exhibit real behaviors, even if they are not real brains.

It is best at this point to return to our electronic brains and the ideas of the cyberneticians who built them. For it was clear to them that working models functioned simultaneously in different ways to further the emerging theory of mechanistic psychology. On the one hand, these models were demonstrations that such a theory could work, even if its details remained mysterious, and on the other hand the devices provided a basis for performing experiments and extending theory. We shall look at each aspect, and how the Homeostat and Tortoise mediated these roles.

3.5 Working Models as Demonstrations

I now want to explore the account of models and simulations offered by the cyberneticians themselves, and in particular the one offered by W. Ross Ashby. This will bring us back to the specific virtues of working models that Cordeschi is concerned with, as well as the significance of the principles of construction for such models. Ashby gave a great deal of serious thought to how biologically-inspired machines could serve scientific discourse. One of the key values of models, especially working machines, he arrived at was the vivid communication of specific scientific principles:

> **Simulation for vividness.** Simulation may be employed to emphasize and clarify some concept. Statements about machines in the abstract tend to be thin, unconvincing, and not provocative of further thinking. A model that shows a point vividly not only carries conviction, but stimulates the watcher into seeing all sorts of further consequences and developments. No worker in these subjects should deprive himself of the very strong stimulus of seeing a good model carry out some of these activities "before his very eyes." (Ashby 1962, p. 461).

[8] We will consider the relationship between working models and simulations in more detail below in Section 4.2.

As the use of the term "vividness" conveys, concepts are here considered to be like images, either literally or analogically. Much can be said about the metaphorical and analogical strength of theories, and it might be debated whether this compels acceptance of the theory, or aids in its application to new problems, or in fact does only superficial work. Vividness points again to the cognitive component of science, and that the ease of comprehension and understanding is important for the proliferation of knowledge.

The analogical strength of a model is somewhat disparaged by Cordeschi, in favor of the pragmatics of working models, but the value of a good analogy cannot be completely denied. Metaphors, analogies, and images have a relational component, the similarity relations that are used to bring out aspects of real phenomena. They also serve as a sort of connective tissue that draws together different ideas and shows their relation–analogy is itself a form of mediation. While much of the epistemic work lies outside the specific relation, the similarity relation serves an important role in organizing various practices–instantiating, demonstrating, measuring, verifying, extending, etc. A central metaphor or analogy, like isomorphism, can be the goal towards which those epistemic practices aim, and either achieve or fail to achieve. Models in this sense can be the compelling exemplars described by Kuhn (1962) around which scientific paradigms are organized. These are simultaneously demonstrations of accepted explanations of some controlled phenomena, and the basis of theoretical and experimental extension through puzzle-solving.

Ashby's notion of simulation for vividness is closely related to the notion of demonstration in science. Demonstration devices go back to the beginnings of modern science. Schaffer (1994) has written on the use of mechanical demonstrations of Newtonian physics as being crucial to their adoption in 17th century England. While much of his history focuses on the academic political matrix in which Newtonian physics sought to establish itself, it was the engineers who built the demonstration devices, and the showman-like scientists who demonstrated them that thrust Newton's mechanics on the English scientific community of the time. Indeed the Homeostat and Tortoise were often used by their builders to promote the concepts and promise of cybernetics to the public at large, and to fellow scientists. As a means to

popularize the emerging science of Cybernetics, the Tortoises were hugely successful, appearing at the Festival of Britain, on BBC television, and several *Life* magazine articles from 1950-1952.

While demonstrations often aim to convince skeptical members of the scientific community of some theoretical understanding, there is also a strong pedagogical function inherent in models. Their ability to convey ideas vividly applies not only to other members of the scientific community, but also to those being introduced to it as students, and the wider social world. By having students observe, or even better *build*, such devices, they come to appreciate the power of simple feedback mechanisms. Thus, the communicative power of working models was to be used not merely to convey new ideas to the existing scientific community, but to train and inspire a new generation of scientific researchers through the transfer of scientific and engineering *practices*. The pedagogical advantage of a working model, as opposed to a theoretical model, is that students actually learn laboratory techniques by building such a model themselves, and can interact with the finished model in a multi-modal way. That is, while some students learn best through text and some through equations, some may learn best through visual forms, others through tactile forms. Hands-on experience conveys practical and often tacit knowledge in ways that formal expressions of knowledge can rarely achieve.

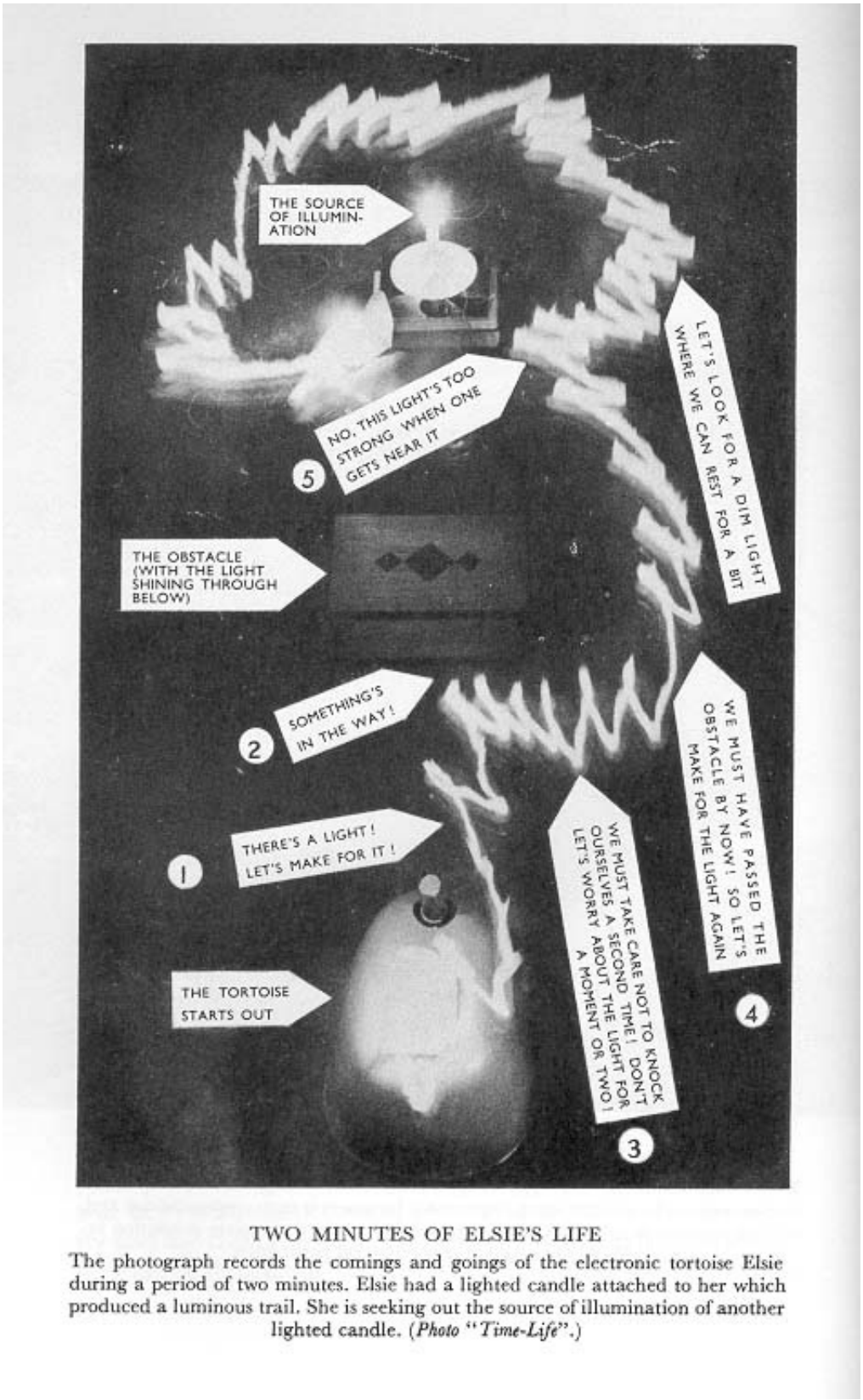
The Homeostat was, as its name implied, intended to demonstrate the biological principle of homeostasis. According to this principle, an organism would adjust its condition through whatever means were accessible in order to maintain certain critical conditions within the organism. If an animal was cold, it would seek out a warmer place, if its blood pressure got too low, it might increase its heart rate, etc. Ashby's insight was that a general purpose learning mechanism could be derived from this principle. If an organism or machine were allowed to search randomly through its possible actions in the world, it could find for itself ways to maintain the critical conditions of its internal environment. This would work even when the system is disturbed by unexpected outside influences, including malicious experimenters. The bottom line is that a mechanism with a feedback-induced random search could exhibit the same principle of homeostasis that living creatures did. It was thus an adaptive system, and demonstrated Ashby's extension of it to the concept of ultrastability.

Indeed, in reading *Design for A Brain* (Ashby 1952) it is often difficult to distinguish Ashby's empirical arguments from his pedagogical illustrations. He moves so rapidly between arguments in favor of the homeostatic principle as being an explanation of certain aspects of the brain's behavior, to analogies between the behavior of the Homeostat and various brain phenomena, to arguments that the Homeostat embodies the principle of homeostasis, that these all seem to hang together. It is difficult to say whether this style of rhetoric is a contribution to the synthetic method or a product of it, as the distinctions between brain, theory, and working model all begin to blur.

Unlike Ashby's Homeostat, which was built on a explicit set of formal equations, W. Grey Walter's robotic Tortoises were built upon a rather loose set of principles of construction, and his work on these models correspondingly focused more on the phenomena that they generated than on the demonstration of any specific underlying principles. The general principle that was demonstrated was simply that a small number of feedback loops, two actually, could generate a large number of behaviors through their interactions with one another and the environment (see Figure 4). Walter frequently argued that much of their value lay in the fact that it was a *demonstration proof* that simple mechanisms could exhibit complex biological phenomena. But it was the character and range of behaviors that he focused on.

In particular, he argued that the various behaviors of his robots were most easily described using biological and psychological terminology. The biological principles that he claimed could be found in these various behaviors included: parsimony (they had simple circuits, yet exhibited numerous and apparently complex behaviors), speculation (they explored the world autonomously), positive and negative tropisms (they were variously attracted and repulsed by lights of different intensities, depending on their internal states), discernment (the machine could sublimate long-term goals in order to achieve short-term goals like avoiding obstacles), optima (rather than sit motionless between two equally attractive stimuli, like Buridan's ass, they would automatically seek out one stimulus, and then perhaps the other), self-recognition (by sensing and reacting to its own light in a mirror), mutual recognition (by sensing the light of another Tortoise), and internal stability (by returning to their hutch to recharge when their batteries go low) (Walter 1953).

The Homeostat was a demonstration proof of the power of the homeostatic principle to explain brain phenomena. The principle of homeostasis was already well known in biology, and feedback controllers were widely studied by electrical and mechanical engineers. What was new to the Homeostat was the notion that a set of four interconnected feedback controllers could not only be stable, but would always tend towards stability if each unit were allowed the additional



TWO MINUTES OF ELSIE'S LIFE

The photograph records the comings and goings of the electronic tortoise Elsie during a period of two minutes. Elsie had a lighted candle attached to her which produced a luminous trail. She is seeking out the source of illumination of another lighted candle. (*Photo "Time-Life".*)

Figure 4. Time-lapse photo of Tortoise behavior (de Latil 1957, p. 275).

capability of randomly changing its relation to the others. Marvin Minsky (1961) would later credit Ashby with originating the concept of random search that was exploited in much of AI research. His device coupled two key theoretical concepts: behavioral adaptation, and trial-and-error search, and thereby served as a mediator between psychological and computational theories of the brain and behavior. While not explaining any new data, or relating data to theory, this model actually mediated between theories in two disciplines, and at two levels of abstraction. In doing this, it provided a conceptual bridge between the disciplines.

But did this *require* a working model? From a purely theoretical perspective, the Homeostat proved nothing that could not be shown mathematically on paper. There have subsequently been formal proofs for the convergence (or lack of convergence) of a great many search algorithms and random search strategies. In fact, Ashby developed the mathematical equations defining the Homeostat before he began designing the device to realize them, and spent several years attempting various schemes for constructing a device which would both conform to these equations and provide a vivid demonstration of its underlying principles. But unlike a purely mathematical demonstration, the Homeostat made it vividly explicit just how such a random search could be coupled to behavior in an explanatory way.

And so the initial interest and most striking aspect of these working models of the brain was their very existence and demonstration of the fundamental principles of mechanistic psychology. But the real power of these models lay in their ability to be subjected to experimentation. In part this was because they were more efficient than mathematical analysis at a time when the first digital computers were still under construction. Eventually, the increasing power, ease of use and accessibility of computers would make digital simulations far more attractive than the construction of analog machines for many purposes. But these electronic brains also allowed for a material practice for the brain sciences without the usual difficulties of dealing with real brains. While there was work done on primarily non-human or pathological human brains, or individual cells, the study of the living brain was ill-equipped to answer many of the questions about the fundamental structure and organization of information in the brain. It was the electronic brains that provided a route to an experimental inquiry into these questions.

3.6 Working Models as Experiments

> *The outlook for an experimental model brightened at once with the problem reduced to the behaviour of two or three elements. Instead of dreaming about an impossible 'monster', some elementary experience of the actual working of two or three brain units might be gained by constructing a working model in those very limited but attainable proportions.* (Walter 1953, p. 109).

Now that we have considered some of the often overlooked demonstrative functions of models, it is time to consider the more traditional roles of models in scientific explanation and verification, and what working models can offer here that theoretical models cannot. It is these aspects which Cordeschi suggests are the specific advantage of working models over "mere metaphors." Ashby characterized this function in terms of deduction and exploration:

> **Simulation for Deduction and Exploration.** Perhaps the most compelling reason for making models, whether in hardware or by computation, is that in this way the actual performance of a proposed mechanism can be established beyond dispute. . . . Note, for instance, the idea that the molar functioning of the nervous system might be explained if every passage of a nervous impulse across a synapse left it increasingly ready to transmit a subsequent impulse. Such a property at the synapse must impose many striking properties on the organism's behavior as a whole, yet for fifty years no opinion could be given on its validity, for no one could deduce how such a system would behave if the process went on for a long time; the terminal behavior of such a system could only be guessed. (Ashby 1962, p. 463).

There are several ideas packed into this passage that bear commenting upon. The first thing to note about this passage is that it points to how the availability of a specific set of material practices influences what can be tested, and thus what can become knowledge. Even though some of the mathematical techniques for testing various hypotheses might be available, the work required to carry them out often was too great to be realized. The more important aspect is that working models support the empirical testing of certain hypotheses. Even though the applicability of the hypothesis to the brain might be tenuous, it is still possible to make progress in the details of theory using such models.

As one popularizer of cybernetics put it:

> When we have thus constructed a model which seems to copy reality, we may hope to extract from it by calculation certain implications that can be factually verified. If the verification proves satisfactory, we are entitled to claim to have got nearer to reality and

perhaps sometimes to have explained it. (de Latil 1957, p. 225).

Whether this constitutes explanation or not might be stretching things, but there are certainly characteristics of a working model that enliven speculation into both the structure and organization of the natural phenomena, and the possible extensions to the working model. I believe that it is the ability to move back and forth between deductions from theories and extensions to models that makes the synthetic method such a powerful research methodology. Though it is in some sense dependent on antecedent and subsequent analytic methods, it greatly enhances these by offering new ideas and directions for extension when the analytic methods are at a loss. This also helps to explain why it is often difficult to separate the scientific from the technological advances made by the synthetic method–the two become intertwined as technoscience.

There are two senses in which a model can be extended. The first sense does not necessarily involve any changes to the model itself:

> Once the model has been made, the work of the model-maker has reached a temporary completeness, but usually he then immediately wishes to see whether the model's range of application may be extended. The process of extension, if we are to stay within the framework of the ideas expressed in this paper, will be subject to just the same postulate as the other processes of selection; for, of all possible ways of extending, the model-maker naturally wants to select those that have some special property of relevance. Thus, a model of the brain in gelatin, that vibrates just like the brain under concussion, is hardly likely to be worth extension in the biochemical direction. From this point of view the process of extension is essentially an exploration. So far as the worker does not *know* the validity of the extension, to that degree must he explore *without* guidance, i.e., "at random." (Ashby 1972, p. 110).

Extension can consist of two different processes: validation (or verification) and exploration. In the first sense, a new model is extended by showing that it models a wider range of phenomena than first believed, or it is modified to achieve this effect. In the second sense, a model can be explored to find new kinds of phenomena in it, or built upon and altered to produce novel phenomena. In exploration, one does not seek out some specific correspondence between model and world, but rather one treats the model as itself the object of investigation and extension.

If a model actually generates a genuine behavior in the world, that is a phenomenon itself to be explained. It is not, in this sense a prediction about what might happen, it is itself a

happening. Moreover, as in experimental apparatus, phenomena can be played with, changed slightly without any particular expectations as to what new phenomena may result, just to see what might happen. In this sense it is not predictive, nor does it necessarily attempt to fit the data obtained from some other phenomenon. It does not tell you what you might find in the world, it *is* what you find. There is a connotation to the use of the word "experiment" which means open-ended exploration of this sort–a "playing around" or fiddling with things just to see what might happen (Pickering 1995). A material model like the Homeostat can be played with by students and novices as well as experts to reveal all sorts of interesting phenomena. Hypotheses about its behavior can also be rigorously tested, with no change in the nature of the model.

Ashby performed a number of experiments on the Homeostat, both to prove that it could do what he had hypothesized it could, and to discover what else it was capable of. In terms of demonstrating that it was a valid model of the nervous system, Ashby replicated on his Homeostat a number of experiments performed previously on living systems. Among these was an experiment by Sperry in which the muscles in the arm of a monkey had been surgically cut, swapped around and reconnected such that they would now move the arm in the opposite direction than they previously had. In the experiment, the monkey relearns quite quickly how to control its arm in this new configuration. To replicate the experiment, Ashby used a switch on the front of a Homeostat unit which reversed the current passing through it. He then noted in the trace of the needle's movement, the exact opposite reaction to incoming current as before, resulting in instability. However, after a short search by the Uniselector, the unit was able to stabilize once again (Ashby 1952, pp. 105-7). This sort of experiment was of the most basic kind.

More complicated experiments were also performed. These included conditioning experiments in which the needle was manually manipulated by the experimenter as a form of "punishment" (Ashby 1952, p. 114). Ashby also describes an experiment in which he changes the rule governing punishment systematically, sometimes enforcing one rule, and sometimes the other. From this he observes that the device is able to adapt to two different environments simultaneously (Ashby 1952, p. 115). He also made new discoveries of the machine's capabilities through these explorations. These included ways of materially manipulating the

machine that were made possible by its design, but were not intended features of the design. Most significant among these was the possibility of tying together with string, or binding with a stiff rod, the needles of two of the units. The result was to force them to find an equilibrium under the strange constraint of their material entanglement, as well as their electronic couplings (Ashby 1952, p. 117). The Homeosat was quite capable of finding such an equilibrium. The other aspect of the material device that was not foreseeable on paper was its particular temporal dimensions–how quickly the Uniselector flipped, how quickly it found a new equilibrium, etc. Ashby also experimented with the delay in timing between the unit's behavior and the administration of punishments in conditioned training (Ashby 1952, p. 120). These aspects of the machine were certainly incidental to its conception, but quite significant to is real operations and provided the basis for further exploration of the Homeostat and its behavioral repertoires.

