CHAOTIC LOGIC:

Language, Mind and Reality from the Perspective of Complex Systems Science

By Ben Goertzel

Logic ... an imperative, not to know the true, but to posit and arrange a world that shall be called true by us.

-- Friedrich Nietzsche

PREFACE

This book summarizes a network of interrelated ideas which I have developed, off and on, over the past eight or ten years. The underlying theme is **the psychological interplay of order and chaos**. Or, to put it another way, the interplay of **de**duction and **in**duction. I will try to explain the relationship between logical, orderly, conscious, rule-following **reason** and fluid, self-organizing, habit-governed, unconscious, chaos-infused **intuition**.

My previous two books, *The Structure of Intelligence* and *The Evolving Mind*, briefly touched on this relationship. But these books were primarily concerned with other matters: *SI* with constructing a formal language for discussing mentality and its mechanization, and *EM* with exploring the role of evolution in thought. They danced around the edges of the order/chaos problem, without ever fully entering into it.

My goal in writing this book was to go directly to the core of mental process, "where angels fear to tread" -- to tackle all the sticky issues which it is considered prudent to avoid: the nature of consciousness, the relation between mind and reality, the justification of belief systems, the connection between creativity and mental illness,.... All of these issues are dealt with here in a straightforward and unified way, using a combination of concepts from my previous work with ideas from chaos theory and complex systems science.

My approach to the mind does not fall into any of the standard "schools of thought." But neither does it stand completely apart from the contemporary scientific and intellectual scene. Rather, I draw on ideas from a variety of disciplines, and a host of conflicting thinkers. These ideas are then synthesized with original conceptions, to obtain a model that, while, fundamentally novel, has many points of contact with familiar ideas. Perhaps the most obvious connections are with Kampis's (1991) component-system theory, Edelman's (1987) theory of neuronal group selection, Nietzsche's (1968) late philosophy of mind, Chaitin's (1988) algorithmic information theory, Whorf's (1948) well-known analysis of linguistic thought, and the dynamical psychology of Ralph and Fred Abraham (1992). But there are many other important connections as well.

The ideas of this book range wide over the conceptual map; indeed, the selection of topics may appear to the reader to obey a **very** chaotic logic. And the intended audience is almost equally wide. The ideas contained here should be thought-provoking not only to theoretical psychologists and general systems theorists, but also to anyone with an interest in artificial intelligence, applied mathematics, social science, biology, philosophy or human personality. Unfortunately, the nature of the material is such that certain sections of the book will not be easy going for the general reader. However, I have done my best to minimize the amount of technical terminology, and I have flagged with **(*)**'s those few sections containing a significant amount of formalism. These sections can be skipped without tremendous loss of understanding.

In sum, I am well aware that this book will draw criticism for its ambitious choice of topic. I also realize that my approach defies the norms of every academic discipline (sometimes quietly, sometimes ostentatiously). However, I believe that one must follow one's scientific intuition where it leads. All that I ask of you, as a reader, is that you consider the ideas given here based on their own intrinsic merits, rather than how "orthodox" or "unorthodox" they may appear.

The symbiosis between logic and intuition is a very tricky thing; perhaps the subtlest phenomenon we humans have ever tried to comprehend. In order to make progress toward an understanding of this strange, fundamental symbiosis, we must summon all our powers of analysis and imagination -- and check our preconceptions at the door.

ACKNOWLEDGEMENTS

The ideas presented here were developed as a solo project. There was very little collaborative thinking involved, and what little there was involved peripheral issues. Over the years, however, many people, institutions and organizations have helped my work in less direct ways.

First of all, a few sections of this book overlap significantly with previously published articles. Thanks are due to the relevant editors and publishers for their permission to duplicate the odd section, page or paragraph here. *The Journal of Social and Evolutionary Systems*, Volume 15-1, edited by Paul Levinson, contains the papers "Psychology and Logic" and "Self-Reference, Computation and Mind" which overlap considerably with Chapter 4 and Section 7.3 respectively. Paul Levinson is an excellent editor who has been very supportive of my work. The *Proceedings of the First, Second and Third Annual Conferences of the Society for Chaos Theory in Psychology*, edited by Robin Robertson and Allan Combs (to be published shortly by Erlbaum, perhaps with a more felicitous title), contains the papers "A Cognitive Equation of Motion" and "Belief Systems as Attractors," which overlap with parts of Chapters Eight and Nine.

Next there are more personal acknowledgements. My previous two books did not include "acknowledgements" sections, so the thanks given here apply not only to *Chaotic Logic* but also to its prequels: *The Structure of Intelligence* (Springer-Verlag, 1993), and *The Evolving Mind* (Gordon and Breach, 1993).

In no particular order, I would like to acknowledge debts of one kind or another to:

Simon's Rock College, which I attended from 1982-85 and where I was introduced to Nietzsche, Whorf, Peirce, formal logic, dynamical systems theory and the philosophy of science, among other things. My unorthodox approach to intellectual work owes a lot to two Simon's Rock instructors, George Mandeville and Ed Misch, and also to the remarkably intelligent group of students who were my classmates at the Rock, especially Dave Goldberg, Bill Meinhardt, John Hancock, Mike Glanzberg, Scott Hughes, Ed Keller and Mike Duncan.

The mathematics faculty of Temple University -- their friendliness helped to restore my passion for mathematics, which, after a year-and-a-half at the Courant Institute, had nearly vanished for good. In particular, Donald Newman supported me at every stage of the arduous process of obtaining a Ph.D. in Mathematics.

Those few members of the UNLV Mathematics department who have supported me in my unusual choice of research topics: Harold Bowman, Malwane Ananda, Rohan Dalpatadu, Ashok Singh and George Miel.

The computer science department of Waikato University, where I am currently lecturing, particularly Lloyd Smith, the former department head, who made my schedule for this year! This book was written in Las Vegas but it was proofread in Hamilton; if it has fewer errors than my previous books this is because of the research-friendly New Zealand work schedule.

Fred Abraham, Sally Goerner, Larry Vandervert, Robin Robertson and Terry Marks, all affiliated with the Society for Chaos Theory in Psychology, for being so supportive of my work (and also for helping me to improve my sometimes too-dense exposition). Thanks especially to Sally and Fred.

My mother Carol Goertzel and my grandfather Leo Zwell for their unflagging warmth and encouragement; also my father, Ted Goertzel, for his encouragement and for reading and critiquing my manuscripts despite their distance from his areas of expertise.

George Klir and George Kampis for placing my book in this series.

And finally, my son Zarathustra, my wife Gwen, and my brand new son Zebulon, for providing a warm, comfortable atmosphere in which to think, write and live.

Ben Goertzel

Hamilton, New Zealand

April 1994

TABLE OF CONTENTS

Chapter One

INTRODUCTION

"Chaos theory" has, in the space of two decades, emerged from the scientific literature into the popular spotlight. Most recently, it received a co-starring role in the hit movie *Jurassic Park*. Chaos theory is billed as a revolutionary new way of thinking about complex systems -- brains, immune systems, atmospheres, ecosystems, you name it.

It is always nice to see science work its way into the mass media. But I must admit that, as a mathematician trained in chaotic dynamics, I find this sudden interest in chaos theory a little odd. The excitement about chaos theory stems from a perception that it somehow captures the complex "disorganized order" of the real world. But in fact, **chaos theory** in the technical sense has fewer well-developed real world applications than obscure areas of applied math like Lotka-Volterra equations, Markov chains, Hilbert spaces, and so forth. Where chaos is concerned, there is a rather large gap between the philosophical, prospective hype and the actual, present-day science.