Ashby went on to develop a much more sophisticated synthetic brain, the Dynamic and Multistable System, or DAMS. While he had great hopes for DAMS, the technical design, cost, and above all the complexity of behavior that the device exhibited frustrated Ashby for years. As Pickering (forthcoming) has recounted in his analysis of the ill-fated machine, there was also a clearly developmental aspect of his work on DAMS. Beginning with his great expectations for the device, Ashby's notebooks reveal the various causes of his frustration, from the problem of having to devise a clearer notion of "essential variables" than had sufficed for the Homeostat, to the technical difficulties of building such a large and complicated device, to the costs running into his lack of research support funds, to Ashby's inability to comprehend the patterns of behavior exhibited by the complex device, his understanding of the brain and what he was trying to do with DAMS evolved dynamically over time. It is clear that he was not certain how its would behave specifically, but only sought a general sort of performance, yet what he got from it was still frustratingly complex, so much so that after nearly seven years of work, he abandoned it, and largely edited it out of the literature.

Walter's Tortoises were also subject to a number of experiments and explorations, and yielded unexpected results. Primarily the unexpected results were due to interactions between the robots based on their lights and light sensors:

Some of these patterns of performance were calculable, though only as types of

> behaviour, in advance, some were quite unforeseen. The faculties of self-recognition and mutual recognition were obtained accidentally, since the pilot-light was inserted originally simply to indicate when the steering-servo was in operation. . . . The important feature of the effect is the establishment of a feedback loop in which the environment is a component. (Walter 1953, p. 117)

What seems most clear from Walter's various discussions of the Tortoises is that he believed they were physiological models of biological phenomena. Of course, he was not claiming that the mechanisms which produced the behaviors in animals were identical to, or even isomorphic to, the circuits of the Tortoises, but just that those behaviors could be produced by simple circuits which employed feedback and modulation mechanisms similar to neural circuits–*i.e.,* that they shared some relevant functional properties. An important scientific insight lay in the fact that very simple circuits could produce such complex and interesting phenomena provided that they instantiated these functions:

> The electronic "tortoise" may appropriately be considered as illustrating the use of models. But models of what? In the first place they must be thought of as illustrating the simplicity of construction of the cerebral mechanisms, rather than any simplicity in their organization. . . . In short: reality may indeed be more complex than the model, but it is legitimate to consider that it may be equally simple. In a matter of which we know so little such a hypothesis is not without importance. (de Latil 1957, pp. 227-8).

The Tortoises were thus used as a basis to argue against those who hypothesized that neural circuits were necessarily or interminably "complex" because the behavior of organisms was complex.

Walter stopped short, however, of saying that these qualities were purely in the eye of the beholder. Instead, the phenomena of life and mind were posited to subsist in the negative feedback loops which held the organism in various stable configurations even as it moved through a dynamic environment. For these biologically-inspired and inspiring machines to become legitimate scientific models, they would have to do more than appear lifelike and evoke curiosity and wonder.

Despite the fact that they lacked a detailed theory or explained any specific set of data, these synthetic brain models produced useful and illuminating contributions to brain science. What becomes clear from looking at the history of these electronic models of the brain in the

1940s are some of the ways in which they served as mediators between relatively vague theories and informal, often impressionistic, observational data, ultimately serving as a stepping stone to more formal theories and rigorous observations. In short, they did this by providing models around which new theorization and observation could be organized. As a result, these rather simplistic models served as mediators between initially vague and unrelated theories, and more concrete and coherent theories of brain organization and behavior which followed.

The Tortoises and Homeostat succeeded in doing this because of their ability to demonstrate and communicate scientific knowledge in practical and accessible ways. These are pragmatic elements of epistemology, which is not simply a matter of truth and justification, but also depends upon pragmatic elements of the retention, transmission, and utilization of knowledge. This is what makes models so crucial to the epistemic efficacy of scientific knowledge. Such issues do not arise in traditional philosophy of science, which considers only the individual mind seeking justified theories, but do become crucial once we take a view of knowledge production as a social practice, in which multiple agents must generate knowledge together and transmit it to others. In these processes, issues of communication are not isolated from issues of knowledge, but are instead crucial to it. It is in this regard that the ability of a model to store and transmit knowledge, as image, idea and practices, become relevant. Ashby and Walter were both self-consciously aware of this fact.

It should also be clear that these examples of working models employ the synthetic method, and succeed scientifically in virtue of being mediators of a particular sort. Both the Homeostat and Tortoises are examples *par excellance* of the synthetic method–machines built according to specific principles of construction to offer a proof by example that a mechanistic approach to the brain could describe mechanisms that produce plausible behaviors. In the case of the Homeostat, it was shown that merely seeking equilibrium points and stability through random trial-and-error searches could lead to adaptive behavior. Whether a "Uniselector" was exactly the same mechanism that real creatures used to adapt to their environments was not at issue because of the level of abstraction at which the model operated. But it succeeded in forging a powerful conceptual linkage between learning and random search which still drives much of the research in AI and machine learning. In this sense the model was successful *because* it was a

mediator between theories in biology (homeostasis) and theories in engineering (random search).

The Tortoises, with their elaborated set of behaviors demonstrate another aspect of how working models act as mediators within the synthetic method. Once a few major proofs by demonstration are achieved, it becomes necessary to explore and extend the working model. Walter did this both by extending the model to new psychological phenomena by observing these phenomena in the behavior of the Tortoises, but also by improving and extending the machinery of the working model. He built versions of the Tortoise which incorporated an adaptive learning mechanism, called CORA (Walter 1953). Using this mechanism he "trained" his robot to respond in certain ways to a whistle by building an association between the whistle and another stimulus, much like Pavlov's dogs. Here the synthetic method works by drawing on the constraints of the real world–we could imagine all sorts of mechanisms that "might" produce such-and-such behavior, but here is one that actually *does it* in this working model. Here the model is mediating in the more traditional sense, between theories and empirical data. But the empirical data are the behavior of the model itself, and the extension of the range and similarity of those behaviors to the target phenomena, animal behaviors, further reinforces the mediating bridge between theories, in this case psychological theories of behavior and interacting feedback control mechanisms.

3.7 Conclusions

> *What we have just reduced to absurdity is any prospect of reproducing all its elaboration of units in a working model. If the secret of the brain's elaborate performance lies there, in the number of its units, that would be indeed the only road, and that road would be closed. But since our inquiry is above all things a question of performance, it seemed reasonable to try an approach in which the first consideration would be the principles and character of the whole apparatus in operation.* (Walter 1953, p. 107)

While the shift in biological and psychological sciences to the study of synthetic brains and mechanical systems was strange to many, it raises the question whether the synthetic method thereby constituted a new scientific epistemology. It is clear, however, that methods of demonstration, and even elaborate pedagogical mechanisms have played a role in scientific debate and the establishment of knowledge since the dawn of modern science. The real question

is how to describe scientific development in ways that respect the valuable roles played by such devices.

As simulations, or stand-ins, working models were not completely new, but they were new to the brain sciences. The brain and mind sciences faced a challenge in the first half of the 20th century to become "empirical" through the use of the experimental method and to distance themselves from "metaphysical" forms of explanation. Physics was at the time considered to be the science which other sciences should emulate methodologically, especially if they wanted to bolster their empirical legitimacy. There were two severe difficulties in doing this which were ultimately overcome through the synthetic method. The first was a requirement for the observability of the data-generating phenomena upon which theories were based. The second difficulty was that real functioning brains were very difficult to work on for both technical and ethical reasons. The result was a behaviorist psychology that largely ignored the inner workings of the brain, and brain sciences that focused on the physiology of single or small numbers of cells, or the gross anatomy of mostly dead brains. The exception to this was psychiatry, which confronted sick and damaged brains on a regular basis, and managed to intervene in rather drastic ways upon them, in the hope of both curing the unhealthy brain and understanding the healthy brain.

As empirical experiments, working models offered a new basis for proceeding in areas of brain science that were otherwise difficult to address using the analytic "controlled experiments" method of physics because individual variables could not be so easily isolated. Despite efforts to shore up the epistemic basis of the brain sciences, the resulting approaches did not quite manage to integrate the various levels of analysis in a compelling way. It was the cybernetic brain models which managed to do this. But why should we think that building working models should be more useful than experimenting on neurons directly, or with the conditioned behavioral patterns of living animals?

This might be read as the "paradigm shift" from behaviorism and traditional neurophysiology to a computational cognitive neuroscience. It might also be possible to view this history as instead seeking a rigorous mathematical and mechanical account, and as a consequence of this arriving at theories which applied equally to the organic and inorganic. But

there is another aspect of the use of these models that is a central theme of this dissertation. It is the notion that these models acted as bridges or mediators between existing theories at different levels of analysis. Cordeschi acknowledges that working models established a new intermediary level of analysis between behavior and physiology. However, he emphasizes the autonomy of this new level of analysis, rather than its constructive and mediating aspects:

> The earliest behavioral models, designed as simple physical analogs, began to suggest that there might exist a *new* level for testing psychological and neurological hypotheses, one that might coexist alongside the investigations into overt behavior and the nervous system. This was the core idea of what Hull and Craik had already called the "synthetic method," the method of model building. This idea comes fully into focus with the advent of cybernetics, and especially when the pioneers of AI, who wanted to turn their discipline into a new science of the mind, found themselves coming to grips with the traditional sciences of the mind–psychology and neurology–and with their conflicting relationships. That event radically affected the customary taxonomies of the sciences of the mind. (Cordeschi 2002, p. xviii)

For present purposes it is less significant what level of analysis is claimed to be the right one, as it is how a level of analysis is supposed to connect up to other empirical theories at different levels. The synthetic method was one way of establishing working models as mediators, mediators between theories espoused by very different disciplinary traditions, and mediators between very different sorts of phenomena and material performances.

The issue facing the mind sciences was how to proceed in devising and testing hypotheses about the biological basis of psychology. The first step in doing this was to give a plausible account that bridged low-level physiological mechanisms and high-level behavioral mechanisms. But there is a way of viewing the synthetic brains in which they might again appear puzzling. One can arrive at this puzzlement by noting that even while the brain sciences were seeking to construct a sound empirical basis for their work, they did this by positing a new level of analysis of the mind that was not itself directly observable. From this perspective, the synthetic brains are models of *mental functions*, not behavior or neurons, but rather functions that are not themselves directly observable. In fact, it is hard to conceive of just what entities these models were supposed to be modeling. The way out of this puzzle is to understand the empirical power of the mediation that they were performing between different levels of analysis. The two levels of analysis between which these new brain models tried to span were the traditional sciences of

behavior and neurons, and in doing this they were able to combine the empirical force of both fields to bolster the hypothesized bridge between them. Without an existing theory of the interaction of the principle levels of analysis in the brain sciences, it was the particular advantage of the working synthetic brain models that they instantiated certain theories about *neurons* and were able to display characteristic *behaviors* as a consequence. It was the models themselves which mediated between the levels, not theoretically, but through their working and exhibiting behavioral phenomena that could stand in similarity relations to animal behavior.

Even if we take a view of science as ultimately about "knowing that," rather than "knowing how," working models are a practical means to knowledge and may very well be dispensable once the fruits of knowledge are collected. Indeed, it can often be difficult to recognize and express just how models have exerted their agency from this perspective. Yet, the moment we stop to ask why certain techniques and technologies are in place, and how they arrived at the forms they take, we would be at a loss to explain these in the absence of the history of material culture. In principle, no technological solution in use is the only one possible, and only rarely is it an optimal solution. At best, the technology in use is the best available, where the criteria of choice and the alternatives to choose among have evolved over time and in interaction with one another. And thus history is the best way to get at such questions, and there is a sense in which every technological artifact is an archive of its own history–though it may have been subjected to much cleansing and scrubbing to remove its intrinsic historical traces. At the very least, we better understand a technology by knowing its history, as well as its structure.

It is clear that synthesizing and experimenting on working models can achieve many kinds of mediation useful to scientific progress. Working models have received far less attention in the philosophy of science literature than theoretical models have. Yet it seems clear that the pragmatic virtues of models function in many areas of scientific practice. It remains for further study to see to what extent working models can be found in other areas of science, and what roles they play as mediators in those areas. Construed broadly enough, one can find working models all over science. Yet it seems that the working models in each science are more or less unique to the science in which they are found. Indeed, it is their specificity to the scientific practices in which they are embedded that makes them both unique and interesting as a means to studying

those practices. In this sense it may be undesirable to seek out any "general theory" of working models. Rather it seems more opportune to view these models as an obvious point of entry into studying the material culture of science. By asking such questions as "Why are the models built? How are they built? What kinds of experiments are performed? Which models are kept and which are discarded? How are they copied and developed over time and propagated through space?" we might hope to get at the very essence of the material culture of scientific models.

Chapter 4
Computers as Models of the Mind:
On Simulations, Brains, and the Design of Computers

After the war, together with a small group of selected engineers and mathematicians, Johnny built, at the Institute for Advanced Study, an experimental electronic calculator, popularly known as *Joniac*, which eventually became the pilot model for similar machines all over the country. Some of the basic principles developed in the *Joniac* are used even today in the fastest and most modern calculators. To design the machine, Johnny and his co-workers tried to imitate some of the known operations of the live brain. This is the aspect which led him to study neurology, to seek out men in the fields of neurology and psychiatry, to attend meetings on these subjects, and, eventually, to give lectures to groups on the possibilities of copying an extremely simplified model of the living brain for man-made machines.

- Klara von Neumann

Turing knew perfectly well what the job he had to do, which was to manufacture or design a machine that would do the complicated sort of mathematics that had to be done in the Mathematical Division of NPL. But he had all sorts of interesting things that he liked to do: for example, he was really quite obsessed with knowing how the human brain worked and the possible correspondence with what he was doing on computers. . . . Turing thought that the machine should be made quite simple, and at the same time should make everything possible that could be done. His particular purpose was to permit the writing of programs that modify programs, not in the simple way now common but rather in the way that people think.

-Ted Newman

4.1 Introduction

The purpose of this chapter is to clarify some of the important senses in which the relationship between the brain and the computer might be considered as one of "modeling." It also considers the meaning of "simulation" in the relationships between models, computers and brains. While there has been a fairly broad literature emerging on models and simulations in science, these have primarily focused on physical sciences, rather than the mind and brain. And while the cognitive sciences have often invoked concepts of modeling and simulation, they have been frustratingly inconsistent in their use of these terms, and the implicit relations to their scientific roles. My approach is to consider the early convolution of brain models and computational models in cybernetics, with the aim of clarifying their significance for more current debates in the cognitive sciences. The discussion of the current issues will be deferred until the next chapter, though it is strongly foreshadowed here. It is my belief that clarifying the

historical senses in which the brain and computer serve as models of each other will clear some ground for considering the debate over the future directions of cognitive science.

"Model" is a challenging concept, in part because it is both a noun and a verb, and it takes many prepositional forms–*X* models *Y*, *X* is a model of *Y*, *X* is a model for *Y*, *Y* is modeled on *X*, *Y* is modeled after *X* and even *Y* models for *Z*.[9] Thus, we could say that the brain was a *model for the structure* of computer (or the computer was *modeled on* the brain) in the sense that the designers of the early computers, such as John von Neumann, treated the biological brain like an artist's model, and crafted the computer in its image. Of course, once complete we might be inclined to think of the artist's sculpture as a sort of model of the subject it is based upon–and so we might think of the computer as a *model of* the brain. This would seem to be the case for certain electronic devices that were *models of the performance* and behavior of the brain, such as W. Ross Ashby's Homeostats which were meant to be models of the adaptive properties of the brain, or W. Grey Walter"s Tortoises which were meant to be models of the dynamic drives of the living brain. In yet another sense, the first computers were seen by some, such as Alan Turing, as *models of the fundamental structure* of the brain in the sense that the digital computer was an engineered device that worked on the same principles as the brain and could be used to test theories about how the higher level functions of the brain might operate. For each of these cases, the actual process of modeling, the construction of each device, was unique and complex, and involved further models of the brain and behavior, both concrete and abstract. We have already considered two large classes of models: theoretical models and working models. We have also considered the case of "modeling" as a particular kind of scientific practice with goals other than producing theories. This chapter will add to these another class of models, the "simulation," which is a term used in nearly as many ways as "model." For many authors, a "simulation" is simply another word for "model," while for others it is what I have been calling a "working model," but in its more precise usage "simulation" generally refers to a special class of models defined by their use–specifically computational models used to approximate the behavior of a system of equations that are too difficult to solve by analytic techniques. Because of the computer, simulations have come to be one of the most significant kinds of models in scientific

[9]In these examples I have tried to make *X* the model, *Y* the thing modeled, and *Z* the agent doing the modeling.

practice. Yet they occupy a strange place *vis a vis* the working models that were described in the previous chapter. In fact, one might see computational simulations as challenging the distinction between theoretical and working models. On the one hand, simulations appear to be something like automated theories. On the other hand, simulations seem to be the most abundant form of models that *do* something–working models.

I begin this chapter with a review of some crucial aspects of the work of John von Neumann and Alan Turing who were involved in the theorization of mathematical computation and the design of the first general-purpose stored-program computers. They also theorized the ways in which the computer could simulate the brain and mind. What this examination reveals is that the design of the first general-purpose computers drew heavily upon the McCulloch and Pitts (1943) neuron model and other aspects of neurophysiology. That is to say, the first computers were *modeled on* the brain, or more precisely were modeled on theoretical models of neurons. There is also a sense in which the mathematical theory of computation was modeled on the human practices of performing mathematical computations. These basic forms of modeling are not often discussed, yet I believe they have had a significant impact on our theoretical intuitions regarding the more complex forms of modeling of the computer after the brain. This was especially true early on, when computers were commonly referred to as "giant brains."

The computer is a unique artifact among the tools of modeling for several reasons that deserve special attention. Of course, the very notion of a symbolic simulation is tied up with ideas about the nature of computation and the technological performance of working computers. Moreover, these ideas, and the conceptualization and design of the first computers, are intimately related to models of the brain and mind in multiple and complex ways. Part of the purpose of this chapter is to make these multiple and complex relations visible as instances of the different types of scientific modeling and models. Thus in considering the computer and the brain we can find instances of each type of model we have considered in previous chapters, as well as the simulations that will be discussed shortly.

Before embarking on our historical journey to the electronic brains of the 1940s, it is worthwhile to consider the view of the mind that has dominated cognitive psychology since then.

The view that the mind *is* a computer has been dubbed "Computationalism." As the preface to a recent book on Computationalism summarizes the view:

> Are minds computers? . . . Computationalism–the view that mental states are computational states–is based on the conviction that there are program descriptions of mental processes and that, at least in principle, it is possible for computers, that is, machines of a particular kind, to possess mentality. (Scheutz 2002, p. ix).

Despite the many different formulations of it, Computationalism is rooted in the basic analogy between mental states and computational states–the different formulations are the result of different definitions and conceptions of these two types of states. In general, the definition of computational states is based upon the stored-program digital serial computer. Two of the mathematicians who had the greatest influence on the design of the first such computers, Turing and von Neumann, were also engaged in the project of building synthetic minds. As we will see in this chapter, while they shared a common understanding of computation, they differed in their views of how a computer might simulate a brain. Still different from these was W. Ross Ashby's attempts to model mental phenomena. Ashby saw "information processing" as central to the adaptive mind, but not strictly as a form of computation. As a result of this, he built analog simulations (working models) of the brain, whereas Turing and von Neumann were seeking different kinds of symbolic simulations (theoretical models).

We will begin this chapter with a clarification of the distinction between simulations as working models and simulations as automated theoretical models. The usefulness of distinguishing the various types of models and modeling will be demonstrated by examining the early history of the computer and attempts to simulate the brain with these early machines. In so doing, we will consider the automata theory developed by von Neumann as he was designing the EDVAC, the first stored-program computer. We will also consider his use of the McCulloch and Pitts neuron model in this design, and his later thoughts on simulating the brain on the computer. We will then consider Turing's preoccupation with the universal (Turing) machine, as a model of the mind and brain. This will lead us to his suggestion to Ashby to simulate the Homeostat on his ACE computer, and a direct comparison of these two types of simulations. I then conclude with the consequences of the developed view of models and simulations for Computationalism.

4.2 Analog and Symbolic Simulations

In the previous chapter we have seen that there is an advantage to making a careful distinction between theoretical models and working models. While the notion of theoretical models has its roots in the normative philosophy of science, both theoretical and working models are useful to a descriptive naturalistic study of scientific practice. Again, the crucial distinction to be made between them is that working models do something in the world, they have material agency, and this agency is independent of human agency. Theoretical models can also have agency, however this is a disciplinary agency and it is not completely independent of human agency, but requires a human agency willing to conform to its disciplinary rules and constraints. Now we will consider more carefully the distinction between theoretical and working models, and how they relate to the current philosophical discussion of simulations.

In a recent paper on computational simulations, Winsberg (2003) considers three traditional attempts to account for simulations: as metaphors, as experiments, and a third middle mode. The view of simulations as metaphors, while perhaps never expressed clearly as such, nonetheless holds that simulations are essentially just brute-force number crunching procedures, used when analytic techniques are impossible–in other words, a degenerate form of theorizing. The view that simulations are experiments, and the computer is an experimental target, holds that there is some mimetic relation between the simulation and the simulated, such that the simulation can mimic the real and act as a stand-in. The third mode holds that simulations are an entirely different kind of things, lying somewhere between theorizing and experiment.

Winsberg is careful to distinguish between simulations in which analytic solutions produce closed form expressions, and simulations which use numerical methods that produce a "big pile of numbers" that require the usual tools of experimental practice to analyze: visualization, statistics, data mining, etc. (Winsberg 2003, p. 111). The difference is not merely one of the mimetic qualities of the mathematics, but of the practices which scientists use to engage and work with the models. The numerical methods are more like experiments than theory because the same practices are used to study them as are used in experimental investigations.

In the terms we used in the previous chapter to describe working models, these numerical simulations are used to generate, or synthesize, phenomena which are to be investigated,

explored, experimented on, etc.–they are themselves objects of empirical study. This is in contrast to the theoretical models, or simulations based on tractable equations deduced from theory, which lend themselves to simple and straightforward mathematical analysis. In these analytic simulations one does not need to employ the data analysis techniques of experimental practice to discern the structures and patterns in the phenomena–these models only produce data as an instantiation of the theory for a given case, and can the desired results are easily derived from the equations. This distinction also corresponds nicely to the analytic/synthetic distinction drawn in the previous chapter. Here the analytic simulations engage in deriving local models from general theories in a formal way, whereas the synthetic simulations seek to fill in the gaps of missing theory and data by generating something new that can be manipulated, experimented on, and used to generate data to devise and test theories.

Much of the philosophical literature on simulations appears to Winsberg to be hung up on visualization as a key aspect of what makes simulations interesting. This is captured in the notion of the mimetic qualities in the simulation's representations. Like isomorphism, mimetic relations are meant to provide an objective way of expressing the precise relations between the real and the simulated systems, but mimetic relations have the added requirement of preserving the graphical aspects of the original:

> The extensive use of realistic images in simulation is a stepping stone that simulationists use in order to make inferences from their data. It is also a tool they use in order to draw comparisons between simulation results and real systems; a move that is part of the process of sanctioning of their results. It is the drawing of inferences and sanctioning of results that give rise to the interesting philosophical connections between simulation and experimental practice. (Winsberg 2003, p. 113).

Certainly the graphical aspects of such models can provide a means to the employment of visual laboratory practices, but it is not the only such aspect, and it is not essential. As Winsberg notes, it is possible to derive visualizations from many mathematical models, and this appears tangential to what makes them good models. Further, if one puts too much faith in the power of the mimetic features, it tends to lead one to see simulations as purely and truly experimental, taking literally the notion of the "numerical experiment" and interpreting the computer simulation as a stand-in for the real phenomena. This begs the question of just how well a simulation mimics the real phenomena (Winsberg 2003, p. 115). That is, it matters not just that

it stands-in in some respect, but in which respects, to what extent and to what precision and accuracy. Particularly if we are interested in the epistemic status of models and simulations, these questions are of great importance, and we cannot take for granted that the fact that we can perform experiments on simulations means they have the same epistemic, much less metaphysical, status as real experiments. Still, it does not mean they lack any epistemic status, or that they are metaphysically weak, either.

In considering that simulation might be a unique new mode of scientific practice between theorizing and experimenting, Winsberg asserts that this is merely a good place to start thinking about simulation, not an explanation of it:

> What is of interest philosophically is to understand (a) how it is that what is at root a theoretical enterprise, takes on characteristics of experimentation, (b) what those characteristics are–at the abstract, reconstructed level, (c) what consequences there are of such a hybrid for our understanding of the nature of modeling, theorizing, and experimenting, and (d) how simulation produces knowledge and what kind of knowledge that is. (Winsberg 2003, p. 118)

In staking out this middle ground for simulation, Winsberg is careful to note that the techniques and practices of modeling can carry their own epistemic credentials, independently of the theory from which they are produced:

> [T]he credibility of [a] model comes not only from the credentials supplied to it by the governing theory, but also from the antecedently established credentials of the model building techniques developed over an extended tradition of employment. That is what I mean when I say that these techniques have their own life; they carry with them their own history of prior successes and accomplishments, and, when properly used, they can bring to the table independent warrant for belief in the models that are used to build. (Winsberg 2003, p. 122).

Thus, it seems that for Winsberg the epistemic basis for simulations comes from both theory and the practices of modeling. It is important to note, however, that he also holds that those simulations are only autonomous (or semi-autonomous as he says) from theory to the extent that they have these independent epistemic foundations rooted in a tradition of practice. I agree with this approach, and to it just wish to add that working models are those models which are produced in, and support, the development of a tradition of modeling practices.

Before discussing the early attempts at using the computer to simulate the mind and brain, I want to consider one more position on simulation that considers more explicitly how a model

represents the system it models. During the Connectionist debates of the early 1990s, there was a recurring argumentative theme based on the distinction between analog and digital computation. Perhaps too much was made of this distinction, or rather, the real nature of the distinction was not always fully recognized. An exception to this is Trenholme (1994). Developing a line of thought connecting the ideas of Kenneth Craik, Norbert Wiener, Philip Johnson-Laird and Rodney Brooks, Trenholme presents a view of simulation that I believe is compatible with the idea of working models that I have been developing. It is my hope that applying his view to the work of von Neumann and Turing on the early computers will provide a new perspective on how the computer was variously conceived of as a simulation of the brain and mind.

In short, the idea is that while all synthetic simulations are working models in the generic sense of being automatic, just as all models are "representational" in a generic sense, the nature of the scientifically relevant aspect of their being models can differ in significant ways. An analog simulation directly models a natural system, while a symbolic simulation employs an intermediate symbolic system and is thus an indirect model. This level of indirection is highly significant to the extent that it shifts the relevant aspect of a simulation from the realm of working models, epistemic artifacts and material agency, to the realm of theoretical models and disciplinary agency. Thus, what Trenholme calls a naturalistic analog simulation is what I have been calling a working model, and what he calls a symbolic simulation is a special kind of theoretical model, one whose disciplinary agency has been transferred from human hands and minds to automatic computations.

Trenholme's argument begins with a clarification. While much of the debate at the time was couched in terms of "analog vs. digital" computation, Trenholme is careful to point out that the real issue is between "analog vs. symbolic" representation. As we will see in the discussion of von Neumann's automata theory below, "analog vs. digital" is a matter of how *numbers* are represented in a computer. It says nothing, however, about what or how those numbers come to represent or simulate anything else beyond those numbers. The real issue is whether numbers are an essential part of the simulation at all. Hence, the "analog vs. symbolic" dichotomy is meant to capture the notion that analog simulations do not require or depend upon symbolic representations. Rather, they serve as simulations, primarily or completely, in virtue of their

causal structure. Trenholme calls these "naturalistic analog simulations" in order to distinguish them from the more careless definitions of analog computation.

Trenholme goes on to argue that naturalistic analog simulations[10] are not representational in the same sense that symbolic simulations are representational. This is due to an additional "mapping" in the sense discussed in previous chapters. In short, analog simulations relate the causal structure of a natural phenomenon to the causal structure of the simulation by isomorphism–actually the looser "similarity" relation is appealed to (along with a probabilistic causal theory). A symbolic simulation, on the other hand, is derived, or mapped, from a formal theory of a natural phenomena, and then this is mapped into a computational simulation of the formal theory. Missing from the symbolic simulation is the obvious sense of an isomorphism between causal structures. Symbolic simulations are instead representational in the sense of semantic relations like denotation. Analog simulations lack this purely representational layer:

> Symbolic simulation is thus a two-stage affair: first the mapping of inference structure of the theory onto hardware states which defines symbolic computation; second, the mapping of inference structure of the theory onto hardware states which (under appropriate conditions) qualifies the processing as a symbolic simulation. Analog simulation, in contrast, is defined by a single mapping from causal relations among elements of the simulation to causal relations among elements of the simulated phenomenon. (Trenholme 1994, p.119).

While a symbolic simulation depends on causal structures at *some* level, it is the symbolic level that is crucial to its performing as a simulation. Similarly, while an analog simulation depends on representation at *some* level, it is the causal structure, or material agency, which is operative in its performing as a simulation. There is a generic sense of representation involved in both types of simulation, *i.e.*, that the simulation is in some sense a representation of the system being simulated, but the kind of representation involved is not necessarily symbolic representation.

For instance, the mercury barometer "represents" air pressure without employing symbolic systems *in its performance*–the symbols only come at the end, as it were, when the result is read off from the markings on the device. Rather than depend upon the symbolic relations to achieve a well behaved simulation, analog simulations depend upon causal structures. These causal structures are the material agency of working models. On the one hand these causal

[10]I will simply call these "analog simulations," but will be careful to distinguish them from analog representations, analog computations, as is necessary.

structures must be constrained in specific ways so as to constitute a valid simulation (they are disciplined material agencies), while on the other hand there are always ways in which they are unconstrained and open-ended to the extent that they are involved in the potentially infinite causal relations of the world and can always exhibit new emergent properties. There is thus an implicit respect for this dual aspect of material agency in analog simulations that is largely missing in symbolic representations that attempt to completely constrain the behavior of the physical system to produce only the precise symbolic system that is intended. When a symbolic simulation behaves irregularly, its output is essentially meaningless.

Symbolic representation is a very specific form of representation in which symbols are made to represent through denotation and reference. Symbols alone do not have a causal structure, and in order to make an automatic simulation based on them, it is necessary to have what Newell and Simon call a "physical symbol-system," *i.e.,* a computer:

> *The Physical Symbol System Hypothesis.* A physical symbol system has the necessary and sufficient means for general intelligent action. By 'necessary' we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By 'sufficient' we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence (Newell & Simon 1976, p. 41).

Simulations based upon symbolic representations are so closely tied to the development of the modern computer that they are often simply called "computer simulations." There are other expressions which evoke notions of modeling and simulation, but actually serve to confuse the issue:

> Certain ways of speaking–for example, saying that a theory is "modeled on a computer"–risk conflating the two mappings, thus blurring the distinction between analog and symbolic simulation, as does the notion of representation when applied to analog simulations. . . . Thus we may say that a user (or an observer) recognizes an analog simulation as such when properties in the analog device are held to *represent* the corresponding properties (those that play the corresponding causal role under the causal-structural isomorphism). In the case of symbolic simulation, once the user identifies the phenomenon to be simulated, the term of the theory coded into the computer may be held to *represent* the relevant features of the phenomenon under its standard (intended) interpretation; here the notion of *representation* stands for language-world semantical relations such as *reference* and *denotation*. An obvious distinction can be made between these semantical relations and the notion of isomorphism of causal structure used in characterizing naturalistic analog simulation. (Trenholme 1994, p.119).

Once the distinction between analog and symbolic representation is clear, one can begin to see how this argument relates to the larger questions concerning Computationalism.[11]

It is worth noting at this point that there are at least two significantly different ways to elaborate the isomorphism[12] between a simulation and the system or process it simulates. One is the simple input/output correspondence, or what Ashby first called the "black box" simulation. Under this view, one system is a black box simulation of another if it produces similar outputs under similar inputs. In the case of analog simulations, the inputs and outputs are causal, in symbolic simulations they are symbolic. The matter is more complicated in the second type of isomorphism in which the *internal processes* of the simulation and the system or process being simulated are meant to correspond. In this form of simulation not only do the input/output relations matter, but also the internal states and processes of the simulation. The difference between analog and symbolic simulations is made clearer in such cases.[13] While symbolic simulations must realize certain symbolic processes, there is an independence between the symbolic processes and their physical realization. This is generally referred to as "multiple realizability" meaning that different causal systems can realize the same symbolic process. For analog simulations, which lack an independent symbolic level, the causal processes that realize the simulation must be isomorphic to the system being simulated for it to count as a simulation. Let us now turn from this abstract discussion of simulation to some more concrete historical examples. We will see in the ideas of von Neumann and Turing two different kinds of symbolic simulation. For von Neumann, the computer ought to be used to simulate a *physical model* of the brain, while for Turing the computer ought to simulate the same essential algorithm, or program, that the brain itself simulates.

[11]For instance, the kind of argument made by Searle (1980) in his famous Chinese-room thought experiment turns on just these semantic aspects of symbols. In that argument, he aims to show that intentionality is an essential element of symbolic representation and that Computationalism, or at least the strong program in AI, fails to explain it.

[12]I use isomorphism here not in its strict sense of correspondence, but rather in the looser sense of similarity proposed by Teller (2001) and others and discussed earlier in Chapter 2.

[13]For black box simulations, we do not care about internal processes so it makes little difference whether these are achieved symbolically or causally, or even magically.

4.3 Analogies, Digits, and Numbers

The distinction between analog and digital is due to John von Neumann. His work on the first stored-program computers and views on the relationship between the computer and the brain present a complex history involving numerous layers of analogy, modeling and simulation. A complete accounting of the development of his thought is well beyond the scope of the present project, but it is worthwhile to consider how the view of working models and simulations would apply to his various stated positions. It is clear that his design of the first computer memory drew heavily upon his review of research in neurophysiology, and quite explicitly upon the work of McCulloch and Pitts (1943) and their conception of logical neural networks. This much is clear from his EDVAC design proposal (von Neumann 1945). The story becomes more complex when we consider his 1948 theory of automata (von Neumann 1951), his 1946 letter to Norbert Wiener about simulating the brain (von Neumann 1946), and his later reflections on the relationship between the brain and the computer for his posthumously published Silliman lectures (von Neumann 1958). In this section I will briefly review some of von Neumann's thoughts on the issues involved in thinking about the computer as being modeled after the brain, and his approach to simulating the brain with a computer.

The key to understanding the development of von Neumann's ideas on the relationship between the brain and the computer is to keep clear that the computer is a representational system of a specific type for von Neumann. In the case of his automata theory, the computer is fundamentally an automatic system for representing and manipulating *numbers*. Numbers are not the same thing as quantities or numerals, they are abstract mathematical entities. Quantities and numerals are concrete ways of representing numbers for practical purposes. Thus, to automate mathematics it is necessary to develop a physical system which can represent numbers, and the choice between quantities and numerals is an open question before the technological possibilities are considered.

On September 20, 1948, von Neumann presented a theory of automata at the Hixon Symposium on Cerebral Mechanisms in Behavior (von Neumann 1951). In his presentation, "The General and Logical Theory of Automata," he spells out the need he saw for a rigorous theory of computation, and outlines a formal theory of *automata* (axiomatically idealized

computational mechanisms). He began by distinguishing two general classes of automata by their mode of representing numbers. The class of automata built on the "analogy principle" represent numbers by analogy; that is, through certain physical quantities that they exhibit in the way that a thermometer represents temperature by the height of its mercury. If, for example, we are representing two numbers by the electrical currents in two circuits, we can add the numbers by combining the circuits appropriately and the result will be registered as the total current output from the combined circuit. It is possible to do all the basic arithmetic operations (+, -, *,÷) in roughly this way by using the currents in a circuit and providing the appropriate configurations of the circuit's relays and switches. Automata built on the "digital principle" do not represent numbers as physical quantities, but as aggregates of numerical digits in the manner humans typically do when we write them down on paper or count on our fingers (the etymological origin of "digit" itself). Such an automata might have a dial with ten positions on it representing 0-9, or a series of such dials for the ones column, tens column, hundreds column, etc, of the decimal numbering system. The digital representation used by nearly all modern computers is a binary system in which wires carry electrical currents of two sufficiently distinct magnitudes, and employ a set of canonical circuits to perform mathematical and logical operations on the binary representations.

The two types of automata described by von Neumann correspond to Trenholme's (1994) "analog vs. digital" distinction, though Trenholme criticizes it as being spurious in distinguishing between types of simulation. That is, in each case one is seeking to represent numbers and, unless one is seeking to "simulate" pure mathematics, both are forms of symbolic simulations and are caught up in a double-mapping. This is because even if we might be using an analog computer, such as a differential analyzer, we still need a theory about how the causal structure of the analog computer realizes the desired mathematical calculations, as well as a theory about how those calculations relate to the natural phenomena in question. This is just to say that a differential analyzer is not a simulation, *e.g.,* of some hydrodynamic system, but an automated device for solving a set of equations derived from a theory of that hydrodynamic system.

The main consequence of the different kinds of numerical representation in automata are the practical realization of them when building electronic computers. Analog computers are

deeply susceptible to errors by the constant introduction of "noise" into their circuits, though they have the advantage of arbitrarily high precision. That is, when dividing, as when dividing 5 by 7, the result cannot be expressed in a finite number of digits and so a digital computer's answer will be limited by the number of digits it can represent at one time, while an analogy computer can represent the answer to an unlimited precision.[14] The down side is that this is only an ideal, in actuality it was quite difficult to get analog circuits to perform calculations with much accuracy because slight fluctuations in the physical quantities, due to sources outside the computation, build up as noise in the system. Consequently, while our analog division of 5 by 7 might be perfectly precise in theory, even low probabilities of very small disturbances in the current of the circuit will limit the actual precision enormously. Von Neumann insists that it is the difficulty of the engineering problem of maintaining a reliable signal-to-noise ratio that prevents the analog computer from being made more precise than a few decimal places, and it just so happens that this problem does not arise in digital machines–which have the further advantage that precision can be increased indefinitely and economically by simply increasing the number of digits represented by duplicating components.

While this point is often overlooked, von Neumann himself claims there is no essential mathematical difference between the two kinds of automata, though there are great practical differences. Primarily, any kind of calculation that can, in principle, be done on an analog automata can also be done on a digital automata and vice versa. Von Neumann argues that organisms, natural automata, are actually *mixed* automata utilizing both principles for different particular functions, and that high speed computers are strictly digital machines. He is careful to point out that trying to base a theory of mind on the kinds of computations achievable in digital computers should not affect its applicability to the living brain since neuronal activity had been shown to be digital and, even if it were not, the theory could be realized by an appropriately constructed analog machine anyway. In other words, a computational simulation of the brain

[14]Computer programmers call this a "truncation" error–when the number of digits needed for an answer exceeds the number the machine can represent. In division, this generally results in a truncation or rounding-off of the decimal positions after the last available digit is used. In multiplication the result is more disastrous. Consider a machine multiplying two 10-digit numbers, and capable of only representing 10 digits at a time. The result of the multiplication will be 18 digits long, but the machine can only represent half of these, and hence can not represent a meaningful answer at all–or must convert it to scientific notation, truncate it and thereby lose precision.

will be a symbolic simulation, as long as that simulation depends on the computation of theoretical physical models, whether an analog or digital computer is used.