To understand this gap in more detail, consider what one studies in a first course on chaos theory: discrete iterations like the tent map, the Baker map and the logistic iteration (Devaney, 1988); or else elementary nonlinear differential equations such as those leading to the Lorentz attractor. These systems are all "low-dimensional," in the sense that the state of the system at each time is specified by a single number, or a short list of numbers. And they are simple, in the sense that the rule which determines the state of the system at time t+1 from the state of the system at time t has a brief expression in terms of elementary arithmetic.

All these systems have one novel property in common: whatever state one starts the system in at "time zero," the odds are that before long the system will converge on a certain region of state space called the "attractor." The states of the system will then fluctuate around the "attractor" region forever, apparently at random. This is "chaos," a remarkable, intriguing phenomenon -- and a phenomenon which, on the surface at least, appears to have little to do with complex, self-organizing systems. It is obvious that complex systems are not pseudo-random in the same sense that these "toy model" dynamical systems are. Something more is going on.

One way to sidestep this problem is to posit that complex systems like brains present "high-dimensional dynamics with underlying low-dimensional chaos." There is, admittedly, some evidence for this view: mood cycles, nostril cycles and EEG patterns demonstrate low-dimensional chaotic attractors, as do aspects of animal behavior, and of course numerous parameters of complex weather systems.

But at bottom, the recourse to dimensionality is an evasive maneuver, not a useful explanation. The ideas of this book proceed from an alternative point of view: that complex, self-organizing systems, while unpredictable on the level of detail, are **interestingly predictable on the level of structure**. This what differentiates them from simple dynamical systems that are almost entirely unpredictable on the level of structure as well as the level of detail.

In other words, I suggest that the popular hype over **chaos** theory is actually an enthusiasm over the study of complex, self-organizing systems -- a study which is much less developed than technical chaos theory, but also far more pregnant with real-life applications. What most chaos theorists are currently doing is playing with simple low-dimensional "toy iterations"; but what most popular expositors of chaos are thinking about is **the dynamics of partially predictable structure**. Therefore, I suggest, it is time to shift the focus from simple numerical iterations to **structure dynamics**.

To understand what this means, it suffices to think a little about **chaos psychology**. Even though the dynamics of the mind/brain may be governed by a strange attractor, the **structure** of this strange attractor need not be as coarse as that of the Lorentz attractor, or the attractor of the logistic map. The **structure** of the strange attractor of a complex system contains a vast amount of information regarding the transitions from onepatterned system state to another. And **this**, not the chaos itself, is the interesting part.

Unfortunately, there is no apparent way to get at the structure of the strange attractor of a dynamical system like the brain, which presents hundreds of billions of interlinked variables even in the crudest formal models. Therefore, I propose, it is necessary to shift up from the level of physical parameters, and take a "process perspective" in which the mind and brain are viewed as **networks of interacting, inter-creating processes**.

The process perspective on complex systems has considerable conceptual advantages over a strictly physically-oriented viewpoint. It has a long and rich philosophical history, tracing back to Whitehead and Nietszche and, if one interprets it liberally enough, all the way back to the early Buddhist philosophers. But what has driven recent complex-systems researchers to a process view is not this history, but rather the inability of alternative methods to deal with the computational complexity of self-organizing systems.

George Kampis's (1991) *Self-Modifying Systems* presents a process perspective on complex systems in some detail, relating it with various ideas from chemistry, biology, philosophy and mathematics. Marvin Minsky's (1986) *Society of Mind* describes a process theory of mind; and although his theory is severely flawed by an over-reliance on ideas drawn from rule-based AI programs, it does represent a significant advance over standard "top-down" AI ideas. And, finally, Gerald Edelman's (1988) *Neural Darwinism* places the process view of the brain on a sound neurological basis.

Here, however, I will move far beyond neural Darwinism, societal computer architecture and component-system theory, and propose a precise **cognitive equation**, hypothesized to govern the creative evolution of the network of mental processes. When one views the mind and brain in terms of creative process dynamics rather than physical dynamics, one finds that fixed points and strange attractors take on a great deal of psychological meaning. Process dynamics give rise to highly structured strange attractors. Chaos is seen to be the substrate of a new and hitherto unsuspected kind of order.

1.1. COMPLEX SYSTEMS SCIENCE

Chaos theory proper is only a small part of the emerging paradigm of **complex systems science**. In thepopular literature the word "chaos" is often interpreted very loosely, perhaps even as a synonym for "complex systems science." But the distinction is an important one. Chaos theory has to do with determinism underlying apparent randomness. Complex systems science is more broadly concerned with the emergent, synergetic behaviors of systems composed of a large number of interacting parts.

To explain what complex systems science is all about, let me begin with some concrete examples. What follows is a a highly ideosyncratic "top twelve" list of some of the work in complex systems science that strikes me as most impressive. The order of the items in the list is random (or at least chaotic).

1. Alan Perelson, Rob deBoer (1990) and others have developed computer models of the immune system as a complex self-organizing system. Using these models, they have arrived at dozens of new predictions regarding immune optimization, immune memory, the connectivity structure of the immune network, and other important issues.

2. Stuart Kauffmann (1993) has, over the last three decades, systematically pursued computer simulations demonstrating the existence of "antichaos." He has found that random Boolean networks behave in a surprisingly structured way; and he has used these networks to model various biological and economic systems.

3. Gregory Bateson (1980) has modeled a variety of social and psychological situations using ideas from cybernetics. For instance, he has analyzed Balinese society as a "steady-state" system, and he has given system-theoretic analyses of psychological problems such as schizophrenia and alcoholism.

4. Gerald Edelman (1988) has devised a theory of brain function called Neural Darwinism, based on the idea that the brain, like the immune system, is a self-organizing evolving system. Similar ideas have been proposed by other neuroscientists, like Jean-Pierre Changeux (1985).

5. Starting from the classic work of Jason Brown (1988), a number of researchers have used the concept of "microgenesis" to explore the mind/brain as a self-organizing system. This point of view has been particularly fruitful for the study of linguistic disorders such as aphasia.

6. There is a very well-established research programme of using nonlinear differential equations and thermodynamics to study far-from-equilibrium self-organizing systems. The name most commonly associated with this programm is that of Ilya Prigogine (Prigogine and Stengers, 1984).

7. A diverse community of researchers (Anderson et al, 1987) have used ideas from stochastic fractal geometry and nonlinear differential equations to model the self-organization inherent in economic processes (such as the stock market).

8. G. Spencer Brown's classic book *Laws of Form* (1972) gives a simple mathematical formalism for dealing with self-referential processes. Louis Kauffmann (1986), Francisco Varela (1978) and others have developed these ideas and applied them to analyze complex systems such as immune systems, bodies and minds.

9. For the past few years the Santa Fe Institute has sponsored an annual workshop on "Artificial Life" (Langton, 1992) -- computer programs that simulate whole living environments. These programs provide valuable information as to the necessary and sufficient conditions for generating and maintaining complex, stable structures.

10. John Holland (1975) and his colleagues such as David Goldberg (1988) have constructed a research programme of "genetic optimization," in which computer simulations of evolving populations are used to solve mathematical problems.

11. Over the past decade, a loose-knit group of researchers from different fields have been exploring the applications of "cellular automata" to model various self-organizing phenomena, from fluid dynamics to immunodynamics. Cellular automata (Wolfram, 1986) are simple self-organizing systems that display many elegant emergent properties of an apparently "generic" character.

12. Vilmos Csanyi (1990), George Kampis (1991) and Robert Rosen (1992), among others, have kept alive the grand European tradition of General Systems Theory, using sophisticated ideas from mathematics and physical science to demonstrate that complex self-organizing systems must be understood to be **creating themselves**.