4.4 Modeling the Computer on the Brain

> *These repeated excursions into biological information processing and the interdisciplinary study of cybernetics have been ignored in previous accounts of von Neumann's computing, yet they clearly shaped his ideas.* (Aspray 1990, p. 189).

This section will be a digression from our discussion of simulations. It is relevant, however, to the extent that it considers the practice of "modeling" in the development of the computer. In particular, it tells the story of how the computer was modeled on the brain, or at least upon certain specific theories of how the brain and its neurons work. There was something more than numerical representation required for the stored-program computer to become a technological reality. Namely, the stored-program computer would need to represent its own *instructions* as numbers, which it could then store in memory, retrieve, execute, alter, etc. In other words, certain numbers are *representations to the computer* of instructions it is meant to interpret and perform. This was the great leap in design that turned mere calculators into universal computers. It is also the kernel of a deep confusion regarding the representational character of computations. This kernel was fertilized early on by the connections drawn between computational memory and neural structures in the brain. What is clear from von Neumann's description of the EDVAC design is that he looked to the brain and to the McCulloch and Pitts neuron for inspiration in the design of the first computer memory. This led to a peculiar kind of analogy between the computer and the mind/brain which the proponents of Computationalism would soon embrace, but which von Neumann himself would ultimately reject.

While various electromechanical calculators had been constructed before, during, and after World War II, the principle technological impediment to creating a general purpose computer (*i.e.,* a working Universal Turing Machine) was with how to give it a memory.[15] The

[15] The early computing machines were either highly specialized to perform a single class of functions or, like ENIAC, had to be programmed manually by arranging walls of dials and networks of patch cables in a fashion similar to hand-operated switchboards for telephones. For instance, Howard Aiken's Mark I Automatic Controlled Sequence Calculator computer, built for IBM in 1944, used 72 rotary counters for storing *numbers* mechanically, but had no means for storing the calculation *instructions* themselves–the "function" resided in the mechanical configuration of the device.

problem was one of both organization and technological realization. It is generally held that the first machine to solve the memory problem was von Neumann's Institute for Advanced Study machine. Von Neumann solved the problem by introducing a memory unit, or "organ" as he called it, as a central part of the computer architecture. The memory organ was distinguished from numerical counters by its representing, or storing, logical instructions, or "codes," as well as numbers. It thus became possible for the computer to calculate partial solutions, store the intermediate results, reconfigure itself to perform a new function by following the instructions in memory, and resume the calculation with the instructions stored in memory as numbers (Goldstine & von Neumann 1946). This had obvious advantages over having to reconfigure the machine by hand for each portion of a calculation, or having to build a computer complicated enough to be completely preprogrammed for all the steps necessary in a computationally demanding problem.

An often overlooked fact about this early computer design is the extent to which von Neumann drew upon the McCulloch and Pitts neuron model and the neurological language he used to describe the new design. This is a clear case of modeling on–the transfer of a theory in one domain to the design of a technology in another domain. The fact that von Neumann modeled his design after neurophysiological theory and the McCulloch and Pitts neuron is unmistakable in the language he uses to describe the computer:

> The three specific parts CA [Central Arithmetical organ], CC [Central Control organ], and M [Memory organ] correspond to the *associative* neurons in the human nervous system. It remains to discuss the equivalents of the *sensory* or *afferent* and the *motor* or *efferent* neurons. These are the *input* and the *output* organs of the device, and we shall now consider them briefly. (von Neumann 1945, p. 20, *original emphasis*)

The most obvious thing to note about this passage is that he does not refer to the parts of the computer's architecture as "units," "components," "modules," or any other such connotatively neutral engineering terms. As a mathematician, he chose the semantically loaded term "organs" taken from the description of biological systems. Moreover, these are explicitly made out as *corresponding to specific elements in the human brain*. This theme is carried through his description of the machine:

> It is worth mentioning, that the neurons of the higher animals are definitely elements in the above sense. They have all-or-none character, that is two states: Quiescent and

> excited. They fulfill the requirements of 4.1 with an interesting variant: An excited neuron emits the standard stimulus along many lines (axons). Such a line can, however, be connected in two different ways to the next neuron: First: In an *excitatory synapsis*, so that the stimulus causes the excitation of that neuron. Second: In an *inhibitory synapsis*, so that the stimulus absolutely prevents the excitation of the neuron by any other (excitatory) synapsis. The neuron also has a definite reaction time, between the reception of a stimulus and the emission of the stimuli caused by it, the *synaptic delay*. Following W. Pitts and W. S. MacCulloch [*sic*] ("A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophysics*, vol. 5 [1943], pp. 115-133.) We ignore the more complicated aspects of neuron functioning: Thresholds, temporal summation, relative inhibition, changes of the threshold by after effects of stimulation beyond synaptic delay, etc. It is, however, convenient to consider occasionally neurons with fixed thresholds 2 and 3, that is neurons which can be excited only by (simultaneous) stimuli on 2 or 3 excitatory synapses (and none on an inhibitory synapsis). It is easily seen, that these simplified neuron functions can be imitated by telegraph relays or by vacuum tubes. Although the nervous system is presumably asynchronous (for the synaptic delays), precise synaptic delays can be obtained by using synchronous setups. (von Neumann 1945, p. 24)

In this passage, it is clear that von Neumann is making a serious consideration of the structure and function of biological neurons, albeit idealized in the manner of McCulloch and Pitts' neuron model. Here we see how the functional identity of neuron activity and mathematical logic (a unification of neural mechanisms and Turing machines) is being used as a modeling tool in the design of the electronic computer. Its usefulness as a model is metaphorical or analogical, to be sure, but this kind of model is powerful in that it actually influences design *decisions*, and not merely the construction of design *alternatives*. In that sense it is evaluative or normative:

> The analogs of human neurons, discussed in 4.2-4.3 and again referred to at the end of 5.1, seem to provide elements of just the kind postulated at the end of 6.1. We propose to use them accordingly for the purpose described there: as the constituent elements of the device, for the duration of the preliminary discussion. We must therefore give a precise account of the properties which we postulate for these elements. (von Neumann 1945, p. 30).

But this model was not the only criteria operative in design decisions:

> At this point the following observation is necessary. In the human nervous system the conduction times along the lines (axons) can be longer than the synaptic delays, hence our above procedure of neglecting them aside of t would be unsound. In the actually intended vacuum tube interpretation, however, this procedure is justified: t is to be about a microsecond, an electro-magnetic impulse travels in this time 300 meters, and as the lines

> are likely to be short compared to this, the conduction times may indeed be neglected. (von Neumann 1945, p. 30)

It seems clear enough from these passages that von Neumann's conception of physical computation was almost completely circumscribed by the McCulloch and Pitts neuron model.

What conclusions can be drawn from these observations of the significance of the McCulloch and Pitts' neuron model on the design of the first computers? The first point is that the computer was from its very conception a kind of model of the brain. Thus, it is at least mistaken to think that researchers in Artificial Intelligence or Cognitive Science "discovered" any analogy between the mind and computer. It was always there. Rather, they reconfiguring this analogy in an effort to develop computer programs as psychological theories. One way of approaching the question of Computationalism might be to ask whether the computer itself, regardless of the program or simulation it is performing, is a model of the mind, or a good model of the mind. It obviously could be, and has been, such a model. This leaves open the question of whether the mind actually is a computer, or a computer a mind, however. But almost no one holds the view that every computer is a mind. Turing seems to have held the view that minds are Universal Computers, and perhaps also the converse, that Universal Computers are at least capable of being minds. We will consider his views shortly. By and large it is really only necessary to consider whether there might be some way for the computer to simulate the brain so efficiently that it becomes a mind.

4.5 Simulating the Brain on the Computer

It is worth considering at this point how von Neumann envisioned the computer as a simulation of the brain. Even while von Neumann had based his automata theory on a strict distinction between analog and digital numerical representation, and modeled the memory of the stored-program computer after McCulloch and Pitts' essentially digital model of neurons, he was clear in other writings that the brain itself was actually much more complicated than these theoretical idealizations let on. As a result, its simulation by a computer would be far more elusive.

On November 29th, 1946, just after the second Macy Conference, von Neumann wrote a letter to the mathematician and cybernetician Norbert Wiener in which he assessed the situation regarding the theory of biological information processing, and the replication of the brain's abilities in a computer. The letter is interesting for a number of reasons, and warrants more thorough examination than our current purposes permit. Despite his early enthusiasm, von Neumann was perhaps the first to realize the limitations of computers for simulating the brain. The clearest articulation of the shortcomings of this approach comes in his letter to Wiener. Here von Neumann argues that directly modeling the physical structure of the brain will be so complicated as to be nearly hopeless, an idea he would more clearly express in later work. He concludes from this situation that a much better approach would be to turn instead to detailed cytological work. Specifically, he proposes to understand simple organisms in complete atomic detail (literally), by starting with the study of bacteriophagic viruses.

In the letter von Neumann intimates one of the primary reasons for the difficulty in making the analogy between brains and computers:

> Besides, the [brain] system is not even digital (*i.e.* neural): It is intimately connected to a very complex analogy; (*i.e.* humoral or hormonal) system, and almost every feedback loop goes through both sectors. If not through the "outside" world (*i.e.* the epidermis or within the digestive system) as well. (von Neumann 1946, p. 507).

There are two important points to note in this passage. First, von Neumann is keenly aware of the embedded and situated nature of the human brain, and that its information processing relies essentially on feedback loops with its environment. Unfortunately, von Neumann does not direct his energies in pursuing this issue. The other point is that the supposedly neat distinction between analog and digital is not so clear in the living brain and this is the direction in which von Neumann devotes a great deal of his energy.

One significant consequence of the McCulloch and Pitts' neuron model was to establish the digital character of neuronal behavior in the brain as the character relevant for understanding its organization. While this move had great benefits in terms of formalizing neural networks, and thereby treating them using mathematical logic, the idealization was ultimately a gross simplification. Von Neumann states the case most devastatingly in his Silliman manuscript. In

one passage, he assails the notion that it is safe to treat the mechanism of neuronal excitation and inhibition as a straightforward summation function:

> It may well be that certain nerve pulse combinations will stimulate a given neuron not simply by virtue of their number, but also by virtue of the spatial relations of the synapses on a single nerve cell, and the combinations of stimulations on these that are effective (that generate a response pulse in the last-mentioned neuron) are characterized not only by their number but also by their coverage of certain special regions on that neuron (on its body or its dendrite system, cf. above), by the spatial relations of such regions to each other, and by even more complicated quantitative and geometrical relationships that might be relevant. (von Neumann 1958, pp. 54-55).

While it might be tempting to treat things as if all inputs to a neuron were equal, in fact the complex three-dimension geometry of the synaptic connections to the neuron's dendrites are relevant to the electro-chemical processes which trigger a pulse. Similarly, the ideal of the synchronous timing of neuronal activity, essential to McCulloch and Pitts assertion that multi-layered networks can be treated as equivalent to logical propositions, is also untrue:

> On all these matters certain (more or less incomplete) bodies of observation exist, and they all indicate that the individual neuron may be–at least in suitable special situations–a much more complicated mechanism than the dogmatic description in terms of stimulus-response, following the simple patterns of elementary logical operations, can express. (von Neumann 1958, p. 56).

So while he found the suggestions of treating neurons as logical units performing summations useful for designing computer memory circuits, he was deeply disturbed by just how remote this idealization was from real brains when it came to building a simulation.

Not only were the neurons not clearly performing the idealized functions required by McCulloch and Pitts, even von Neumann's own distinctions between analog and digital automata applied to the brain in complex, and by no means straightforward, ways:

> The observation I wish to make is this: processes which go through the nervous system may, as I pointed out before, change their character from digital to analog, and back to digital, etc., repeatedly. Nerve pulses, *i.e.* the digital part of the mechanism, may control a particular stage of such a process, *e.g.* the contraction of a specific muscle or the secretion of a specific chemical. This phenomenon is one belonging to the analog class, but it may be the origin of a train of nerve pulses which are due to its being sensed by suitable inner receptors. When such nerve pulses are being generated, we are back in the digital line of progression again. As mentioned above, such changes from a digital process to an analog one, and back again to a digital one, may alternate several times. Thus the nerve-pulse part of the system, which is digital, and the one involving chemical

> changes or mechanical distortions due to muscular contractions, which is of the analog type, may, by alternating with each other, give any particular process a mixed character. (von Neumann 1958, p. 68-69).

Ultimately, these complexities led von Neumann to believe that the study of the brain would lead to a new mathematics.

Von Neumann begins the letter to Wiener with a devastating critique of his own work, as well as that of Wiener, Turing, McCulloch and Pitts, to formulate a substantive theory of information processing in the brain:

> Our thoughts–I mean yours and Pitts' and mine–were so mainly focused on the subject of neurology, and more specifically on the human nervous system and there primarily on the central nervous system. Thus, in trying to understand the function of automata and the general principles governing them, we selected for prompt action the most complicated object under the sun–literally. . . . The difficulties are almost too obvious to mention: They reside in the exceptional complexity of the human nervous system, and indeed of any nervous system. (von Neumann 1946, pp. 506-7).

From this passage it is clear that von Neumann recognized that the lack of formalized scientific understanding of the operation of the neurons was a major hurdle to constructing synthetic brains with artificial automata. The consequence of his efforts in clarifying these ideas over the course of several years was one of exasperation:

> What seems worth emphasizing to me is, however, that after the great positive contribution of Turing-cum-Pitts-and-McCulloch is assimilated, the situation is rather worse than better than before. Indeed, these authors have demonstrated in absolute and hopeless generality, that anything and everything Brouwerian can be done by an appropriate mechanism and specifically by a neural mechanism–and that even one, definite mechanism can be "universal." (von Neumann 1946, p. 507).

This is perhaps the most devastating critique of Turing's project offered before the rise of AI. It is also the clearest statement of the paradoxical relationship between the universal computer and the synthetic brain, which we will consider in the next section. The critique is this: the universal computer, and its equation with the functioning of the brain is at the same time infinitely potent, yet essentially flaccid. The "absolute and hopeless generality" of the theory is that while it could replicate the processes of mind, if those are formally definable (Brouwerian), it gives absolutely no insight into the nature, structure or organization of those processes. The universal computer can imitate any machine, and also the brain-machine, but this tells us nothing about brains. What is needed instead is a rigorous theory of the biological brain:

> Inverting the argument: Nothing that we may know or learn about the functioning of the organism can give, without "microscopic," cytological work any clues regarding the further details of the neural mechanism. (von Neumann 1946, p. 507).

Ultimately, the study of neural networks in the abstract leaves us with the fundamental problem of understanding the brain empirically. And so the simulation of the brain on the computer is limited by our theoretical understanding of neuroscience. This should not be so surprising when we recognize von Neumann's approach to brain simulation as a form of symbolic simulation, and hence a theoretical simulation. That is to say, his notion of a computer simulation of the brain is completely dependent on a theoretical model of the brain being mapped into a symbolic representation on a computer. It is thus a clear example of the alignment between the symbolic simulation and the theoretical model. We will now turn to a different conception of how the computer can simulate the brain, which is nonetheless a symbolic simulation based on a theoretical model.

4.6 Computer as Universal Modeling Machine

> *It is possible to invent a single machine which can be used to compute any computable sequence.* (Turing 1936, p.127).

Now that we have examined the complex relationship between the computer and the brain in von Neumann's work, we turn to the more straightforward, if more abstract, relationship between the computer and the mind in Alan Turing's work. Put simply, Turing was preoccupied with his idea of the Universal Machine (now called Universal Turing Machines or UTMs). The idea of the UTM, as described in Turing (1936), is that of an abstract formal computer which is able to compute any computable function. Elsewhere he refers to the ability of the UTM to "model" any other machine. This way of talking is of course circumscribed by our earlier distinction between analog and symbolic simulation. The UTM is a purely symbolic conception, and real working computers are only approximations of this mathematical formalism. Turing did, however, seek to build real computers based on this formalism, most notably the Advanced Computational Engine (ACE).

In nearly every speech and paper related to computation that Turing produced from 1944-1950, he cited the similarity between his UTM and a programmable digital computer as being highly significant:

> Some years ago I was researching on what might now be described as an investigation of the theoretical possibilities and limitations of digital computing machines. I considered a type of machine which had a central mechanism, and an infinite memory which was contained on an infinite tape. This type of machine appeared to be sufficiently general. One of my conclusions was that the idea of a 'rule of thumb' process and a 'machine process' were synonymous. The expression 'machine process' of course means one which could be carried out by the type of machine I was considering. It was essential in these theoretical arguments that the memory should be infinite. It can easily be shown that otherwise the machine can only execute periodic operations. Machines such as the ACE may be regarded as practical versions of this same type of machine. There is at least a very close analogy. (Turing 1947, pp. 106-7)

It is clear from this passage and others that Turing sees universality as the critical element of UTMs–it is their ability to model any other machine that is so remarkable. The design of computing machines was then to be modeled after the mathematical theory of UTMs--the only difference in Turing's mind between the mathematical formalism of UTMs and the physical computer was the finite size of the memory, and the additional physical limitations of the machine, such as time. Apart from these practical constraints, Turing's vision of the computer might easily be called a "universal modeling machine."

His focus on universal modeling was also central to his consideration of learning and intelligence, and modeling these on a computer. He cast the problem as being that of an unorganized system becoming organized. But rather than taking a thermodynamic interpretation of organization, as the Cyberneticians would, Turing's approach sought to show how unorganized systems could efficiently organize themselves into UTMs through reinforcement by pleasure and pain.

Turing applied the same conception of universal modeling machines to the human brain as he did to the design of computers:

> All of this suggests that the cortex of the infant is an unorganized machine, which can be organized by suitable interfering training. The organizing might result in the modification of the machine into a universal machine or something like it. This would mean that the adult will obey orders given in appropriate language, even if they were very complicated; he would have no common sense, and would obey the most ridiculous

> orders unflinchingly. When all his orders had been fulfilled he would sink into a comatose state or perhaps obey some standing order, such as eating. Creatures not unlike this can be found, but most people behave quite differently under many circumstance. However, the resemblance to the universal machine is still very great, and suggests to us that the step from the unorganized infant to a universal machine is one which should be understood. When this has been mastered we shall be in a much better position to consider how the organizing process might have been modified to produce a more normal type of mind. (Turing 1948, p. 16/120).

For Turing, it seems that the key to unlocking the secrets of the mind lay with the UTM. For him, the mind had the ability to model other machines, and was thereby a universal model of a particular sort, though perhaps not completely identical to the uncreative computer.

This approach to intelligence and mind can be safely contrasted to that of W. Ross Ashby, who saw behavior not in terms of logical rule-following, but as trajectories in phase space. He sought to embody his ideas in machines which he could directly interact with, such as the Homeostat, and later the DAMS. Learning for Ashby was a never-ending dynamic process in which the goal was survival, and the environment was continually changing. Modeling was a consequence of this, not necessarily the cause. Turing saw learning as a search for a single stable goal–the organized universal computer. For Turing, it was the patterns of symbols in memory, the symbolic "instruction tables" which held the secrets of the mind, while for Ashby, it was the patterns of interactions and feedback loops between the system and its environment. In both cases, of course, it was possible to build a machine to embody their vision. However, for Turing this machine would epitomize the symbolic simulation of mind, while for Ashby the machine would epitomize the analog simulation of mental behavior. This contrast can be best understood by considering their letters to one another.

4.7 Simulating the Homeostat

When Ashby returned from his service with the Royal Medical Corps in India in the Spring of 1946, he had been concerned with the mechanisms of learning for over a decade and had begun thinking about constructing a machine to demonstrate the principles he had concluded to be essential to adaptation. During that year his notebooks are filled with mathematical formalisms for various stability-seeking systems. These eventually turn to diagrams of simple,

and then more complex, electrical circuits to realize the behavior of a set of equations in a system which could be directly engaged with.