Complex systems science is not as yet an official academic discipline; there are no university departments of complex systems science. However, there are a few research institutes and professional organizations. For instance, the Santa Fe Institute has supported a wide variety of research in complex systems science, including the work on immunology, antichaos, artificial life and genetic optimization mentioned above. In recognition of these efforts, the Institute recently received a MacArthur Foundation "genius grant."

The Center for Complex Systems in Illinois has also, as one would expect from the name, been the location of a great deal of complex systems research, mainly dealing with applications of cellular automata. And, finally, the Society for Chaos Theory in Psychology, now in its third year, has served to bring together an impressive number of social, behavioral and physical scientists interested in studying the mind as a complex self-organizing system.

1.1.1. Chaos and "Chaos"

Parenthetically, it is worth noting that the battle for the word "chaos" is not yet over. A few weeks after I wrote the preceding paragraphs, I ran across an interesting discussion on the Internet computer network, which really drove this point home. Someone posted a news item on several computer bulletin boards, declaring the imminent creation of a new bulletin board focusing on chaos theory. The only problem remaining, the news item said, was the selection of a **name**. Many variations were suggested, from "sci.math.nonlinear" to "sci.emergence.chaos" to "sci.nonlinear" to "sci.chaos" to "sci.math.chaos" to "sci.complexity."

Most discussants rejected the names "sci.chaos and sci.math.chaos" as encouraging a mistakenly wide interpretation of the word "chaos." But the fact is that there are already several **unofficial** newsgroups dealing with the subject of complex systems science. And these are all named -- "sci.chaos"! No amount of rational argumentation can counteract a **habit**. This is nothing else but **chaotic logic** at work, in a wonderfully self-referential way. It is chaos regarding "chaos," but only if one accepts the result of this chaos, and calls complex systems science "chaos."

Perhaps one should not shed too many tears over the fact that the name "chaos theory" is at variance with standard mathematical usage. After all, mathematicians did not invent the word "chaos"! In its original theological meaning, "Chaos" simply referred to the void existing between Heaven and Earth. In other words, it had virtually nothing to do with **any** of its current meanings.

But anyhow, I am amused to report that the newsgroup finally took on the name "sci.nonlinear." This is also a misnomer, since many nonlinear systems of equations are neither chaotic nor self-organizing. Also, many complex systems have nothing to do with linear spaces and arehence not nonlinear but **a**linear. But, be that as it may, one may chalk up one for the anti-"chaos" forces!

1.1.2. General Systems Theory

All kidding aside, however, I do think that using the name "chaos theory" for complex systems science has one sigificant disadvantage. It perpetuates an historical falsehood, by obscuring the very deep connections between the modern theory of self-organizing systems and the "General Systems Theory" of the forties and fifties.

Today, it seems, the average scientist's opinion of General Systems Theory is not very good. One often hears comments to the effect that "There **is** no general systems theory. What theoretical statements could possibly be true of **every** system?" In actual fact, however, the General Systems Theory research programme was far from being a failure. Its many successes include Bateson's psychological theories, Ashby's work in cybernetics, McCulloch's groundbreaking work on neural networks, and a variety of ideas in the field of operations research.

The truth is simply that after a decade or two, General Systems Theory collapsed under the weight of its own ambitions. It was not proved "wrong" -- it said what it had to say, and then slowly disappeared. True, it did not turn out to be nearly as productive as its creators had envisioned; but this doesn't contradict the fact that it was **very productive** anyway.

What does modern complex systems science have that General Systems Theory did not? The answer, I suspect, is remarkably simple: computing power. Of the twelve contributions to complex systems science listed above, seven -- immune system modeling, "antichaos" modeling, far-from-equilibrium thermodynamics, artificial life, genetic optimization, cellular automata and fractal economics -- rely almost entirely on computer simulations of one sort or another. An eighth, Edelman's theory of Neural Darwinism, relies largely on computer simulations; and a ninth, Spencer-Brown's self-referential mathematics, was developed in the context of circuit design.

Computing power has not been the only important factor in the development of complex systems science. For example, the revolutionary neurobiological ideas of Edelman, Changeux, Brown and others would not have been possible without recent advances in experimental brain science. And my own work depends significantly not onlyon ideas derived from computer simulations, but also on the theory of algorithmic information (Chaitin, 1987), a branch of computer science that did not exist until the late 1960's. But still, it is fair to say that greater computing power was the main agent responsible for turning relatively sterile General Systems Theory into remarkably fertile complex systems science.

The systems theorists of the forties, fifties and sixties recognized, on an intuitive level, the riches to be found in the study of complex self-organizing systems. But, as they gradually realized, they lacked the tools with which to systematically compare their intuitions to real-world data. We now know quite specifically what it was they lacked: the ability to simulate complex processes numerically, and to represent the results of complex simulations pictorially. In a very concrete sense, today's "chaos theory" picks up where yesterday's General Systems Theory left off.

In the following pages, as I discuss various aspects of language, mind and reality, I will not often be directly concerned with computer simulations or technical mathematics. However, the underlying **spirit** of the book is inextricable from recent advances in mathematical chaos theory, and more generally in complex systems science. And these advances would not have been possible without 1) the philosophy of General Systems Theory, and 2) the frame of mind induced by modern computing power. Science, philosophy and technology are not easily separable.

1.1.3. Feedback Structures

Rather than letting historical reflection get the upper hand, I will end this section with a concrete example. The basic article of faith underlying complex systems science is that there are certain large-scale patterns common to the behavior of different self-organizing systems. And perhaps the simplest such pattern is the **feedback structure** -- the physical structure or dynamical process that not only maintains itself but is the agent for its own increase. Some specific examples of feedback structures are as follows:

1. Autocatalytic reactions in chemistry, such as the Belousov-Zhabotinsky reaction. Once these chemical reactions get started, they grow by feeding off themselves. Often the rate of growth fluctuates chaotically.

2. Increasing returns in economics. This refers to a situation in which the more something is sold, theeasier it becomes to sell. Such situations are apt to be unpredictable -- an historical example is the competition between VHS and Beta format videotapes.

3. Double binds in psychology. Gregory Bateson's groundbreaking theory of schizophrenia postulates feedback reactions between family members, according to which miscommunication leads to more miscommunication.

4. Chaos in immune systems. Mathematical models trace the dynamics of antibody types, as they stimulate one another to reproduce and then attack each other. In some cases this may result in concentrations of two antibody types escalating **each other** by positive feedback. In other cases it may result in low-level chaotic fluctuations.

Of course, feedback structures of a simple sort are present in simple systems as well as complex systems (every guitar player knows this). But the important observation is that feedback structures appear to be a crucial part of self-organization, regardless of the type of system involved. Parallels like this are what the complex-systems-science researcher is always looking for: they hint at general laws of behavior.

And indeed, the cognitive equation of Chapter Seven came about as an attempt to refine the notion of "complex feedback structure" into a precise, scientifically meaningful concept -- to rigorously **distinguish** between the intricate feedback structures present in economies and mind and the relatively simple feedback involved in a guitar solo.

1.2 LANGUAGE, THOUGHT AND REALITY

In this book I will be concerned with four types of psychological systems: linguistic systems, belief systems, minds and realities. All of these systems, I suggest, are strange attractors of the dynamical system which I call the "cognitive equation." And they are, furthermore, related by the following system of "intuitive equations":

Linguistic system = syntactic system + semantic system

Belief system = linguistic system + self-generating system

Mind = dual network + belief systems

Reality = minds + shared belief system

The meanings of the terms in these four "equations" will be explained a little later. But the basic idea should be, if not "clear," at least not completelyblurry. The only important caveat is as follows: the use of the "+" sign should not be taken as a statement that the two entities on the right side of each equation have significant independent functionality. For instance, syntactic systems and semantic systems may be analyzed separately in many respects, but neither can truly function without the other.