In the fall of 1946, Ashby wrote to Turing about the adaptive machine Ashby was designing.[16] As with all other machines and processes, Turing believed that a programmable computer such as the ACE could model Ashby's Homeostat, as he explained in his reply to Ashby on November 20, 1946:

> The ACE will be used, as you suggest, in the first instance in an entirely disciplined manner, similar to the action of the lower centres, although the reflexes will be extremely complicated. The disciplined action carries with it the disagreeable feature, which you mentioned, that it will be entirely uncritical when anything goes wrong. It will also be necessarily devoid of anything that could be called originality. There is, however, no reason why the machine should always be used in such a manner: there is nothing in its construction which obliges us to do so. It would be quite possible for the machine to try out variations of behaviour and accept or reject them in the manner you describe and I have been hoping to make the machine do this. This is possible because, without altering the design of the machine itself, it can, in theory at any rate, be used as a model of any other machine, by making it remember a suitable set of instructions. The ACE is in fact, analogous to the 'universal machine' described in my paper on computable numbers. This theoretical possibility is attainable in practice, in all reasonable cases, at worst at the expense of operating slightly slower than a machine specially designed for the purpose in question. Thus, although the brain may in fact operate by changing its neuron circuits by the growth of axons and dendrites, we could nevertheless make a model, within the ACE, in which this possibility was allowed for, but in which the actual construction of the ACE did not alter, but only the remembered data, describing the mode of behaviour applicable at any time. I feel that you would be well advised to take advantage of this principle, and do your experiments on the ACE, instead of building a special machine. (Turing 1946).

This letter did not stop Ashby from pursuing his own machine and, in fact, it was shortly after he received this letter that he came up with one of the basic elements of the Homeostat circuit. The difference of perspective and approach between Ashby and Turing can now be made more clear. Of particular interest is his insistence that the universal character of the ACE means that there is no point in building a special machine of the type Ashby was proposing, even while he admits the importance of learning to intelligent behavior in general. There is also a tension between the rigid determined behavior of the ACE and the adaptive behavior of the brain. Considering how the ACE could model the Homeostat will bring into focus the differences between analog and symbolic computation.

[16] Unfortunately, this letter has been lost.

While Turing's letter seems to move casually from talking about the ACE modeling the brain to talking about the ACE modeling the Homeostat, we should pause to consider the difference. This difference is the essence of the analog and symbolic distinction insofar as Turing does not believe that the unprogrammed computer is a model of the brain. This is because it is too rigid and disciplined to be brain-like, though it has some resemblance to the "lower centres." What is required are the instruction tables, the program that defines the machine of the brain. And while Turing has some idea that this might be based on conditioned behavior, he has little idea how to program his machine to behave in this way. That is to say, the ACE might be able to perform a symbolic simulation of the brain, but only if it is given the proper symbolic organization. Ashby's Homeostat is itself a model of the brain, however. It is such a model because of its causal structure, not because of its symbolic structure. What Turing sees in the Homeostat, however, is the potential to model the brain indirectly. Because the Homeostat is a machine whose design is understood, it should be possible to simulate it symbolically on the ACE. Because the Homeostat is a simulated brain, the ACE's simulation of the Homeostat would be a simulation of the brain (once removed). The level of indirection in this case is the Homeostat itself as a theory of the brain's adaptive behavior. Any simulation performed by the ACE will only be as good as the theory upon which it is based. It tells us something about the Homeostat, and only indirectly tells us something about the brain. But we may still wonder what advantages there are to the directness we get from the analog simulation provided by the Homeostat. There are various reasons why the analog Homeostat is a better model of the brain than the symbolic Homeostat simulated by the ACE. First there are the practical issues which Ashby and Turing faced. Then there are methodological issues of scientific practice. And finally there is the issue of the effectiveness of the two simulations as a demonstrations. It is to these questions that I now turn.

First is the question of the practicality of building each simulation. As it turned out, the Homeostat was far easier to build than the ACE. Pilot ACE would not become fully operational until May of 1950, two years *after* the Homeostat was demonstrated to the EEG Society in Bristol (von Neumann's machine was not fully operational until 1952, though this was due more to its parts being continually redesigned). In addition, the Homeostat was a far more economical

solution than the ACE, even if it had only one application. In this sense Ashby was an intellectual entrepreneur who built a significant machine on his kitchen table out of war surplus parts, in a self-financed project. While this was true then, the opposite is true now, when computers sit on the majority of desks and finding the analog circuit components that Ashby used would be extremely difficult. Therein lies the true advantage of the symbolic computational simulation–because the computer potentially suits everyone's modeling needs, they can be produced cheaply in quantity and thus despite their vastly more complex structure turn out to be cheaper and easier than building specialized models. This much was clear to Turing by 1946. Less clear is the quality of the symbolic simulation when compared to the analog.

In Winsberg's (2003) analysis of simulations, he argues that it is often the case that the mathematical theory underlying the simulation is not in question, only the local outcomes of a specific situation. One of the great advantages of simulations for these cases is the ability to visualize the model, what Hughes (1997) called their mimetic properties. This advantage is due to the ability of researchers to employ experimental practices (most notably observation) and laboratory techniques in understanding the local simulations, as opposed to depending upon mathematical practices and analytic techniques. Yet these mimetic properties are dependent upon the development of instrumentations within the simulation to make those properties visible. The symbolic processes themselves are rather opaque and not susceptible to direct observation. This entails more levels of indirection in the interpretation and visualization of data derived from simulation. From an epistemic perspective, one is getting further and further away from the phenomena in question. Consider the Homeostat and its ACE simulation. The Homeostat can be directly interacted with and observed, any interactions with the ACE simulation must be preprogrammed, and the results must be interpreted from the symbolic output of the machine. While it is true that we could write a simple Java program with a colorful on-screen visualization of the Homeostat in 2005, this was hardly conceivable in 1946–which just proves the point that any such symbolic visualization introduces another layer of interpretation.

There are other important practical issues of time which are similarly becoming obscured by the increasing sophistication of computing machines. First, there are issues of mathematical complexity and computation time. To simulate the behavior of the Homeostat's circuits using

the equations of electrical engineering, or physics, would probably have taken far too long for a machine like ACE to perform in real-time (something approximating the response time of a real Homeostat). The significance of having a real-time simulation is related to the mimetic property mentioned above. Namely, that one can interact with a real-time system in ways that must themselves be simulated in false-time systems. This is not always what is needed in science, but is extremely important for understanding certain phenomena. Real-time here applies to the observers' time frame and ability to act upon the system, receive feedback from that action, and evaluate the relationship of the two. This feedback loop must be real-time for the observer, even if the time-frame of the simulated phenomena is greatly sped-up (as in astronomical phenomena like the movement of the planets) or slowed-down (as in high-speed phenomena like protein-folding). The significance of this aspect of analog models will become apparent in the next chapter when we consider the idea of a regulator as model.

4.8 Conclusions: Simulation and Computationalism

Von Neumann saw the power of mathematics to simulate detailed physical models, and ultimately to allow for the control of enormously complex systems involving many dynamically interdependent variables. His computer was built to automate the mathematic calculations needed to simulate these systems so as to allow mathematicians to develop ever more complex and detailed simulations. The brain, too, was a natural phenomenon which could be so simulated, once it was scientifically understood. The difference between Turing and von Neumann was that von Neumann felt that Turing's universal machine concept added almost nothing to our understanding of the mind because it told us nothing about the brain. Without a clear understanding of the brain's operation, the computer would never embody an artificial mind.

Von Neumann ultimately came to see the task at hand to be one of modeling the *detailed microstructures of brain*. For him, the computer was not itself a brain, though his own design for the EDVAC computer drew heavily on analogies to brain structures and neuronal functioning. The crucial aspects of the design of the computer were technical considerations of scalability, reliability, and numerical accuracy as computers were given more memory and faster processing

speeds. This was because he quickly realized that the scale of the simulations that would be needed for brains would require vast numbers of calculations, and even small or infrequent errors would result in critical failures. He also quickly became skeptical about rapid progress in building simulations of the brain.

Turing's view of the computer was as the universal modeling machine. But rather than worry about modeling the brain in physical detail, he sought to simulate its behavior symbolically. Thus both Turing and von Neumann sought after symbolic simulations of the brain, but had very different approaches to these. Ashby remained committed to analog simulations through most of the 1950s, but seems to have given up on them after his frustration with DAMS and the increasing capabilities of digital computers better suited his needs.

Another way of thinking about the difference between analog and symbolic simulations is to note that the symbols of a simulation inside a computer are not connected to the world in the same way as the causal structures of an analog simulation are. For the Homeostat, these connections were highly relevant to Ashby's theoretical views of the mind. Turing's claim that the distinctions are irrelevant thus bespeak a significant theoretical disparity between the two men. The best and simplest way of characterizing this difference between Ashby and Turing is in terms of the relationship between a system and its environment, which for Turing is irrelevant, yet for Ashby is central. It is a distinction of both theoretical and practical import. Theoretically, Ashby starts from the rich complexity of relationships between system and environment, while Turing considers the environment only in the simplest and most idealized terms possible. Practically, Ashby was interested in building a machine that an observer could directly interact with in real-time. Such a direct interaction was essential both to his pedagogical and rhetorical aims, and to his experimentalist desires. For Turing, logical proof was the ultimate form of demonstration, experimentalism was not a priority, and the details of the physical machinery were inconsequential. These distinctions became even more pronounced over time. For even as Turing turned to deeper considerations of learning and adaptation, he became more frank about his reasons for avoiding issues of general intelligence, human behavior, and human-like bodies for computers. In the next chapter we will consider the significance of situated and embodied

cognition, dynamic systems models, and regulators as models, for thinking about models of the brain and mind.

Chapter 5
Regulators as Models of Mind:
A Performative Model of How the Brain Represents the World

A dynamical account of cognition promises to minimize the difficulties in understanding how cognitive systems are real biological systems in constant, intimate dependence on, or interaction with, their surrounds.

-Timothy van Gelder

If the organism carries a "small-scale" model of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilise the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to emergencies which face it.

-Kenneth Craik

5.1 Introduction

In this chapter we shall consider a particular model of the mind that might yet prove to be as useful to cognitive science as the computer model has been the feedback-controlled regulator. The potential significance of the feedback-controlled regulator was recognized early on by the cyberneticians. It figured prominently in such important books as Wiener's *Cybernetics* (1948) and *The Human Use of Human Beings* (1950), and was the central connecting theme of the Macy Conferences (von Foerster 1949-52). In this chapter, I want to explore some of the ways that the idea of the feedback-controlled regulator might still be useful as a model of mind, in particular as a working model. I choose this model because it leads us to thinking about the mind as performative, embodied, and situated in an environment. As such, it offers a concrete example of how a causal mechanism can "model," or represent, the world without being an interpreted physical-symbol system. I hope to show how this cuts across some of the current debates in the philosophy of mind, in particular causal theories of mental content.

This chapter draws upon some of the later work of W. Ross Ashby, and in particular his work with Roger Conant on how a good regulator models the aspect of the world that it regulates. This work goes beyond the traditional cybernetic view of the error-controlled regulator, and considers the "cause-controlled" regulator as a natural extension of those ideas. The distinction is significant because it allows us to think about cognitive models as being causal models without resorting to computationalism, or the physical-symbol system hypothesis. This

allows us to develop a notion of the performative representation, with a different sort of semantics than symbolic representations. The value of doing this is that it offers a straightforward way of thinking about the brain as connected to and constantly interacting with the world, and of the mind having models of that world whose semantic content is tied up in the combined agencies of the mind and the world without independent symbolic structures.

The regulator has been considered at length by proponents of the Dynamicist approach as well. Timothy van Gelder (1995,1998) has focused attention on the steam engine governor–a type of mechanical regulator. He argues that the governor can be treated as a paradigmatic dynamic system as we seek to model the brain using dynamic systems theory. He further argues that dynamic systems theory challenges the basic assumptions of computational theories of mind, and even that it offers an alternative to traditional symbolic mental representation. This proposal has attracted some debate, which I shall examine.

The feedback controlled regulator, usually in the guise of the servomechanism, is a sort of ideograph for cybernetics. It embodies two of its deepest principles–mechanistic materialism and information feedback. That is, it was a material manifestation of information feedback in action–a performative, working model. In fact, the words "governor" and "cybernetics" share a common etymology, tracing back to the Greek *kubernetes* which referred to the person who steered a ship. This leads us to another way of distinguishing the performative and computational approaches to models: by their different conceptions of information. That is, there is a way of thinking about information as *coded data,* and another way of looking at information as a *causal loop of regulation*. The former leads us to thinking about minds as consisting of symbolic representations and computations, while the later leads us to think about minds as consisting of performative interactions with, and adaptation to, the world. The performative idiom depends on the materialist and feedback aspects of the regulator ideograph.

As we saw in Chapter 3, mechanistic materialism was a crucial element of the synthetic method and the new science of the mind which took the brain to be the embodiment of mind. In other words, psyche was to be found in the brain, and its nature could be discovered by the methodologies of the physical sciences. It is thus not surprising that three of the principle founders of cybernetics were brain scientists–two psychiatrists (Warren McCulloch & W. Ross

Ashby) and a physiologist (W. Grey Walter). That is, there was a strong interest in discovering how the mechanisms of the brain related to the organization of behavior. The error-controlled regulator was a clearly mechanistic system that nonetheless was able to exhibit robust and even goal-directed behavior. It thus offers a bridge between theories of physical mechanisms and theories of organized behavior.

Information feedback was seen by Cybernetics as essential for unlocking the secrets of the brain and behavior in mechanistic terms. The key idea of information feedback was that of an error-correcting signal being fed-back to some controlling mechanism. Thus, as the behavior of some system diverged from its desired behavior, this information would be fed-back to the controlling mechanism which would then regulate its output signal to direct the system towards the desired behavior. The servomechanism was the classic example from engineering. In Wiener's account of the anti-aircraft gun, the servomechanism's goal is to aim the gun in a particular direction (Galison 1994). Due to differences in the viscosity of the grease, temperature, wind, and many other factors, the same amount of force would turn the gun different distances at different times. It was thus necessary to include a feedback signal to correct over-steer and under-steer of the gun. Early attempts were not immediately successful, as it was possible to induce a situation where the gun would bounce back and forth between the over-steer and under-steer conditions, a behavior called hunting, and never arrive at the target. The solution to this problem was to "dampen" the feedback signal.

Let us consider for a moment what "information" means in the context of feedback. While the cyberneticians often cited Claude Shannon's (1948) definition of information, this was not the only way that the cyberneticians thought about information, nor does it completely suit our example of the feedback-controlled regulator. Shannon presents a measure of the quantity of information transmitted–given an expected set of messages and a degree of uncertainty for each message in the set. The measure of information is the reduction in uncertainty from the expectation of receiving a given message to its actual reception (thus a highly unexpected message transmits relatively more information than one which is expected).[17] This formulation

[17]Note that under Shannon's measure, the amount of data required to encode the message is only indirectly related to the amount of information transmitted. If all the messages in the set require the same number of bits to encode, this does not imply that they all contain the same amount of information.

presupposes a strict notion of the encoding and decoding of messages. That is to say, there is a strong distinction between the signal and the message, where the signal is a physically encoded version of the message and, apart from encoding and decoding, the message is irrelevant. In this sense, the communication of the signal is independent of its meaning, semantics, truth, etc., yet retains the qualities of fidelity (in its structural organization), of being quantifiable (in terms of its informational content) and of being empirically measurable (as a signal to noise ratio). All of these are critical to electrical engineering and signal processing, but are not often considered in theories of cognitive information. Indeed, many cognitive scientists were frustrated that this definition of information is completely devoid of any useful notion of semantic content.

While it is incredibly useful to be able to quantify the complexity of the required structure of a signal, such as when measuring the capacity of a channel to transmit it on, there are limitations to its applications to any model of mind. Primarily these follow from the required assumptions: that there is a fixed and finite set of messages to be transmitted, that these be known completely and in advance, and that the signal be properly encoded for transmission. These are rather difficult conditions for minds to meet, because the world appears to be open to the introduction of new messages and new signals, and both appear to be potentially infinite sets. Moreover, even though signals arrive at the mind through sensations and cognition, it is difficult to assume that they were specifically encoded as messages, even if they are interpreted as such. Even though we accept that natural signals, *e.g.,* the refraction of light off of surfaces, transmit information, it is possible to see this as being a matter of causal history rather than a symbolic encoding from a determinate set of signals–and causal history need not produce unambiguous encodings. That is, it seems much more natural that organisms attenuated to some particular signals in the world and *learned* to make use of them, even if the world does not always produce them unambiguously. While they may interpret this information, learning how to do this requires interaction with the environment in multiple modalities, and above all requires the agency of the mind as a regulator of certain aspects of its environment. This need not be construed as "learning the code" but can instead be understood as "making practical use of" the information received. And this was what W. Ross Ashby did with Shannon's definition; he simply used it as a quantifiable measurement in his calculations, not as a model of mind itself.

The idea of feedback loops and their implicit relation to circular causality were also a key element of the servomechanism ideograph. The idea of feedback is utterly essential to both cybernetics and the discussion of performative representations which follows. Feedback is best thought of as a causal loop. The loop starts with an action, which causes changes in the world, resulting in signals that are fed-back to the agent. It is important to note that it is a strictly causal, material sequence. It does not require the strict notion of symbolic encoding and decoding that is associated with Shannon's definition. It does require an interplay of the agency of a regulator, and the agency of the world. There is an implicit sense of structure in this causal loop–there has to be an ability to distinguish the consequences of an action on the part of the agent, *i.e.,* a processing of the feedback information. In general this is the error-signal–but it need not be an encoded signal.

The loop also implies the persistent involvement of the environment in the actions of an agent. This involvement is simultaneously causal and semantic–and this is what I shall mean by "performative." If the feedback loop is the causal mechanism which determines the significance of the agent's actions as is argued, then it is an essential element of its semantics. The agent and the environment are always coupled to each other, and the semantics of the agent's representations of the world are enmeshed and embedded in the world, because the agent's actions and perceptions are too. Thus the problems of mental content can be construed very differently once we are not limited to resolving them while looking at only half of the interaction–causal theories of mental content can be future-oriented as well as, or instead of, historically oriented. In other words, the feedback loops imply some form of ecological cognition.

My analysis will focus on the nature of the feedback-controlled regulator as a model of the brain and mind. This will begin with a brief overview of van Gelder's (1995) use of a particular feedback-controlled regulator, the steam-engine governor, as an example of a dynamicist model for mind. After noting some of the representational tensions latent in van Gelder's position, I examine the clearest formulation of the regulator-as-model in a famous paper by Conant and Ashby (1970), which argues that every good regulator is a model of the system it regulates. After interpreting their view of regulator-as-model we then reconsider in greater depth

the dynamicist program as articulated by van Gelder (1995, 1998), and his conception of simulation. I will argue that the regulator-as-model could inform cognitive science as to how performative representations might operate in the brain in ways that are not apparent from van Gelder's analysis of the dynamic aspects of the governor. It is his insistence on dynamic models as a particular class of abstract mathematical models that leads him into a difficult and perhaps indefensible philosophical position–arguing that cognition is non-representational. I argue instead that cognition is representational, but that these representations need not be symbolic–a cognitive system can represent how to get along in the world, without having any explicit symbols that correspond to specific aspects of the world. We begin first with a review of the major critiques of Computationalism.