A slightly more detailed explanation of the terms in these "equations" is as follows:

1) A **linguistic system** consists of a deductive,

transformational system called a **syntactic system**, and an interdefined collection of patterns called a **semantic system**, related according to a principle called **continuous compositionality**. This view explains the role of logic in reasoning, and the plausibility of the Sapir-Whorf hypothesis.

2) A **self-generating system** consists of a collection of stochastically computable processes which act on one another to create new processes of the same basic nature. The dynamics of mind may be understood in terms of the two processes of self-generation and **pattern recognition**; this idea yields the "cognitive equation."

3) A **belief system** is a linguistic system which is also a self-generating system. Belief systems may be thought of as the "immune system" of the mind; and, just like immune systems, they may function usefully or pathologically. They are a necessary complement to the fundamental **dual network** structure of mind (as outlined in *The Evolving Mind*).

4) **Reality** and the **self** may be viewed as two particularly powerful belief systems -- these are the "master belief systems," by analogy to which all other belief systems are formed.

Each of the "equations," as these explanations should make clear, represents a novel twist on a reasonably well-known idea. For instance, the idea of linguistics as semantics plus syntax is commonplace. But what is new here is 1) a pragmatic definition of "semantics," and 2) the concept of "continuous compositionality," by which syntactic and semantic systems are proposed to be connected.

Similarly, the idea that beliefs are linguistic is not a new one, nor is the idea that beliefs collectively act to create other beliefs. But the specific formulation of these ideas given here is quite novel, and leads to unprecedentedly clear conclusions regarding the **validity** of belief systems.

The idea that mind consists of a data structure populated by belief systems is fairly common in theAI/cognitive science community. But the **relation** between the belief system and the data structure has never been thoroughly examined from a system-theoretic point of view. Neither the role of feedback in belief maintenance, nor the analogy between immune systems and belief systems, has previously been adequately explored.

And finally, the view of reality as a collective construction has become more and more common over the past few decades, not only in the increasingly popular "New Age" literature but also in the intellectual community. However, up to this point it has been nothing more than a vague intuition. Never before has it been expressed in a logically rigorous way.

The cognitive equation underlies and guides all of these complex systemic dynamics. Elements of mind, language, belief and reality exist in a condition of constant chaotic fluctuation. The cognitive equation gives the overarching structure within which this creative chaos occurs; it gives the basic shape of the "strange attractor" that is the world.

More specifically, the assertion that each of these systems is an attractor for the cognitive equation has many interesting consequences. It implies that, as Whorf and Saussure claimed, languages are **semantically closed**, or very nearly so. It implies that belief systems are **self-supporting** -- although the nature of this self-support may vary depending on the **rationality** of the belief systems. It implies that perception, thought, action and emotion form an **unbroken unity**, each one contributing to the creation of the others. And it tells us that the relation between mind and reality is one of **intersubjectivity**: minds create a reality by sharing an appropriate type of belief system, and then they live in the reality which they create.

All this is obviously only a beginning: despite numerous examples, it is fairly abstract and general, and many details remain to be filled in. However, my goal in this book is not to provide a canon of unassailable facts, but rather to suggest a new framework for studying the remarkable phenomena of language, reason and belief. Three hundred years ago, Leibniz speculated about the possibility of giving an **equation of mind**. It seems to me that, with complex systems science, we have finally reached the point where we can take Leibniz seriously -- and transform his dream into a productive research programme.

1.3 SYNOPSIS

In this section I will give an extremely compressed summary of the main ideas to be given in the following chapters. These ideas may be somewhat opaque without the explanations and examples given in the text; however, the reader deserves at least a vague idea of the structure of the arguments to come. For a more concrete idea of where all this is leading, the reader is invited to skip ahead to Chapter Eleven, where all the ideas of the previous chapters are integrated and applied to issues of practical human and machine psychology.

Chapters Two and Three: a review of the concepts of pattern, algorithmic information, associative memory and multilevel control. These ideas, discussed more thoroughly in *SI* and *EM,* provide a rigorous basis for the analysis of psychological phenomena on an abstract structural level. A special emphasis is placed here on the issue of parallel versus serial processing. The mind/brain, it is argued, is essentially a parallel processor ... but some processes, such as deductive logic, linguistic thought, and simulation of chaotic systems, involve **virtual serial processing** -- networks of processes that simulate serial computation by parallel operations.

Chapter Four: the first part of a multi-chapter analysis of the relationship between language and thought. Using the concept of a **structured transformation system**, I consider a very special kind of linguistic system, Boolean logic, with a focus on the well-known "paradoxes" which arise when Boolean logic is applied to everyday reasoning. I argue that these "paradoxes" disappear when Boolean reasoning is considered in the context of associative memory and multilevel control. This implies that there is nothing problematic about the mind using Boolean logic in appropriate circumstances -- a point which might seem to be obvious, if not for the fact that it has never been demonstrated before. The standard approach in formal logic is simply to ignore the paradoxes!

Chapter Five: this analysis of Boolean logic is extended to more general linguistic systems. It is argued that, as a matter of principle, a linguistic system cannot be understood except in the context of a particular mind. In this spirit, I give a new analysis of **meaning**, very different from the standard Tarski/Montague possible worlds approach. According to the newapproach, the meaning of a phrase is the set of all patterns associated with it. This implies that meaning is fundamentally systematic, because many of the patterns associated with a given phrase have to do with **other phrases**. In this view, it is not very insightful to think about the meaning of a linguistic entity in isolation. The concept of meaning is only truly meaningful in the context of a whole **linguistic system** -- which is in turn only meaningful in the context of some particular mind.

Chapter Six: the connections between language, logic, reality, thought and consciousness are explored in detail. First, the pattern-theoretic analysis of language is applied to one of the more controversial ideas in twentieth-century thought: the Sapir-Whorf hypothesis, which states that patterns of thought are controlled by patterns of language. Then I discuss the role of consciousness in integrating language with other thought processes. A new theory of consciousness is proposed, which clarifies both the biological bases of awareness and the fundamental relation between mind and the external world.

Chapter Seven: a brief excursion into the most impressive modern incarnation of General Systems Theory, George Kampis's theory of component-systems, which states that complex self-organizing systems **construct themselves** in a very basic sense. After reviewing and critiquing Kampis's ideas, I introduce the novel concept of a **self-generating system**.

Chapter Eight: formulates a "dynamical law for the mind," the **cognitive equation**. This is a dynamical iteration on the level of **processes** and **structures** rather than numerical variables. It is argued that complex systems such as minds and languages are **attractors** for this equation: they supply the structure overlying the chaos of mental dynamics. Learning, and in particular language acquisition, are explained in terms of the **iteration** of the cognitive equation.

Chapter Nine: having discussed linguistic systems and self-generating systems, I introduce a concept which synthesizes them both. This is the **belief system**. I argue that belief is, in its very essence, systematic -- that, just as it makes little sense to talk about the meaning of an isolated word or phrase, it makes little sense to talk about a single belief, in and of itself. Using examples from psychology and the history of science, I develop the idea that a belief system is a **structured transformation system**, fairly similar in construction to a language.

And in this context, I consider also the question of the **quality** of a belief system. If one takes the system-theoretic point of view, then it makes little sense to talk about the "correctness" or "incorrectness" of a single belief. However, it is possible to talk about a **productive** or **unproductive** belief system. Complex systems thinking does not prohibit normative judgements of beliefs, it just displaces these judgements from the individual-belief level to the level of belief systems.

Chapter Ten: continuing the analysis of belief, I put forth the argument that belief systems are functionally and structurally analogous to immune systems. Just as immune systems protect bodies from infections, belief systems protect expensive, high-level psychological procedures from input. A belief permits the mind to deal with something "automatically," thus protecting sophisticated, deliberative mental processes from having to deal with it. In this context, I discuss the Whorfian/Nietzschean hypothesis that **self** and **external reality** itself must be considered as belief systems.

Next, I propose that beliefs within belief systems can survive for two different reasons:

a) because they are useful for linguistic systems such as logic, or

b) because they are involved in a group of beliefs that mutually support each other regardless of external utility: i.e., because they are **in themselves attractors of the cognitive equation**

Good reasoning, I argue, is done by logical systems coupled with belief systems that support themselves mainly by process (a). On the other hand, **faulty** reasoning is done by logical systems coupled with belief systems that support themselves mainly by process (b).

Chapter Eleven: the relation between mind and reality is discussed from several different perspectives. First it is argued that self and reality are **belief systems**. Then hyperset theory and situation semantics are used to give a mathematical model of the universe in which mind and reality reciprocally contain one another. Finally, I present a series of philosophically suggestive speculations regarding the relation between psychology and quantum physics.

Chapter Twelve: the phenomenon of dissociation is used to integrate the ideas of the previous chapters into a cohesive model of mental dynamics. It is argued that minds naturally tend to separate into partially disconnected subnetworks, with significantly independent functionality. This sort of dissociation has traditionally been associated with mental disorders such as multiple personality disorder and post-traumatic stress syndrome. However, I argue that it is in fact necessary for **normal, effective logical thought**. For the competition of dissociated personality networks provides a natural incentive for the creation of self-sustaining belief systems -- which are the only type of belief systems capable of supporting creative deduction.

As well as supplying a new understanding of human personality, this idea also gives rise to a design for a new type of computer program: the **A-IS**, or "artificial intersubjectivity," consisting of a community of artificial intelligences collectively living in and **creating** their own "virtual" world. It is suggested that A-IS represents the next level of computational self-organization, after artificial intelligence and artificial life.

Chapter Two

PATTERN AND PREDICTION

Language, thought and reality form an inseparable triad. Each one is defined by the others; you can't understand any one of them until you have understood the other two. But in order to speak about this triad, I must noneless begin somewhere. I will begin with **thought**, mainly because this is the topic with which most of my previous writings have been concerned. In this chapter I will review the model of mind presented in my earlier works, embellishing where necessary and placing an emphasis on those aspects that are most relevant to the task ahead.

The key phrases for understanding this model of mind are **pattern**, **process**, and **global structure**. The mind is analyzed as a network of regularities, habits, patterns. Each pattern takes the form of a **process** for acting on other mental processes. And the avenues of access joining these processes adhere roughly to a specific global structure called a **dual network**. This is an abstract, computational way of looking at the mind. But it fits in well with the qualitative nature of current neurological data. And, as we shall see, it gives a great deal of insight into many **concrete** issues regarding the human mind.

2.1. THE LOGIC OF PATTERN

Pattern-symbolic expressions are exact, as mathematics is, but are not quantitative. They do not refer ultimately to number and dimension, as mathematics does, but to pattern and structure. Nor are they to be confused with the theory of groups orsymbolic logic, though they may be in some ways akin.

-- Benjamin Lee Whorf

Before I can talk about the structure of the mind, I must first develop an appropriate vocabulary. In this section and the next, following *The Structure of Intelligence*, I will present a general mathematical vocabulary for discussing **structure**. These ideas, while abstract and perhaps rather unexciting in themselves, are essential to the psychological ideas of the following sections.

Before getting formal, let us first take a quick intuitive tour through the main concepts to be discussed. The natural place to begin is with the concept of **pattern**. I define a pattern, very simply, as a **representation as something simpler**.

A good example is computer image generation. Suppose one wants to tell one's PC to put a certain picture up on the screen. There are many ways to do this. One is to write a program telling the computer exactly what color to make each little pixel (each dot) on the screen. But this makes for a very long program -- there are around twenty thousand pixels on the average screen.

A better way to do it is to come up with some algorithm that exploits the internal structure of the picture. For instance, if one is dealing with a figure composed of four horizontal stripes, alternatingly black and white, it is easy to tell the computer "fill in the top quarter white, the next quarter down black, the next quarter down white, and the bottom quarter white." This program will be much much shorter than the program giving a pixel-by-pixel description of the picture. It is a **pattern** in the picture.

The same approach works with more complicated pictures, even photographs of human faces. Michael Barnsley (1989), using fractal image compression techniques, has given very a short program which generates realistic portraits and landscapes. In general, computer graphics experts know how to write short programs to generate very close approximations to **all** ordinary pictures -- houses, people, dogs, clouds, molecules,.... All of these things have a certain internal structure, which the clever and knowledgeable programmer can exploit.

A screen filled with static, on the other hand, has **no** internal structure, and there is no short-cut to generating it. One can rapidly generate "pseudo-random" static that will look to the human eye like random static, but one will not be getting a close approximation to the **particular** screen of static in question.

In general, a pattern is a **short-cut** -- a way of getting some entity that is in some sense simpler than the entity itself. A little more formally, suppose the **process** y leads to the **entity** x. Then y is a pattern in x if the complexity of x exceeds the complexity of y.

2.1.1. Structure

From pattern, it is but one small step to structure. The **structure** of an entity may be defined as the set of all patterns in that entity. For instance, in a figure consisting of a circle next to a square, there are at least two patterns -- the program for generating the circle and the program for generating the square.

Next, consider a picture of 100 concentric circles. Cut the picture in half to form two parts, A and B. Neither of the two parts A or B contains a pattern involving concentric circles. But the combination of the two **does**! A pattern has emerged from the combination of two entities. In general, the **emergence** between two entities A and B may be defined as the set of all processes that are patterns in the **combination** of A and B but **not** in either A or B individually.

In all this talk about pattern, one technical point repeatedly arises. Two processes, that are both patterns in the same entity, may provide different degrees of simplification. The **intensity** of a process relative to a given entity, defined formally in the following section, is a measure of how much the process simplifies the entity -- how **strongly** the process is a pattern in the entity. It has to do with the **ratio** of the complexity of the process to the complexity of the entity.

If one considers each pattern to have an intensity, then the structure of an entity becomes what is known as a "fuzzy set." It contains all the patterns in the entity, but it contains some more "intensely" than others. And, similarly, the emergence between two entities becomes a fuzzy set.

The **structural distance** between two entities A and B may then be defined quite naturally as the total intensity of all the patterns that are either in A or notB, but in B or not A. This measures how much **structure** differentiates A from B. Thus, for instance, the structural distance between two random entities would be zero, since there would be no structure in either entity -- the amount of structure differentiating two structureless entities is zero.

These concepts may be used to measure the **total amount of structure** in an entity -- a quantity which I call the **structural complexity**. The definition of this quantity is somewhat technical, but it is not hard to describe the basic idea. If all the patterns in an entity were totally unrelated to one another (as, perhaps, with this picture of the square next to the circle discussed above), then one could define the structural complexity of an entity as the sum of the complexities of all its patterns. But the problem is, often all the patterns will **not** be totally unrelated to each other -- there can be "overlap." Basically, in order to compute the structural complexity of an entity, one begins by lining up all the patterns in the entity: pattern one, pattern two, pattern three, and so on. Then one starts with the complexity of **one** of the patterns in the entity, adds on the complexity of whatever part of the second pattern was not already part of the first pattern, then adds on the complexity of whatever part of the third pattern was not already part of the first or second patterns, and so on.