5.2 Computationalism and Its Critics

While computational models are ubiquitous in cognitive science, it remains a point of contention whether this state of affairs is desirable or not. For many, computation has always been, and shall remain, definitive of cognitive science. For others, computation has in various ways blinded or retarded development in cognitive science, or otherwise outlived its usefulness. Members of the latter camp have frequently drawn upon concepts, ideas, and models from the cybernetic era in order to challenge the computational hegemony. Among these critiques of computationalist cognitive science we find Connectionism, Situated and Embodied Cognition, and Dynamicism.[18] Connectionism draws its critique from the neural network tradition that began with the McCulloch and Pitts neurons, and developed along a trajectory at times wedded to classic AI, and at times set against it as a theoretical adversary. Situated and Embodied cognition draw upon the idea that cognition occurs in agent-environment interactions. Though it is rarely invoked as the basis for the critique, this idea is deeply connected to the feedback-controlled regulators that are the subject of this chapter. This chapter will focus on Dynamicism because it most clearly stakes its claims on the nature and use of models in the sciences of the mind, and specifically the feedback-controlled regulator. Although I believe that such a focus

[18]There are, of course other sources of anti-computationalist psychology. Among these we find phenomenology, gestalt psychology, as well as specific figures like Heidegger, Wittgenstein, Vygotsky, and Gibson, whose ideas have been developed into various critiques of computational psychology, or are combined with elements of the cybernetically-derived critiques.

could also illuminate the challenges presented by Connectionism and Situated and Embodied Cognition, a full consideration of these issues is beyond the scope of the present project. But to the extent that these views contributed to the Dynamicist approach, they merit a brief overview.

Connectionism arose as the philosophical wing of certain engineering advances in neural network research during the 1970s. As we saw in the previous chapter, the idea of computation by neural networks actually predates, and in important ways inspired, the design of the stored-program computer by von Neumann. During the 1980s, Connectionism was drawing primarily on the parallel-distributed processing (PDP) concepts espoused by Rumelhart *et al*. (1986) and others as constituting a new theory of mental representation. The central contrast drawn was that Computationalism insisted on "symbolic" representations, whereas Connectionism insisted on "distributed" or "sub-symbolic" representations. In both cases the representations were computed, but the causal and semantic relationships were argued to be very different. Moreover, the distributed representations of Connectionism were argued to be more "brain-like," and therefore more biologically plausible.

According to Computationalism, intelligence resides in the rules that govern the processing of symbols. These rules only apply to symbols and processes "inside the head," as it were. This is one way of looking at the boundary between system and environment–namely that it is absolute when it comes to theorizing the mind. Whatever meanings exist in the mind, and whatever processes are involved in cognition, these must all take place inside the mental system. In other words, it promotes a vision of the mind as decontextualized and disembodied. Of course, few Computationalists would deny that contexts and bodies are important to cognition, but they do maintain that it is possible to give complete psychological descriptions or theories without going beyond the mind/brain. That is, to the extent that contexts or bodies are relevant, then they are represented symbolically within the mind, tokened on perceptual inputs, and are relevant in just the same way as any other computational element.

Situated cognition challenges this in a fundamental way by arguing that a crucial aspect of representation lies outside of the computational system–or alternatively that the computational system includes the environment. The system has behaviors as outputs, but the semantics of those outputs are not completely determined by the system that generates them. Rather, their

semantics are in part determined by the environmental context. For instance, whether or not a behavior is "intelligent" might be determined by whether or not it promotes the survival of an organism in its environment–but that just means it depends in non-trivial ways on the environment. One of the main points of the Situated critique is that reactions, such as reflexes, can be intelligent even when there is no internal representation involved in them–their semantic content resides in the external referent that the reflex is reacting to.

Reflexes also play a key role in Embodied cognition, which is so closely related to Situated cognition that I will group them together as a single critique. The difference between them rests largely on where outside of the mind they locate semantics. Situated cognition places a portion of intelligence in the situation, which includes the system and its environment. Embodied cognition looks to the more immediate environment of the mind, namely the unconscious processing of the body.[19] Embodied cognition also raises a representational challenge to Computationalism, but develops its arguments from a different starting point–from embodied knowledge as implicit knowledge, and from a notion of embodied knowledge as decentralized, and embodied cognition as distributed cognition, as well as often drawing inspiration from philosophical phenomenology.

Dynamicism is more explicit about its cybernetic roots than the other critiques of Computationalism. Though some proponents of the position are critical of various aspects of the early attempts by cybernetics, in particular the inadequate mathematical models of control mechanisms (Eliasmith 1997, 2003), the dynamicist program frequently invokes the central model of cybernetics, the feedback-controlled regulator. It is for this reason that I focus on the dynamicist critique in this chapter. Though it should be noted that my conclusions could be extended to offer insights into the Situated and Embodied cognition approaches as well. The aim of this chapter is to bring together the various threads introduced and clarified in the previous chapters for the purpose of re-examining the feedback-controlled regulator as a working model for studying the mind. We begin by first examining the problematic interpretation of regulators in van Gelder's account of Dynamicism and the recent debates that have arisen from it.

[19] Another way of looking at this is *wide computationalism* (Wilson 1994, 2004) which takes a different spin on the intuitions of situated and embodied cognition. Instead of arguing that certain aspects of cognition occur outside of mind, it extends mind out as far as the computations go, *e.g.,* your appointment calender is part of your cognitive memory, and your computer is an extension of your mind for certain kinds of cognitive tasks.

5.3 The Governor as Dynamic Model of Mind

Before we examine the conception of the regulator-as-model, it is worthwhile to consider the general outline of van Gelder's (1995) critique of the computational approach to cognition. His critique is built upon a strong distinction between "computational models," and what he calls "dynamic models." He defines computationalism accordingly:

> A general form of the computational hypothesis, then, is that cognitive systems such as people are computational systems in the sense just defined, and that cognition is the behavior of such systems, that is, sequences of configurations of symbols. An alternative form is that for any given cognitive process, the best model of that process will be drawn from the computational subcategory of systems. (van Gelder 1995, p. 367).

This characterization of computational systems as symbolic systems is in keeping with the previous chapter's characterization of symbolic simulations. It is also interesting to note that the issue is focused quite sharply upon *which models are best*–though it is less clear how "best" ought to be determined. That is, there is some ambiguity as to whether one is interested in models which offer the best scientific explanation of cognitive phenomena, or whether we might be able to substitute other, perhaps more pragmatic, criteria for models here.

Van Gelder's characterization of what stands in opposition to symbolic computational models is quite different than the characterization of naturalistic analog models we developed in the previous chapter. He develops this characterization by examining the centrifugal governor–a cleverly engineered mechanism developed by James Watt to regulate the speed of steam engines in the 18th century. Van Gelder contrasts the mechanical governor's operation to that of a hypothetical computationally engineered governor that "computes" the engine speed and adjusts the steam pressure accordingly to regulate it. The contrast is meant to illuminate the distinct nature of dynamical systems:

> The centrifugal governor is a paradigm example of a dynamical system. Perhaps the most pertinent contrast between it and the computational governor is that the states through which it evolves are not configurations of symbols but rather numerically measurable arm angles and rates of change of arm angle. Generalizing this feature, and, of course, looking over the shoulder at other textbook examples of dynamical systems and the kind of systems that are employed by dynamicists in cognitive science, we can define 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, pp. 367-8).

Not only does van Gelder invoke a central icon and metaphor of cybernetics, the governor, in his elaboration of dynamic systems, he also employs the concept of dynamics as trajectories in a state-space (also called a phase-space) that were a central feature of Ashby's (1952) *Design for a Brain*. We should not for this reason alone, however, assume that van Gelder's philosophical commitments are identical to Ashby's or cybernetics more generally.

Van Gelder's formulation of dynamic systems is couched in the notion of a range of different mathematical models for the system in question. These models are all state-dependent systems, a notion we will explore shortly, but they are distinguished by the mathematical structures they take as characterizing cognitive phenomena:

> There is a vast range of abstract state-dependent systems. Schools of thought which differ over the nature of cognition can be seen as differing over which of these abstract systems are realized by cognitive systems; or, put differently, as differing over *where* in the range of all possible systems the best models of cognition are to be found. (van Gelder 1995, 365).

Thus for van Gelder, all models of cognition can be characterized as state-dependent systems, the issue is what kind they are–or the mathematical assumptions that underlie their descriptions. It is these mathematical assumptions which he makes central to his critique of computationalism: their atemporal nature, the rigidity of their state transitions, and their discrete logical rules for system evolution. The claim is that the dynamic, flexible, continuous character of dynamic systems theory is better able to describe the behavior of cognitive systems than traditional computational methods.

For reasons which will become clear after my analysis of regulators-as-models, I believe that van Gelder's critique is well-intentioned, but loses sight of the nature of feedback regulation, in favor of mathematical nuance, and thereby develops an untenable notion of dynamical systems as being essentially nonrepresentational. Van Gelder states unequivocally that he believes the governor is a nonrepresentational system:

> While the centrifugal governor is clearly a nonrepresentational dynamical system, and while it was argued above that representation figures in a natural cluster of deep features that are jointly characteristic of computational models, in fact 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).

While certain features of the dynamical approach seem desirable, van Gelder's treatment of the governor confuses some of the central issues. That is to say, the centrifugal governor is not "clearly a nonrepresentational system." I argue to the contrary that it is a model of the most rudimentary form of a *performative representational* system. As such, it does offer "opportunities for dramatically reconceiving the nature of representation," but this is as embodied non-symbolic performative representation.

The representational suggestions that van Gelder then offers further betray the fundamental mistake of his argument. These suggestions include characterizing the governor's arm position as an abstract numerical representation, which he means to be opposed to a symbolic representation. That is, he believes that the arm position is a numerical quantity whose values can be mapped into a phase-space, and the trajectories of those values define its behavior. This behavior is then representable as a mathematical equation, or system of equations, which reconstructs the governor's behavioral trajectory. While this hints at the analog computations described in the previous chapter, it is an instance of *analog numerical* representation that is better placed alongside digital numerical representations within the class of symbolic simulations. While these representations have a different character than, *e.g.,* the symbolic representations of the computational governor, they are still symbolic representations. Granted, they may be more closely related to some concrete physical properties of a system, in this case the angle of the governor's arm, they may also relate abstract variables without any straightforward relation to concrete quantity. My suggestion is that we can instead pursue the embodied aspects of feedback-controlled regulators as being representational but not symbolic, as in the naturalistic analog simulations of the previous chapter. It is to a careful consideration of these issues that we now turn, before returning to consider van Gelder's treatment of what he calls "concrete models" as opposed to "abstract models."

5.4 Regulators as Models

While the Tortoise and Homeostat have already been shown to be specific examples of working models, and naturalistic analog simulations, they also represent a class of models that were argued by the cyberneticians to constitute a new kind of brain model. That is, they are

examples of feedback-controlled regulators. And not only were feedback-controlled regulators meant to be models *of* the brain, they were also argued to model the world in the same way that the *brain modeled the world.* It is here that the "working" aspect of the "working model" actually come into play. That is, the material capabilities of a mechanism, as demonstrated in the working model, becomes the basis for theorizing that the same mechanism may be at work in the brain. So here we find another case of double-modeling, wherein the feedback-controlled regulator is a model of its environment, and the regulator-as-model is argued to be a model of the brain and its ability to model the world.

While it seems clear that the mind does represent the world, and that these representations are structured in some way, there remains a question of how it happens that the structure of the brain can come to represent something outside itself. While it is probably the case that not all mental representation has the character of the feedback regulator, the suggestion from cybernetics is that it is worthwhile to consider what aspects of the brain's behavior might be effectively characterized this way. This form of representation seems quite appropriate for many kinds of bodily coordination, the functions of lower organisms, and the low-level cognitive processes of higher organisms. The argument made by Conant and Ashby (1970) is that *at least* as much representation as the feedback-controlled regulator offers is required for cognition.

In many cases, the relation was clear and explicit, as in Norbert Wiener's descriptions of neuro-muscular diseases which result in the "hunting" motions of human limbs that were argued to result from the same mechanical control conditions that caused "hunting" in the servomechanisms used for aiming anti-aircraft guns. There were also powerful arguments made about teleology and intentionality as being explicable in terms of the feedback-control of various goal-oriented behaviors, which bolstered a vision of these as being candidates for explaining higher-level mental behavior (Rosenblueth *et al.* 1943). Feedback mechanisms were seen by cybernetics as not only the basis of intelligent behavior in organisms, but as the general form of all communication, control and organized behavior in natural and social systems. In general, this was seen as a result of the flow of control information, and did not have anything to do with specific conceptions of models or representations *per se*.

In their classic 1970 paper, W. Ross Ashby and his student Roger Conant present a compelling case that "every good regulator of a system must be a model of that system." This paper contains within it not only an explanation of how a feedback-controlled regulator can become a model, but also offers insight into the nature of a type of embodied non-symbolic representation that may be common in biological intelligence. What this paper presents is a way of thinking about the relationships between the world, a regulator and the informational transactions between them that is largely independent of traditional psychological terminology:

> The suggestion has been made many times that *perhaps* the brain operates by building a model (or models) of its environment; but the suggestion has (so far as we know) been offered only as a possibility. A proof that model-making is necessary would give neurophysiology a theoretical basis, and would predict modes of brain operation that the experimenter could seek. The proof would tell us what the brain, as a complex regulator for its owner's survival, *must* do. We could have the basis for a theoretical neurology. (Conant & Ashby 1970, p. 90).

This brief passage reveals much about the mental theory pursued by Ashby, and to a large extent the rest of cybernetics. It claims nothing less than to offer a new foundation for cognitive neuroscience! They go on to present a formal proof which appears to be tautological, but the concepts upon which it is based are what interest us here. First there is the science-mind move, in which the concept of models taken from an understanding of how science and engineering are supposed to work is turned to the task of understanding how the mind is supposed to work. The difference from other attempts to do this is that the notion of model is presented in sufficient detail that it has tangible consequences for how we conduct mental investigations. That is to say that "model" here is not just another word for representation, as in the truism "the mind represents the world" = "the mind builds models of the world." Here the model is a consequence of the activity of a concrete system that regulates a portion of the physical world–it is performative.

There is something to this kind of an internal working model that is essentially different from symbolic ways of representing the world–tokening symbols, asserting propositions, or storing images or maps, etc. It is not that symbols, propositions, images, and maps cannot be models or parts of models, but that alone they do not do the whole job of representing–they require a situated and structured architecture to interpret them and act on them. While

computationalism assigns this task to the computer, this move alone does not resolve the issue of how symbols come to represent. The existence of the necessary perceptual, performative, and environmental processes are largely taken for granted in the computational approach. Moreover, the necessary elements of a digital computer are not always plausibly realized in a given biological cognitive system. The characterization that Conant and Ashby give is that models are the consequence of another phenomena which is not itself necessarily symbolic or computational–namely effective regulation. They also suggest that this way of looking at the brain provides a theoretical basis for neurophysiology, and leads to testable hypotheses. In this sense, they are offering a scientific model of mental models.

This brings us to the central idea of the quote above, that the brain is essentially a complex regulator, and thereby models its environment. Under this naturalistic mechanistic, materialistic view, the brain evolved not to understand the world objectively, but to regulate the behavior of the organism. As such, the natural measure of success, as judged by survival and reproductive success, is the *effective regulation* of behavior. This is opposed to a rationalist mentality in which truth is the measure of success, or the accuracy of correspondence between internal representations and the external world, or hybrid views in which truth or accuracy of belief is argued to ultimately lead to survival. It also suggests that the mind creates working models of the world rather than theoretical models of it.

Conant and Ashby formulate regulation as a function from external stimuli (what they call "disturbances") and internal states of the regulator, to actions in the world:

> We consider the regulatory situation . . . in which the set of regulatory events R and the set of events S in the rest of the system (*i.e.* in the 'reguland' *S*, which we view as *R*'s opponent) jointly determine, through a mapping ψ the outcome events Z. By an optimal regulator we will mean a regulator which produces regulatory events in such a way that *H*(Z) is minimal. (Conant & Ashby 1970, p. 95).

Effectiveness is then to be measured by the efficiency of a system as it approaches an optimal function in which the system presumably expends minimal amounts of energy in taking actions that achieve desirable events in the world. Whether the regulator itself could "know" that it has an optimal function is of course another question, as is the question of whether it has any "beliefs" about the world, but their approach does offer the objective engineering criteria as a means to evaluate performance and direct adaptive improvements.

Note also that the joint determination of future events by the world and the regulator presupposes an ontology in which the regulator and the world are causally enmeshed through feedback loops–*i.e.,* the effective regulator is a situated mind. There is thus a constant flow of information between the regulator and the world which is not necessarily symbolically encoded, but is causally realized. The output of the regulator is action, which presupposes actuators or other causally-effective means for influencing the world–*i.e.,* the effective regulator is an embodied mind. Beyond this, we have said nothing about the structure of the information that it is supplied with, or the structure of the internal states that determine its behavior. So the effective regulator has several things going for it as a non-computational mechanism. It is still free of structural commitments, but offers a means to study the specific mechanisms which determine behavior. It can also be seen as a naturalistic analog simulation as described in the previous chapter–it represents its target causally rather than symbolically.

But what then can we mean by "representation" in the context of effective regulation? Regulation is primarily about the control of outwardly directed behaviors, and it is not clear in what sense the external world could be involved in this apart from traditional notions of representation. Recalling that there are two ways of understanding "information,"[20] let us consider that "information processing" is not a matter of processing symbols, but of regulating behavior. Under both interpretations, information processing is a matter of translating inputs to outputs, environmental information to agent behavior. But this can be a purely mechanical, non-symbolic, performative translation. This is precisely what the governor does, and why it is such an important ideograph. It translates engine speed to steam-valve control mechanically, and non-symbolically. It is possible to interpret it symbolically, and thus model it computationally, but that does not make the governor an intrinsically symbolic computational system–indeed there may be other non-symbolic ways to structurally translate engine speed into steam regulation. Thus, we could consider the brain as transforming environmental information into neuronal patterns which do not derive their meaning in virtue of being symbolic representations, but rather in virtue of being causally engaged in performing a causal neural network that regulates behavior

[20]The symbolic interpretation sees information as "coded data," while the performative interpretation sees information as a "causal loop of regulation."

effectively. This seems to be the way most known neural systems work, including all reflex mechanisms–a network is formed from stimulus to response such that an action is taken.

This view comes very close to Block's (1987) conceptual role semantics, but differs in significant respects. Primarily, performative representations aim to provide an explanation of behavior, not conceptual content. Indeed, the point of arguing for non-symbolic representations is to get at a level of cognition that is below the conceptual, yet operative in cognitive behavior. Because of this different aim, performative representations do not run into the same criticisms that have been levelled against conceptual role semantics. Specifically, there is no problem in choosing between wide and narrow content of these representations, because their content is circumscribed by the causal feedback loops they participate in. As such, they either succeed or fail in effective performance, and the question of their "true referent" is irrelevant. Performative representations are not the same as functional roles either. While such representations will certainly have some functional roles, they are either effective or ineffective (and in need of refinement or correction). Again, there is no need to theorize the semantics of the internal representation apart from its external behavioral manifestations. The difference is also expressed in the distinction between cause-controlled and error-controlled regulators described below.

This construal of how the brain works may also seem perilously close to classical behaviorism–simple stimulus-response and conditioned behavior. While the conditioned behavior mechanism is certainly one way for such mechanisms to operate and develop, it need not be the only one. Thus, the view is broader than traditional behaviorism. It can also contain complicated and detailed mechanisms lying between stimulus and response, as cognitive psychology requires. The difference is that the intervening mechanisms are causal, not symbolic. It is not as if perceptions eventually reach a Turing machine deep in the brain where they are processed into actions. Perceptions are just a stage in a causal process that will interact with other neural processes and influence the behavior of the organism, or be ignored. Nor does this view preclude the possibility that there are symbolic operations in the brain. It simply requires that such symbolic processing will necessarily be derived from lower level, performative mechanisms. And this is not the sort of account that computationalism has sought.

So what makes the effective regulator representational? While there has been a trend toward non-representational cognition, rallying around claims like the "world is its own best model" and epitomized by van Gelder's Dynamic hypothesis, there is an important sense in which the effective-regulator-as-model is representational. To see this, we must begin by making an important distinction between the cause-controlled and the error-controlled regulator. This difference can best be understood by considering the classic examples of a thermostat as a feedback-controlled regulator of the temperature in the room and the steam-engine governor as we consider two key figures from Ashby and Conant's (1970) article:

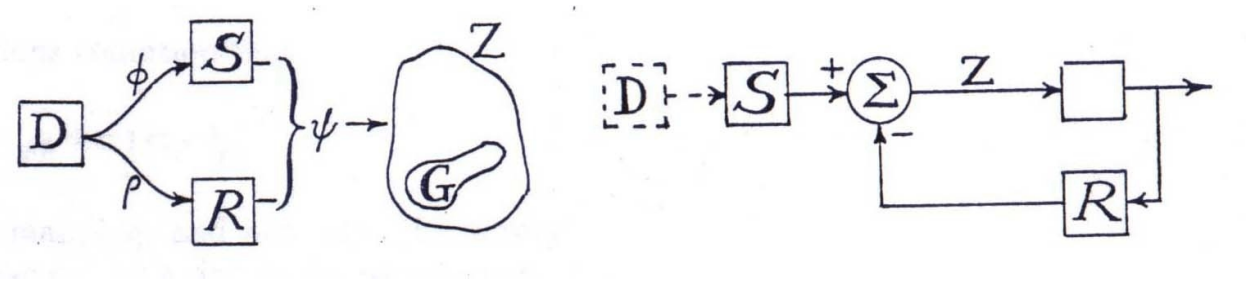


Figure 5. Two types of feedback controlled regulators (Conant & Ashby 1970, pp. 90, 92).

In the first of these we see a disturbance connected directly to the regulator, in the second, the disturbances act on the environmental system before that information reaches the regulator. There are several interesting things to note in the difference.

First there are the points raised by the authors. A regulator of the second type will never be perfect. This is because it must wait for the system to depart from a stable or desirable state before it can be returned to that state by the regulator. It only recognizes actual changes in its immediate field of sensitivity. The thermostat and governor are examples of this type: the furnace will not kick on until the temperature of the room drops, and the steam pressure will not begin to build until the steam engine slows. Conant and Ashby call such systems *error*-controlled, because they act on a deviation from a desired state, *i.e.,* an error. The other type of system detects the *causes* of disturbances directly, without waiting for the system to diverge from its desired state, and acts on these causes before the effects are realized. They call such systems *cause*-controlled. For this to happen, of course, there must be some causal structure within the

regulator which translates causes of proximate future disturbances into effective actions. But theses causal structures need not be symbolic in the computational sense.

Now let us consider where representation fits into these two diagrams. What appear here as simple arrows "p" indicate a flow of information. Aspects of D and/or S transmit some amount of information about environmental states to the regulator. Just how this processing occurs is left unexplained by Conant and Ashby. There are, however, ways we might construe this process. Let us consider one construal of the perfect and optimally efficient regulator. Such a regulator would always act to maintain the system in its ideal state (whatever that is), without perturbance. The regulator must be cause-controlled in order to achieve perfect regulation. This is because it must anticipate changes in the world before they happen. It does this by reacting to information it receives about potential causes of change to the environment–it connects input information to actions causally. This is not to say that the regulator has an explicit concept of causality, or that it represents causal rules that relate causes to effects. It does mean, however, that it reacts to causes so as to achieve effective regulation, and so there is something implicit in the causal structure of its functioning that relates to the causal structure of the system it regulates. What that causal structure is, is the mechanism we wish to understand. It is a model, but there is no reason to think it is computational.

How might an effective cause-controlled thermostat or governor work? Perhaps by reacting to the outside temperature, the humidity, the wind speed, the airflow through opening doors and windows, the current and near-future amounts of sunlight coming through windows or falling on the exterior of the building the thermostat could approximate an optimal regulator. Similarly, by reacting to the factors which determine the build-up of steam-pressure in the boiler, sensing impending changes in the workload demand on the steam-engine, and the like would improve the performance and stability of the steam-engine governor. These are surely not all the relevant inputs for these cases, indeed there might be a limitless number of possible disturbances, but they are probably sufficient for effective regulation most of the time. Still, despite the fact that there could be a very large amount of information which a cause-controlled regulator might need to obtain in order to choose the appropriate action, doing this amounts to the "processing" of this information for making the appropriate selection of actions. The regulator's model of the

environment is thus cashed out in terms of the regulator's appropriate response to changes in the environment–it is performative.

In both the thermostat case and the governor case we can imagine how a mechanism might achieve effective information processing without employing symbolic representations or computations by simply being causally coupled to relevant aspects of the environment in the right ways. Moreover, to the extent that the system is anticipatory, there is a sense in which the cause-controlled system is a model of the potential future states of the world–otherwise it would always be reacting to errors and not causes. For it to do this, it must have an internal structure that relates external causes to necessary actions prior to the realization of external effects. As a dynamic, situated, embodied model, the effective regulator is a compelling model for cognition, and, I argue, for non-symbolic representation. However, the sense in which the regulator is a naturalistic analog simulation is that *it is causal*, and this is not equivalent to the numerical dynamic model which van Gelder (1995) promotes, as we shall now consider more carefully. In doing this, we shall consider more carefully what kinds of internal representations we are considering.

5.5 Cognition and Internal Representations

In order to understand what is at stake in appealing to internal representations to explain cognition, it will be helpful to consider Andy Clark's (1997) critique of van Gelder's dynamicism and the formulation of internal representation he develops for that critique. The critique is especially significant because Clark is one of the leaders of the situated and embodied cognition approach. According to his critique, dynamicism fails in its stated goals of being a non-representational account of cognition, because it misunderstands the nature of representation in question. Clark distinguishes between strong and weak positions on internal representation, arguing that no one, including the dynamicists, rejects the weak form. And so in a trivial sense, dynamicism actually has to appeal to internal representations. Of course, this sort of representation may not turn out to be very interesting:

> Let us begin with the weakest possible sense–one with which no one takes issue, except to note that it is so weak as to be totally uninformative. This is just the bare idea of internal state. It is agreed on all sides that flexible, adaptive, intelligent behavior often

> requires a creature to respond to current situations in ways informed by past experience, on-going goals and the like. Systems that merely react, in a pre-determined way, to immediate stimuli (that will always react the same to the same stimulus) are unable to achieve this flexibility. What is needed is, at a minimum, the use of inner state to allow the agent to initiate and organize behavior without immediate environmental input, to anticipate future environmental inputs, and so on. In short, merely reactive agents are clearly inadequate to the full range of intelligent adaptive behaviors exhibited by biological organisms. Complex persisting and updatable inner state is thus at the heart of many (probably all) genuinely cognitive phenomena. (Clark 1997, p. 463).

If by weak internal representation we mean simply that a system has an internal state, then both types of regulators described by Conant and Ashby (1970) qualify as having internal representations of the systems they regulate. That is, the simplest error-controlled thermostat represents the temperature of the room as an "internal state." And this is precisely the argument Clark makes against van Gelder's claim that the governor does not represent the speed of the steam-engine in its swinging arms. Clark instead argues that the governor, in virtue of its acting to regulate the steam valve effectively, *does* represent the speed of the engine in its mechanism. I would take this further and add that the representation is non-symbolic and *performative*.[21]

Van Gelder makes the odd claim that the arm angle does not represent the speed of the steam engine, but rather it presents a numerical quantity which can be treated as a variable in a dynamic system. This strikes me as odd because it argues that the relation to an abstract model is the relevant aspect of the governor, not the concrete causal structure that regulates the steam engine. He makes his distinction between concrete and abstract systems quite clear:

> A concrete dynamical system is, of course, any concrete system that realizes an abstract dynamical system. The realization relationship here is quite different than in the computational case, however. Rather than configurations of tokens of symbol types, the concrete dynamical system is made up of quantities changing in a way that corresponds to the numerical sequences specified by the rule of evolution. This correspondence is set up by measuring the quantities, that is, by using some yardstick to assign a number to each quantity at any given point in time. For example, in the case of the centrifugal governor we set up a correspondence between the actual device and the abstract mathematical system by using the "degrees" yardstick to assign a number (for example 45) to the angle of the arm at each point in time. (van Gelder 1995, 368-9).

[21] Cummins and Poirier (forthcoming) distinguish this kind of weak representation as being what they call an "indicator" rather than a full-blown representation. For them, indicators, like a fuel gauge, are causally coupled to what they indicate, but cannot be manipulated in any useful ways–among which we might include off-line processing.

What strikes me as odd about this argument is that it seems to employ obvious symbolic and representational practices to describe the governor.[22] That is, the abstract model is obviously meant to be a representation of the concrete system, or the reverse if we start from the abstract system. So when the dynamicist models the governor mathematically, or builds a working version of her abstract model, she is engaged in a representational practice that employs either isomorphisms or similarity relations. This applies whether the abstract model has a computational or a dynamical structure.

When van Gelder comes to define what he means by representation, he avoids functional or structural definitions:

> A useful criterion of representation–a reliable way of telling whether a system contains them or not–is to ask whether there is any explanatory utility in describing the system in representational terms. If you really can make substantially more sense of how a system works by concretely describing various identifiable parts or aspects of it as representations in the above sense, that is the best evidence you could have that the system really does contain representations. Conversely, if describing the system as representational lets you explain nothing over and above what you could explain before, why on earth suppose it to be so? (van Gelder 1995, p. 352).

Instead of saying what representation is, he appeals to a notion of scientific explanation, and the utility of including representational terms in such an explanation. There are several things wrong with this approach. First, it fails to acknowledge the representational practices involved in scientific explanation and modeling. Second, it fails to consider the relevant differences between natural objects and artificial objects, and how representation might play a role in the construction of artificial objects that it does not play in natural objects. And finally, it misses what makes representation significant to cognition–it is not so clear that representations do not explain something useful in these cases, namely how thermostats and governors achieve effective regulation as causal systems.

I cannot see how the nature of the abstract formal models used in the construction of a governor alone could determine whether or not it is representational. The fact that the abstract model is sensitive to quantitative changes in the arm angle does not change the fact that the variable in the abstract model *represents arm angle*. Now it may not be a feature of the model

[22]According to cybernetic theory, this is a case of the observer determining the system under observation. Every system thus presupposes an observer, and is implicitly representational in the weak sense.

that engine speed is a function of arm angle, or vice versa, and it only appears as a correlation in their behavior. So *in that model* arm angle may not represent engine speed. But it could if we made a model which did compute arm angle as a function of engine speed, or if we specified some other means for a variable to be representational and applied it in this case. The point is that the *representational character of a variable* in an abstract model is the product of the practices of model building, which are representational in various and complicated ways. Whereas the *representational character of the abstract dynamical model* seems to be an essential aspect of its being a model of the governor.

The reverse process is a bit more confusing. If we start with an abstract dynamic model and then build a concrete model of that, are we engaged in a representational practice? Is that concrete model a simulation of the abstract model? If the concrete model is a computer simulation, there would seem to be a great deal of representational work going on. Building a computational governor would be a case in point. One makes explicit representations of various aspects of the system for the governor to use in computing some regulatory outputs. But this is not a very good characterization of James Watt's governor, how it works, or how he may have gone about building it. It probably would be mistaken to think that Watt had an abstract mathematical model of the governor as a dynamic numerical system which he sought to realize in a concrete model. Nor did he have the practical option of building the mechanism computationally. Perhaps he had an abstract idea that he needed to couple current engine speed to the steam valve, and sought out a mechanism that would do this effectively. But even then, his choice to build the governor probably came from his understanding, in an intimate practical way, of how the spindle arms behaved mechanically, and how they could be rigged up in such a way that they could regulate the steam valve on the engine. As van Gelder notes, Watt knew of the basic spindle mechanism from windmill technologies. In an instance of "modeling after," in the sense of borrowing and extension, Watt applied this mechanism in a new way to a new task. While we might want to argue that there is an abstract idea of the spindle mechanism, that is not the same as saying it was an *abstract mathematical model*. While Watt was surely aware that the arm angles changed with the speed of the engine, this was simply part of its mechanical functioning. There seems to be little reason to speculate that he saw the arm angles as

representing engine speed, whether this was symbolic or numeric. It was a causal system, and what mattered is that it worked, not that it represented anything.

So while van Gelder points to the governor as a paradigmatic dynamic system, in its traditional role it does not operate in ways that van Gelder actually describes dynamic systems theory as operating. That is, there is not necessarily any abstract dynamical model that Watt first theorized, and then realized in his concrete working governor. Watt went straight for the concrete, most likely from an understanding of the concrete. Thus, in terms of an historical account of the governor, representation seems to play no role of significance.

From the perspective of giving a scientific explanation of how a governor works, it is not clear that we need to invoke representational language, though we could. But then, Watt was an engineer worried about making steam engines that give out steady power, not a psychologist interested in intelligent behavior, so it is also not out of the question for us to consider a governor as possibly having representational properties. The issue at hand is what these representational properties might be.

If we are concerned with cognitive systems, and understanding how they generate intelligent behaviors, we could look to the governor as a model for representation in another sense. Notice that there is an important difference in the kind of internal representations involved in Conant and Ashby's error-controlled and cause-controlled regulators–the ability to anticipate future states. And this difference is captured by Clark's formulation of "strong internal representations" in what he calls off-line problem solving:

> For it does not matter what the mechanisms are (neural networks, object-oriented programs, expert systems, etc.) As long as they display certain key features. First, they must involve inner states or processes whose functional role is to coordinate the system's activity with its world (no mere correlations). Second, we must be able to identify specific inner states or processes with specific representational roles–we must be able to isolate, within the system, the inner parameters or processes that stand-in for particular extra-neural states of affairs (otherwise we confront only complex inner states implicated in successful agent-environment coordination). And lastly, the system must be capable of using these inner states or processes so as to solve problems off-line, to engage in vicarious explorations of a domain, and so on. It is this last capacity that distinguishes the genuine model-using agents from the rest. Strong internal representation is thus of a piece with the capacity to use inner models instead of real-world action and search. Inner states and processes that function as stand-ins in such models are, I suggest, genuinely

> representations for the agent and not simply useful glosses imposed from outside. (Clark 1997, p. 465).

The most obvious notion of an internal representation, that of a representation "standing-in" for something else, is captured in this formulation of strong internal representations. They "stand-in" to the extent that they can be used off-line, when their usual environmental cues are not present. This is a powerful form of representation for obvious reasons. It does not seem unreasonable to ask whether this sort of representation can be derived from the weaker internal representations which are achieved by being directly coupled to sensors and actuators. In other words, there might be a structural continuum between the two forms of representation, and the only difference is the degree to which each is coupled to the world, and can be decoupled from action. The decoupled representation might be considered "symbolic" in a sense, simply because it would be symbolic *to the agent* using it for off-line processing. It would not necessarily be symbolic in the sense of instantiating a symbolic system of the sort required by computationalism, namely by being defined by a set of syntactic features or instantiating a Turing machine. Moreover, such a representation would symbolize not an objective external object, but subjective sets of sensory-motor couplings. They are thus better thought of as being *performative representations*–representing potential performances rather than actual performances due to their being decoupled from action.

It seems that once van Gelder gets past his anti-representationalist arguments, he actually endorses a form of weak representation in cognition and dynamic systems:

> Within the conceptual repertoire of dynamics, however, there is a vast range of entities and structures that might be harnessed into representational roles; individual state variables and parameters are merely the simplest of them. For example, it is known how to construct representational schemes in which complex contents (such as linguistic structures) are assigned in a recursive manner to points in the state space of a dynamical system, such that the representations form a fractal structure of potentially infinite depth, and such that the behavior of the system can be seen as transforming representations in ways that respect the represented structure. Yet even these methods are doing little more than dipping a toe into the pool of possibilities. For example, representations can be trajectories or attractors of various kinds, or even such exotica as transformations of attractor arrangements as a system's control parameters change. (van Gelder 1995, pp. 376-7).

While these are ways for abstract dynamic models to be representations, I think the greatest insight to be gained from the governor and other feedback-controlled regulators is how a concrete system can represent without being a computational symbolic system.

There are several things I want to emphasize from this brief sketch of performative representations. First and foremost, the notion of the regulator is fundamentally a matter of control of action. In the absence of action, or at least potential action, it does not make sense to talk about regulation–it is fundamentally bound up with behavior. Compared to more traditional cognitive theories which emphasize mental images, subjective experience, and structured knowledge of the world–all open to passive observers and not necessarily requiring actors–this conception redistributes the relative weights placed upon perception and motor control. Some might even think they have succeeded in explaining cognition once we have explained how thoughts get into the mind, without worrying at all about what consequences they have in action apart from "decisions to act." By thinking about the regulation of behavior, however, motor control is to be elevated to being the ultimate purpose of cognition, while perception is neither more nor less than a means to assess the world in the service of regulating motor activity. This emphasis on behavior does not make it a purely behaviorist position, either. Primarily this is because the regulator can be viewed as temporally extensive, always receiving stimuli and generating actions in feedback loops that are larger and more complex than the stimulus-response model of classical behaviorism, and can have sophisticated causal structures lying in between stimuli and actions. In short, when the mind builds models of the world, it first builds working models, not theoretical models.

5.6 Working Models and Scientific Explanation

I now want to take a step back from these hair-splitting philosophical debates over representation. While they are interesting and I believe that the notion of a performative representation may prove worthwhile for work in artificial intelligence and cognitive neuroscience, it has already taken us on a detour from understanding the use of models in science.

It is now time to return to some basic questions facing the philosophy of science. Our detour into representation will not have been a waste of time if we are able to draw from it an important lesson about models and simulations. The lesson is this: that what makes the governor and feedback-controlled regulator good models is that they are *working* models.

I want to make this argument by pointing out that where van Gelder goes wrong is in adopting a D-N model of scientific explanation and becoming concerned with the structure of abstract models at the expense of a concern over the concrete models that do much of the work. This will draw upon the ideas we developed in previous chapters. In the course of untangling van Gelder's ideas on concrete and abstract models in cognitive science, which are mostly equivalent to what I have been calling theoretical and working models respectively, I want to further clarify the ideas of theoretical and working models, simulations, and the synthetic method that I developed in earlier chapters.

The ultimate goal of van Gelder's dynamicist project is to promote a particular way of making models in cognitive science. While he frequently draws upon our intuitions about how cognitive systems do or might work, the real issue at hand concerns what kind of explanations cognitive science seeks to provide. It is here that van Gelder's project seems to miss the lessons of cybernetics and the synthetic method. In other words, he seems to discredit the significance of concrete working models, as much of the philosophy of science has done. Moreover, cognitive science would probably do better to spend more of its energy trying to build new working systems than debating the structure of its abstract models. As the synthetic method has shown, we might be able to come up with different explanations of how a given system works, or whether another system works in the same way, but we can usually agree that it works in some respect.

For his part, van Gelder is clear that he believes theoretical explanation is the goal of building models of cognitive systems:

> Now, when cognitive scientists come to study cognitive systems, whose basic nature is a matter for empirical investigation, they often proceed by providing models. Generally speaking, a model is another entity which is either better understood already, or somehow more amenable to exploration, and which is similar in relevant respects to the explanatory target. Scientific models are concrete objects, or–more commonly–abstract mathematical entities; very often, they can be understood as state-dependent systems. If a model is

> sufficiently good, then we suppose that it somehow captures the nature of the explanatory target. What does this mean? Well, if the model is an abstract state-dependent system, then we suppose that the target system realizes the abstract system, or one relevantly like it. If the model is a concrete system, then we suppose that the model and the target system are systems of the same kind, in the sense that they both realize the same abstract system (or relevantly similar systems). Thus, even when providing a concrete model, what the scientist is really interested in determining is the abstract structure in the behavior of the target system. (van Gelder 1995, pp. 364-365).