These concepts, as described here, are extremely general. Very shortly I will outline a very specific way of developing these concepts in the context of binary sequences computing machines. A few chapters later, this analysis of the complexity of sequences and machines will be extended to deal with mathematical entities called **hypersets**. However, these technical specifications should not cause one to lose sight of the extreme generality of the concepts of "pattern," "structure" and "emergence." These concepts, in and of themselves, have nothing to do with sequences, machines, or hypersets -- they are completely general and philosophical in nature. It is essential to have concrete models to work with, but one must always keep in mind that these models are only secondary tools.

Finally, one comment regarding is in order regarding complexity. I have been speaking of the complexity of an entity as though it were an "objectively" defined quantity, which an entity possessed in itself independently of any observer. But the story does not end here. One may define a process to be a pattern in A **relative to a given observer** if the result of the processis A, and if the process appears simpler to A **relative to that observer**.

2.2. PATTERN AND INFORMATION (*)

Now it is time to make the concept of pattern more precise -- to give a specific, "objective" measure of complexity. The best way to do this is with the obscure but powerful branch of mathematics known as **algorithmic information theory**.

The concept of algorithmic information was conceived in the 1960's, by Kolmogorov (1965), Chaitin (1974) and Solomonoff (1964). Where U is a universal Turing machine understood to input and output potentially infinite binary sequences, and x is a finite binary sequence, it may be defined as follows:

Definition: The **algorithmic information** I(x) contained in x is the length of the shortest self-delimiting program for computing x on U given the (infinite) input string ...000...

A **self-delimiting** program is, roughly speaking, a program which explicitly specifies its own length; this restriction to self-delimiting programs is desirable for technical reasons which we need not go into here (Chaitin, 1974). It is not hard to show, using simulation arguments, that as the length of x approaches infinity, the quantity I(x) becomes machine independent.

Bennett (1982) has criticized this definition, on the grounds that what it really measures is "degree of randomness" and not "degree of structure." It assigns a random sequence maximum complexity, and a completely repetitive sequence like ...000... minimum complexity. He defines the **logical depth** of a binary sequence x, relative to a universal Turing machine U, to be the **running time** on U of the shortest self-delimiting program which computes x on U from the (infinite) input ...000... .

The sequence consisting of the first billion digits of pi has low algorithmic information, but, apparently, high logical depth. It can be proved that, as n goes to infinity, the vast majority of binary sequences of length n have near-maximal algorithmic information and logical depth.

Moshe Koppel (1987) has formulated a third measure of complexity, which he calls "sophistication" or "meaningful complexity." He has shown that for large n its behavior is similar to that of logical depth. Anapproximate **opposite** of the sophistication of a sequence is given by the **crudity** defined as follows. Instead of simply considering a program y for computing the sequence x, let us consider a program y that computes the sequence x from the input sequence z. Then the crudity of a pair (y,z) may be defined as |z|/|y|, where |z| denotes the length of the sequence z and |y| denotes the length of the sequence y.

SI discusses in detail the qualitative properties of sophistication, algorithmic information, logical depth, crudity and a number of hybrid complexity measures. It also introduces a completely new measure of complexity, called the **structural complexity**. Structural complexity differs significantly from all of the algorithmic information based complexity measures discussed above. It does not refer to one distinguished way of computing a sequence -- the shortest, the most "sophisticated," etc. Rather, it considers all possible economical strategies for computing a sequence, where an economical strategy for computing x -- or more succinctly a **pattern in x** -- may be defined as follows, given a fixed universal Turing machine U.

Definition: A **pattern** in x is a self-delimiting program y which computes x on U from the input ...000z000... (it is understood that z extends to the right of the tape head of U), so that the length of y plus the length of z is less than the length of x. Where | | denotes length, this may be written

$$|y| + |z| < |x|$$

The **intensity** of (y,z) in x is the quantity

$$1 - (|y| + |z|)/ |x|$$

(note that intensity is always positive if (y,z) is actually a pattern in x, and it never exceeds one).

Note that no generality would be lost if z were set equal to 0, or some other constant value. However, in many applications the (y,z) notation is useful.

We have introduced algorithmic information as an "objective" complexity measure, which makes the theory of pattern concrete. But this "objective" measure may be used to generate other, "subjective" complexity measures. To see how this can be done, assume some standard "programming language" L, which assigns to each program y a certain binary sequence L(y). The specifics of L are irrelevant, so long as it is computable on a Turing machine, and it does not assign the same sequence to anytwo different programs. Where U is a universal Turing machine and v and w are binary sequences, one may then propose:

Definition: The **relative information** I(v|w), relative to U and L, is the length of the shortest self-delimiting program which computes v, on U, from input L(y), where y is a minimal-length self-delimiting program for computing v on U.

Obviously, if v and w have nothing in common, I(v,w)=I(v). And, on the other hand, if v and w have a large common component, then both I(v,w) and I(w,v) are very small. If one sets |y| = I(y|x), one has a measure of complexity relative to x.

2.2.1. Fuzzy Sets and Infons

Intuitively, a "fuzzy set" is a set in which membership is not "either/or" but gradual. A good example is the set of tall people. Being nearly six foot, I belong to the set of tall people to a somewhat higher degree than my friend Mike who is five nine, and to a much higher degree than Danny DeVito, but to a much lower degree than Magic Johnson.

Formally, a fuzzy subset of a given set E is defined as a function d_E from E into some interval [0,a]. Where x is in E, I will write $d_E(x)$ for the degree of membership of x in E. $d_E(x)=0$ means that x is **not** an element of E. Unless it is specified otherwise, the reader should assume that a=1, in which case $d_E(x)=1$ means that x is **completely** an element of E. The usual algebra placed on fuzzy sets is

d_E(x **union** y) = max{ $d_E(x)$, $d_E(y)$ },

d_E(x **intersect** y) = min{ $d_E(x)$, $d_E(y)$ }

but I shall not require these operations (Kandel, 1986). The only operation I will require is the **fuzzy set distance** |E - F|, defined for finite sets as the sum over all x of the difference $|d_E(x)-d_F(x)|$.

In Chapter Three I will introduce a few ideas from situation semantics (Barwise and Perry, 1981; Barwise, 1989), which speaks about **infons** and **situations**. I will define an infon as a fuzzy set of patterns, and will make sporadic use of the following quasi-situation-theoretic notations:

s |-- i means that i is a fuzzy set of patterns in s

s |-- i //x means that i is a fuzzy set of patterns in s, where

complexity is measured **relative** to x, i.e. by I(,x)

s |-- i //x to degree a,

(s,i,x,a), and

d(s,i,x) = a all mean that the intensity of i in s, according to the complexity measure I(,x), is d. Here the intensity of i in s may be defined as the average over all w in i of the product [intensity of w in s] * [degree of membership of w in i].

In later chapters, I will call a quadruple such as (s,i,x,a) a **belief**. x is the belief-holder, i is the entity believed, s is what i is believed about, and a is the degree to which it is believed.

2.2.2. Structure and Related Ideas

As in Section 2.1, having formulated the concept of pattern, the next logical step is to define the **structure** of an entity to be the set of all patterns in that entity. This may be considered as a fuzzy set -- the degree of membership of w in the structure of x is simply the intensity of w as a pattern in x. But for now I shall ignore this fuzziness, and consider structure as a plain old set.

The structural complexity of an entity, then, measures the size of the structure of an entity. This is a very simple concept, but certain difficulties arise when one attempts to formulate it precisely. An entity may manifest a number of closely related patterns, and one does not wish to count them all separately. In words: when adding up the sizes of all the patterns in x, one must adhere the following process: 0) put all the patterns in a certain order, 1) compute the size of the first pattern, 2) compute the size of that part of the second pattern which is not also part of the first pattern, 3) compute the size of that part of the third pattern which is not also part of the first or second patterns, etc. One may then define the size |S| of a set S as the average over all orderings of the elements of S, of the number obtained by the procedure of the previous paragraph.

Where St(x) is the set of all patterns in x, one may now define the **structural complexity** of x to be the quantity |St(x)|. This is the size of the set of all patterns in x -- or, more intuitively, the total amount of regularity observable in x. It is minimal for arandom sequence, and very small for a repetitive sequence like 000...0. It deems 0101010...10 slightly more complex than 000...0, because there are more different economical ways of computing the former (for instance, one may repeat 10's, or one may repeat 01's and then append a 0 at the end). It considers all the different ways of "looking at" a sequence.

For future reference, let us define the **structure** St(D;r,s) of a discrete dynamical system D on the interval (r,s) as the set of all approximate patterns in the ordered tuple [D(r),...,D(s)], where D(t) denotes the **state of S** at time t.

And, finally, let us define the **emergence** Em(x,y) of two sequences x and y as the set St(xy) - St(x) - St(y), where xy refers to the sequence obtained by juxtaposing x and y. This measures what might be called the **gestalt** of x and y -- it consists of those patterns that appear when x and y are considered together, but not in either x or y separately. This is an old idea in psychology and it is now popping up in anthropology as well. For instance, Lakoff (1987,p.486-87) has found it useful to describe cultures in terms of "experiential gestalts" -- sets of experiences that occurs so regularly that the whole collection becomes somehow simpler than the sum of its parts.

2.3. STRUCTURE AND CHAOS

Before diving into computational psychology, let us briefly return to a topic raised in the Introduction: the meaning of "chaos." In Chapter Three it will be shown that the concept of chaos is related quite closely with certain psychological matters, such as the nature of intelligence and induction.

In mathematics, "chaos" is typically defined in terms of certain technical properties of dynamical systems. For instance, Devaney (1988) defines a time-discrete dynamical system to be chaotic if possesses three properties: 1) sensitivity to initial conditions, 2) topological transitivity, and 3) density of periodic points. On the other hand, the **intuitive** concept of chaos -- apparent randomness emergent from underlying determinism -- seems to have a meaning that goes beyond formal conditions of this sort. The mathematical definitions approximate the idea of chaos, but do not capture it.

In physical and mathematical applications of chaos theory, this is only a minor problem. One identifies chaos intuitively, then uses the formal definitions for detailed analysis. But when one seeks to apply chaos theory to psychological or social systems, the situation becomes more acute. Chaos appears intuitively to be present, but it is difficult to see the relevance of conditions such as topological transitivity and density of periodic points. Perhaps these conditions are met by certain low-dimensional subsystems of the system in question, but if so, this fact would seem to have nothing to do with the method by which we make the **educated guess** that chaos is present. "Chaos" has a pragmatic meaning that has transcends the details of point-set topology.

2.3.1. Structural Predictability

In this section I will outline an alternative point of view. For starters, I define a temporal sequence to be **structurally predictable** if knowing patterns in the sequence's past allows one to roughly predict patterns in the sequence's future. And I define a static entity to be structurally predictable if knowing patterns in **one part** of the entity allows one to predict patterns in other parts of the entity. This allows us to, finally, define an **environment** to be structurally predictable if it is somewhat structurally predictable at each time as well as somewhat structurally predictable over time.

One may give this definition a mathematical form, modeled on the standard epsilon-delta definition of continuity, but I will omit that here. The only key point is that, if an environment is structurally predictable, then patterns of higher degree have in a certain sense a higher "chance" of being found repeatedly. This shows that the assumption of a structurally predictable environment implies Charles S. Peirce's declaration that the world possesses a "tendency to take habits." The more prominent and rigid habits are the more likely to be continued.

It is interesting to think about the relationship between structural predictability and chaos. For example, one key element of chaotic behavior is **sensitive dependence on initial conditions** (or, in physicists' language, positive Liapunov exponent). Sensitive dependence means, informally, that slightly vague knowledge of the past leads to extremely vague knowledge of the future. In practical terms, if a system displays sensitive dependence, this means that it is hopeless to try to predict the exact value of its future state. Structural predictability is compatible with sensitive dependence. It is quite possible for a system to possess sensitive dependence on initial conditions, so that one can never accurately predict its future state, but still display enough regularity of overall structure that one can roughly predict future patterns. Intuitively, this appears to be the case with complex systems in the real world: brains, ecosystems, atmospheres. Exact prediction of these systems' behavior is impossible, but rough prediction of the regularities in their behavior is what we do every day.

But sensitive dependence does not, in itself, make chaos -- it is only one element of chaotic behavior. There are many different definitions of chaos, but they all center around the idea that a chaotic dynamical system is one whose behavior **is deterministic but appears random**.

A pattern-theoretic definition of chaos is as follows: an entity x is **structurally chaotic** if there are patterns in x, but if the component parts of x have few patterns besides those which are also patterns in the whole. For instance, consider the numerical sequence consisting of the first million digits of the pi: 3.1415926535... There are patterns in this sequence -- every mathematical scheme for generating the expansion of pi is such a pattern. But if one takes a subsequence -- say digits 100000 through 110000 -- one is unlikely to find any **additional** patterns there. There may be some extra patterns here and there -- say, perhaps, some strings of repeated digits -- but these won't amount to much.

Structural chaos is a weak kind of chaos. All the commonly studied examples of chaotic dynamical systems have the property that, if one records their behavior over time, one obtains a structurally chaotic series (the easiest way to see this is to use symbolic dynamics). But on the other hand, the interesting structurally predictable series are **not** structurally chaotic.

2.3.2. Attractors, Strange and Otherwise

To probe more deeply into the relation between chaos and prediction, one must consider the notion of an "attractor." Let us begin with the landmark work of Walter Freeman (1991) on the sense of smell. Freeman has written down a differential equations model of the olfactory cortex of a reptile (very similar to that of a human), and studied these equations via computersimulations. The result is that the olfactory cortex is a dynamical system which has an "attractor with wings." Recall that an **attractor** for a dynamical system is a region of the space of possible system states with the property that:

1) states "sufficiently close" to those in the attractor lead eventually to states within the attractor

2) states within the attractor lead immediately to other states within the attractor.

An attractor which consists of only one state is called a "fixed point." It is a "steady state" for the system -- once the system is close to that state, it enters that state; and once the system is in that state, it doesn't leave. On the other hand, an attractor which is, say, a circle or an ellipse is called a "limit cycle." A limit cycle represents oscillatory behavior: the system leaves from one state, passes through a series of other states, then returns to the first state again, and so goes around the cycle again and again.

And a "strange attractor," finally, is a kind of attractor which is neither a fixed point nor a limit cycle but rather a more complex region. Behavior of the system within the set of states delineated by the "strange attractor" is neither steady nor oscillatory, but continually fluctuating in a chaotic manner. More specific definitions of "strange attractor" can be found in the technical literature -- for instance, "a topologically transitive attractor" or "a topologically transitive attractor with a transversal homoclinic orbit." But, like the formal definitions of "chaos," these characterizations seem to skirt around the essence of the matter.