There are several indications in this passage that van Gelder is caught up in the spell of the theoretical models and the D-N model of scientific explanation. Most clearly, he concludes by stating that scientists are "really interested in determining the abstract structure." But this need not be true. Scientists may be interested in building concrete systems, as in the synthetic method. Or they may be interested in determining the mechanisms that realize the abstract behavior structures, as William Bechtel (1998a, 1998b) has convincingly argued.

Bechtel (1998a, 1998b) argues that van Gelder is preoccupied with the "covering law" view of scientific explanation, or what I have been calling the D-N model of explanation. That is, the goal of modeling for van Gelder is to capture deep theoretical truths that are generalizable to other cognitive phenomena. This is not the only goal possible, and Bechtel promotes his own (Bechtel & Richardson 1993) theory of mechanistic explanation, which seeks to decompose a systems behavior and localize the mechanisms which achieve it. The goal of this sort of explanation is not to capture general descriptions of the system's behavior, but to understand the roles and structure of the specific physical mechanisms that realize those general behaviors. Thus while a dynamicist approach might seek to explain the dynamics of muscular contractions in a rat's leg as a function of nerve pulses from a central pattern generator in rat's spine, the mechanist approach might try to explain the structure of the nerve clusters in the rat's spine that allows it to operate as a central pattern generator. And while the dynamicist is happy to move between various levels of analysis, the mechanist is always explaining one level of behavior by examining the mechanisms at a lower level of analysis (Bechtel 1998a, p.629).

While Bechtel challenges the D-N model for ignoring the role of mechanistic explanation, we can also challenge it for undervaluing the role of working models in scientific practice. Van Gelder is clearly aware that working models play a role, yet he systematically favors abstract models as being the ultimate goal of science:

> Although, as mentioned above, their primary interest is in the abstract structure of the target phenomenon, for various reasons researchers in this approach standardly provide a concrete model: an actual computer programmed so that (hopefully) it realizes the same (or a relevantly similar) abstract computational system as is realized by the cognitive systems under study. If the concrete model appears able to perform actual cognitive tasks in much the way people do, then the hypothesis that people are such systems is supported. One reason to provide a concrete model is that the abstract systems themselves are much too complex to be studied by purely analytical means. In order to determine whether the model has the right properties, the theorist lets a concrete version run from a variety of starting points (initial conditions), and observes its behavior. Another reason for providing a concrete model is that, given the complexity of the abstract systems, it is very difficult actually to discover that structure except through an iterative procedure of constructing a concrete model, testing it, making improvements, and so on. (van Gelder 1995, p. 367).

Here we find van Gelder making a strong case for the role of concrete models, but prefacing it by saying that such models are merely a means for getting at the abstract models. Much as he recognized the governor was a good model for a kind of cognition, van Gelder seems to recognize the iterative method of constructing concrete models is integral to science, though in both cases he works to undermine these important insights. And it is at just this point that his argument goes wrong, but not so far wrong that we cannot gain some insights from setting it straight.

Even as van Gelder acknowledges the pragmatic epistemic virtues of concrete working models, that they offer practical means to exploration and understanding, he still seems to maintain that this is a fundamentally inferior form of knowledge and understanding:

> As in the computational case, although the theorist's primary goal is to identify the relevant abstract structure, it is often necessary in practice to explore particular concrete models. It tends to be difficult, however, to set up and explore the behavior of a concrete dynamical system with the right properties. Fortunately there is a convenient alternative: program (that is physically configure) a computer (a concrete computational system) so that it produces sequences of symbol-configurations which *represents* points in the state trajectories of the abstract dynamical model under consideration. In such a situation, the computer does not contain numerically measurable aspects changing over time in the way that aspects of the target system are hypothesized to be changing. That is, the computer does not realize the abstract dynamical model; rather it *simulates* it. (van Gelder 1995, p. 369).

And contrarily, the digital computer does not simulate computational models, but rather "realizes" them:

> In both cases [dynamical modeling and computational modeling], there is a target system, an abstract model, and a digital computer. In the latter case, however the target is assumed to be a digital computer; the abstract model is not a dynamical system but a digital computer; and the concrete digital computer does not *simulate* but rather *realizes* the abstract system. Indeed the abstract model is often specified *by* providing the concrete computer which realizes it. (van Gelder 1998, p. 620).

The real difference is based on the metaphysics of models, and their causal structure. But it is due to the practical aspect of *working with models* not the formal structure of the mathematics employed in the abstract models. Ultimately, model building develops understanding through practical experience, trial and error, and the development of intuition.

How does this work? Let's assume for a moment that the D-N model of explanation were essentially correct, and the goal of scientific practice is to construct precise formal models that capture true laws of nature. Even so, these formal models may be so complex that they cannot be dealt with effectively on paper, or in the minds of researchers. It is at this point that they must try different practical strategies for developing models. One approach is to automate the calculations involved so as to solve equations that would take too long by hand. This was the goal of most early computing projects. A more sophisticated approach is to approximate the solution of analytically unsolvable equations. This is what Eric Winberg (2001) defines as simulation. The difference is perhaps only one of degree, as this type of simulation is only an approximation of the abstract model, but one which is pragmatically useful. We might call this kind of simulation an instance of the synthetic method, where one synthesizes a system in order to observe its behavior.

Of course, if we reject the D-N model and instead take a more pragmatic approach, then we might conceive of all models as always being practical tools for dealing with reality. In this case, the difference between automating an analytic model and simulating a synthetic model becomes simply a matter of practice, and not of kind or even degree. The abstract model is not a goal in and of itself, but a means to building various working models. Since no model is "closer" to reality in any metaphysical sense, the real differences are in the techniques used to synthesize the various models and the ways in which they work. Thus, there may be more respect for the rigor of analytic methods, but when they are needed, the synthetic methods are invaluable for

developing knowledge of complex natural systems. I favor this approach and believe it leads naturally to a symmetric view of the relative value of theoretical and working models in science.

5.7 Conclusions

What more needs to be said about models of the mind? There are clearly different ways to approach these models. There are different ways to formalize abstract models using different mathematical frameworks. There are also different ways to realize these abstract models in concrete systems. What I hope I have shown in this dissertation is the importance of modeling and simulation practices and the pragmatic epistemic virtues of the synthetic method. That is, the real benefits of these models from the experience, intuition, understanding, and technical aptitude gained in constructing them is an integral and non-trivial aspect of science. There is thus a need for a more symmetric treatment of working models relative to theoretical models in science studies. Similarly, the construction of working models in the brain is a non-trivial aspect of cognition, and there is a similar need for symmetry in the study of these kinds of models. I believe that the feedback-controlled regulator, seen as a non-symbolic causal representation, can help promote the study of these kinds of models. While theory has a role to play in science, and symbolic representation in the brain, these are ultimately dependent upon the practices of modeling in science and the effective regulation of action in the world. We thus do science and cognitive science a disservice by ignoring working models.

Works Cited

Ashby, W. Ross (1952). *Design For a Brain,* London: Chapman and Hall.

— (1956). *An Introduction to Cybernetics,* London: Chapman and Hall.

— (1961). "Brain and Computer," in *Proceedings of the 3rd International Congress of Cybernetics*, Nemur, Belgium: Association Internationale de Cybernetique, pp. 778-793.

— (1962). "Simulation of a Brain," in H. Borko (ed.), *Computer Applications in the Behavioral Sciences*, New York: Plenum Press, pp. 452-466.

— (1967). "The Place of the Brain in the Natural World," *Currents in Modern Biology*, **1**(2), pp. 95-104.

— (1968). "Information Processing in Everyday Human Activity," *BioScience*, **18**(1), pp. 190-192.

— (1972). "Analysis of the System to be Modeled," in Ralph M. Stogdill (ed.), *The Process of Model-Building in the Behavioral Sciences*, New York: W. W. Norton, pp. 94-114.

Aspray, William (1990). *John von Neumann and the Origins of Modern Computing*. Cambridge, MA: MIT Press.

Babcock, Murray (1960). *Reorganization by Adaptive Automation*. Ph.D. Thesis, Department of Electrical Engineering, University of Illinois, Urbana-Champaign.

Bechtel, William (1998a). "Dynamicists versus Computationalists: Whither mechanists?" *Behavioral and Brain Sciences*, **21**(5), p. 629.

— 1998b). "Representations and Cognitive Explanations: Assessing the Dynamcist's Challenge in Cognitive Science," *Cognitive Science*, **22**(3), pp. 295-318.

Bechtel, William, and R. C. Richardson (1993). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research.* Princeton, NJ: Princeton University Press.

Berkeley, Edmund C. (1949). *Giant Brains: Or Machines that Think*. New York: Science Editions Inc.

Block, Ned (1987). "Functional Role and Truth Conditions," *Proceedings of the Aristotelian Society LXI*, pp. 157-181.

Brooks, Rodney (1999). *Cambrian Intelligence: The Early History of the New AI.* Cambridge, MA: MIT Press.

Cartwright, Nancy (1983). *How the Laws of Physics Lie*. Oxford: Clarendon Press.

Clark, Andy (1997). "The Dynamical Challenge," *Cognitive Science*, **21**(4), pp. 461-481.

Conant, Roger, and W. Ross Ashby (1970). "Every Good Regulator of a System Must Be a Model of that System," *International Journal of Systems Science*, **1**(2), pp. 89-97.

Cordeschi, Roberto (2002). *The Discovery of the Artificial: Behavior, Mind and Machines Before and Beyond Cybernetics*. Dordrecht, The Netherlands: Kluwer.

Crevier, Daniel (1993). *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York: Basic Books.

Cummins, Robert and Pierre Poirier (forthcoming). "Representation and Indication," in Phillip Staines and Peter Slezak (eds.) *Representation in Mind.* New York: Elsevier.

de Latil, Pierre (1957). *Thinking by Machine: A Study of Cybernetics*. Translated from French by Y. M. Golla. Boston: Houghton Mifflin.

Dupuy, Jean-Pierre (2000). *The Mechanization of the Mind: On the Origins of Cognitive Science*. Translated from French by M. B. DeBevoise. Princeton, NJ: Princeton University Press.

Eames, Charles, and Ray Eames (1973). *A Computer Perspective: A Sequence of 20th Century Ideas, Events, and Artefacts from the History of the Information Machine*. Cambridge, MA: Harvard University Press.

Eliasmith, Chris (1997). "Computation and Dynamical Models of Mind," *Minds and Machines*. **7**, pp. 531-541.

— (2003). "Moving Beyond Metaphors: Understanding the Mind for what it is," *Journal of Philosophy*, **C**(10), pp. 493-520.

Feyerabend, Paul (1975). *Against Method.* New York: Verso.

Galison, Peter (1994). "The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision," *Critical Inquiry*, Autumn 1994, pp. 228-266.

Gardner, Howard (1987). *The Mind's New Science: A History of the Cognitive Revolution*. New York: Basic Books.

Giere, Ronald (1988). *Explaining Science: A Cognitive Approach.* Chicago: University of Chicago Press.

— (1999). *Science Without Laws*. Chicago: University of Chicago Press.

Goldstine, Herman H. and John von Neumann (1946). "On the Principles of Large Scale Computing Machines," given on May 15, 1946 to Mathematical Computing Advisory Panel of the Navy's Office of Research and Inventions. Reprinted in William Aspray and Arthur Burks (eds.) (1987). *Papers of John von Neumann on Computers and Computing Theory*. Cambridge, MA: MIT Press.

Hacking, Ian (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

— (1992). "Do Thought Experiments Have a Life of Their Own?," in A. Fine and M. Forbes and K. Okruhlik (eds.), *PSA 1992*, **2**, pp. 302-310.

Heims, Steve J. (1991). *The Cybernetics Group*. Cambridge, MA: MIT Press.

Hughes, R. I. G. (1997). "Models and Representations," *Philosophy of Science*, **64**, pp. S325-S336.

Kuhn, Thomas (1962). *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

McCorduck, Pamela (1979). *Machines Who Think*. San Francisco: W. Freeman & Company.

McCulloch, Warren and Walter Pitts (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics*, **5**, pp. 115-133. Reprinted in Rook McCulloch (ed.) (1989). *The Collected Works of Warren S. McCulloch*, *Volume 1*. Salinas, CA: Intersystems Publications, pp. 343-361.

Minsky, Marvin (1961). "Steps Toward Artificial Intelligence," *Proceedings of the Institute of Radio Engineers*, **49**, pp. 8-30. Reprinted in Edward Feigenbaum and Julian Feldman (eds.) (1963). *Computers and Thought*. New York: McGraw Hill, pp. 406-450.

Morgan, Mary and Margaret Morrison (eds.) (1999). *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.

Newell, Allen and Herbert Simon (1976). "Computer Science as Empirical Inquiry: Symbols and Search," *Communications of the Association of Computing Machinery*, March 1976, **19**, 113-126. Reprinted in John Haugeland (ed.) (1981). *Mind Design: Philosophy, Psychology, Artificial Intelligence*. Cambridge, MA: MIT Press, pp. 35-66.

Pickering, Andrew (1995). *The Mangle of Practice: Time, Agency, and Science*. Chicago: University of Chicago Press.

— (forthcoming). *Ontological Theatre: Cybernetics in Britain, 1940-2000*. Chapter 4: Ross Ashby; Psychiatry, Synthetic Brains and Cybernetics. pp. 1-80. Unpublished manuscript, February 2006.

Rumelhart, David, John McClelland & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volumes 1 and 2.* Cambridge, MA: MIT Press.

Rosenblueth, Arturo, Norbert Wiener and Julien Bigelow (1943). "Behavior, Purpose, and Teleology," *Philosophy of Science*, **10**, pp 18-24.

Salmon, Wesley (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

— (1989). "Four Decades of Scientific Explanation," in Philip Kitcher and Wesley Salmon (eds.), *Minnesota Studies in the Philosophy of Science Vol. XIII: Scientific Explanation*. Minneapolis: University of Minnesota Press, pp. 3-219.

Schaffer, Simon (1994). "Machine Philosophy: Demonstration Devices in Gregorian Mechanics," *Osiris*, 2nd series, **9,** Instruments, pp. 157-182.

Scheutz, Matthias (ed.) (2002). *Computationalism: New Directions*. Cambridge, MA: MIT Press.

Searle, John (1980). "Minds, Brains, and Programs," *Behavioral and Brain Sciences*, **3**, pp. 417-424. Reprinted in John Haugeland (ed.) (1981). *Mind Design: Philosophy, Psychology, Artificial Intelligence*. Cambridge, MA: MIT Press, pp. 282-306.

Shannon, Claude (1948). "A Mathematical Theory of Communication," *Bell Systems Technical Journal*, **27**, 1948, pp. 379-423, pp. 623-656. Reprinted in Claude Shannon and Warren Weaver (1949). *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press, pp. 3-91.

Smith, Brian Cantwell (2002). "The Foundations of Computing," in Matthias Scheutz (ed.), *Computationalism: New Directions.* Cambridge, MA: MIT Press, pp. 23-58.

Teller, Paul (2001). "Twilight of the Perfect Model Model," *Erkenntnis,* **55**(3), December 2001, pp. 393-415.

Teuscher, Christof (2002). *Turing's Connectionism: An Investigation of Neural Network Architectures*. New York: Springer.

Trenholme, Russell (1994). "Analog Simulation," *Philosophy of Science*, **61**, pp. 115-131.

Turing, Alan M. (1936). "On computable numbers, with an application to the *Entscheidungsproblem*." Reprinted in Martin Davis (ed.) (1965). *The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions.* New York: Raven Press.

— (1946). "Letter to W. Ross Ashby," Turing Archives, 11-20-46, Cambridge University.

— (1947). "Lecture to the London Mathematical Society on 20 February, 1947," reprinted in D. C. Ince (ed.) (1992). *Collected Works of A. M. Turing: Mechanical Intelligence.* New York: North-Holland, pp. 87-105.

— (1948). "Intelligent Machinery, Report to the Executive Committee of the National Physical Laboratory." Reprinted in D. C. Ince, (ed.), *Collected Works of A. M. Turing: Mechanical Intelligence.* New York: North-Holland, 1992, pp. 107-127.

— (1950). "Computing machinery and intelligence," *Mind*, **59**, pp. 433-460.

van Fraassen, Bas C. (1980). *The Scientific Image*. Oxford: Clarendon Press.

van Gelder, Tim (1995). "What Might Cognition Be, If Not Computation?," *Journal of Philosophy*, **92**(7), pp. 345-381.

— (1998). "The Dynamical Hypothesis in Cognitive Science," *Behavioral and Brain Sciences*, **21**(5), pp. 615-665.

von Foerster, Heinz (ed.) (1949-1953). *Cybernetics: Circular Causal and Feedback Mechanisms in Biological and Social Systems. Transactions of the Sixth Conference*, March 24-25, 1949, *Seventh Conference*, March 23-24, 1950, *Eighth Conference*, March 15-16, 1951, *Ninth Conference*, March 20-21, 1952, *Tenth Conference*, April 22-24, 1953, New York: Josiah Macy Jr. Foundation.

von Neumann, John (1945). "First Draft of a Report on the EDVAC," Moore School University of Pennsylvania. Reprinted in *IEEE Annals of the History of Computing*, **15**(4), 1993, pp. 27-75.

— (1946). "Letter to Norbert Wiener," *Proceedings of Symposia in Applied Mathematics*, **52**,1997, pp. 505-512.

— (1951). "The General and Logical Theory of Automata," in L. A. Jeffress (ed.), *Cerebral Mechanisms in Behavior, The Hixon Symposium.* New York: John Wiley & Sons, 1951. Reprinted in William Aspray and Arthur Burks (eds.) (1987). *Papers of John von*

Neumann on Computers and Computing Theory. Cambridge, MA: MIT Press, pp. 391-431.

— (1958). *The Computer and the Brain.* New Haven, CT: Yale University Press.

Walter, W. Grey (1950). "The Twenty Fourth Maudsley Lecture: The Functions of Electrical Rhythms in the Brain," *The Journal of Mental Science*, **96**(402), January 1950, pp. 1-31.

— (1953). *The Living Brain*. New York: W. W. Norton & Company.

Webb, Barbara and Thomas R. Consi (eds.) (2001). *Biorobotics: Methods and Applications.* Cambridge, MA: MIT Press.

Wiener, Norbert (1948). *Cybernetics, or Control and Communication in the Animal and the Machine*. New York: John Wiley & Sons.

— (1950). *The Human Use of Human Beings: Cybernetics and Society.* Cambridge, MA: Houghton Mifflin.

Wilson, Robert (1994). "Wide Computationalism," *Mind,* **103**(411), pp. 351-372.

— (2004). *Boundaries of the Mind: The Individual in the Fragile Sciences.* Cambridge: Cambridge University Press.

Winsberg, Eric (2003). "Simulated Experiments: Methodology for a Virtual World," *Philosophy of Science*, **70**, pp. 105-125.

Author's Biography

Peter Mario Asaro graduated *magna cum laude* from Illinois Wesleyan University with a bachelor of arts degree, majoring in philosophy and minoring in computer science. He came to the University of Illinois at Urbana-Champaign in 1994 to pursue graduate work in philosophy. After earning his masters degree in philosophy, he pursed graduate work in computer science, earning a masters degree and becoming a doctoral candidate in the computer science department. In the year between defending and depositing his dissertation he completed a postdoctoral fellowship at the Austrian Academy of Sciences in Vienna, Austria. After completing his Ph.D. in the History, Philosophy and Sociology of Science Program, he plans to continue working on his Ph.D. in computer science and to teach.