Freeman found that the olfactory cortex has a strange attractor -- a fixed set of states, or region of state space, within which it varies. But this strange attractor is not a formless blob -- it has a large number of "wings," protuberances jutting out from it. Each "wing" corresponds to a certain recognized smell. When the system is presented with something new to smell, it wanders "randomly" around the strange attractor, until it settles down and restricts its fluctuations to one wing of the attractor, representing the smell which it has decided it is perceiving.

This is an excellent intuitive model for the behavior of complex self-organizing systems. Each wing of Freeman's attractor represents a certain **pattern recognized** -- smell is chemical, it is just a matter of recognizing certain molecular patterns. In general, the states of a complex self-organizing systems fluctuatewithin a strange attractor that has many wings, sub-wings, sub-sub-wings, and so on, each one corresponding to the presence of a certain pattern or collection of patterns within the system. There is chaotic, pseudo-random movement within the attractor, but the structure of the attractor itself imposes a rough global predictability. From each part of the attractor the system can only rapidly get to certain other parts of the attractor, thus imposing a complex structural predictability that precludes structural chaos.

In other words, the structure of the dynamics of a complex system consists of the **patterns in its strange attractor**. The strange attractors which one usually sees in chaos texts, such as the Lorentz attractor, have very little structure to them; they are not structurally complex. But that is because these systems are fundamentally quite simple despite their chaos. A truly complex system has a highly patterned strange attractor, reflecting the fact that, in many cases, states giving rise to pattern X are more likely to lead to states giving rise to pattern Y than they are to states giving rise to pattern Z. The states **within** the attractor represent patterned states; the patterns **of** the attractor represent patterns of transition. And these two sets of patterns are not unrelated.

Chapter Three

THE STRUCTURE OF THOUGHT

Hundreds of thousands of pages have been written on the question: what is mind? Here I will dispense with the question immediately. In good mathematical form, I will define it away. **A mind is the structure of an intelligent system**.

This definition has its plusses and minuses. One may endlessly debate whether it captures every nuance of the intuitive concept of mind. But it does situate mind in the right place: neither within the physical world, nor totally disconnected from the physical world. If a mind is the structure of a certain physical system, then mind is made of **relations** between physical entities.

The question, then, is whether this system of relations that is mind has any **characteristic structure**. Are all minds somehow alike? Locomotion can be achieved by mechanisms as different as legs and the wheel -- is this kind of variety possible for the **mechanisms of intelligence**? I suggest that it is not. There is much room for variety, but the logic of intelligence dictates a certain **uniformity of overall structure**. The goal of this chapter is to outline what this uniform global structure is.

Of course, one cannot reasonably define mind in terms of intelligence unless one has a definition of intelligence at hand. So, let us say that **intelligence is the ability to optimize complex functions of complex environments**. By a "complex environment," I mean an environment which is unpredictable on the level of details, but somewhat structurally predictable. And by a "complex function," I mean a function whose **graph** is unpredictable on the level of details, but somewhat structurally predictable.

The "complex function" involved in the definition of intelligence may be anything from finding a mate to getting something to eat to building a transistor or browsing through a library. When executing any of these tasks, a person has a certain **goal**, and wants to know what set of actions to take in order to achieve it. There are many different possible sets of actions -- each one, call it X, has a certain **effectiveness** at achieving the goal.

This effectiveness depends on the environment E, thus yielding an "effectiveness function" f(X,E). Given an environment E, the person wants to find X that maximizes f -- that is maximally effective at achieving the goal. But in reality, one is never given complete information about the environment E, either at present or in the future (or in the past, for that matter). So there are two interrelated problems: one must estimate E, and then find the optimal X based on this estimate.

If you have to optimize a function that depends on a changing environment, you'd better be able to predict at least roughly what that environment is going to do in the future. But on the other hand, if the environment is **too** predictable, it doesn't take much to optimize functions that depend on it. The interesting kind of environment is the kind that couples unpredictability on the level of state with rough predictability on the level of structure. That is: one cannot predict the future state well even from a good approximation to the present and recent past states, but one **can** predict the future structure well from a good approximation to the present and recent past structure.

This is the type of partial unpredictability meant in the formulation "Intelligence is the ability to optimize complex functions of partially unpredictable environments." In environments displaying this kind of unpredictability, prediction must proceed according to **pattern recognition**. An intelligent system must recognize patterns in the past, store them in memory, and construct a model of the future based on the **assumption** that some of these patterns will approximately continue into the future.

Is there only one type of structure capable of doing this? I claim the answer is **yes**.

3.1. THE PERCEPTUAL-MOTOR HIERARCHY

My hypothesis is a simple one: every mind is a superposition of two structures: a **structurallyassociative memory** (also called "heterarchical network") and a **multilevel control hierarchy** ("perceptual-motor hierarchy" or "hierarchical network"). Both of these structures are defined in terms of their action on certain **patterns**. By superposing these two distinct structures, the mind combines memory, perception and control in a creative an effective way.

Let us begin with multilevel control. To solve a problem by the multilevel methodology, one divides one's resources into a number of levels -- say, levels ...,3,2,1,0. Level 0 is the "bottom level", which contains a number of problem-solving algorithms. Each process on level N contains a number of subsidiary processes on levels k = 1, 2, ..., N-1 -- it tells them what to do, and in return they give it feedback as to the efficacy of its instructions.

This is a simple idea of very broad applicability. One clear-cut example is the hierarchical power structure of the large corporation. Level 0 consists of those employees who actually produce goods or provide services for individuals outside the company. Level 1 consists of foremen and other low-level supervisors. And so on. The highest level comprises the corporate president and the board of directors.

3.1.1. Perception

A vivid example is the problem of perception. One has a visual image P, and one has a large memory consisting of various images z_1, z_2,..., z_M. One wants to represent the perceived image in terms of the stored images. This is a pattern recognition problem: one wants to find a pair of the form (y,z), where y*z=P and z is a subset of $\{z_1,...,z_M\}$. In this case, the multilevel methodology takes the form of a hierarchy of subroutines. Subroutines on the bottom level -- level 0 -- output simple patterns recognized in the input image P. And, for i>0, subroutines on level i output patterns recognized in the output of level i-1 subroutines. In some instances a subroutine may also instruct the subroutines on the level below it as to what sort of patterns to look for.

It appears that this is one of the key strategies of the human visual system. Two decades ago, Hubel and Wiesel (Hubel, 1988) demonstrated that the brain possesses specific neural clusters which behave as subroutines for judging the orientation of line segments. Since that time, many other neural clusters executing equally specific visual "subroutines" have been found. As well as perhaps being organized in other ways, these clusters appear to be organized in levels.

At the lowest level, in the retina, gradients are enhanced and spots are extracted -- simple mechanical processes. Next come simple moving edge detectors. The next level up, the second level up from the retina, extracts more sophisticated information from the first level up from the retina -- and so on. Admittedly, little is known about the processes two or more levels above the retina. It is clear, however, that there is a very prominent hierarchical structure, although it may be supplemented by more complex forms of parallel information processing (Ruse and Dubose, 1985).

To be extremely rough about it, one might suppose that level 1 corresponds to lines. Then level 2 might correspond to simple geometrical shapes, level 3 might correspond to complex geometrical shapes, level 4 might correspond to simple recognizable objects or parts of recognizable objects, level 5 might correspond to complex recognizable objects, and level 6 might correspond to whole scenes. To say that level 4 processes recognize patterns in the output of level 3 processes is to say that simple recognizable objects are constructed out of complex geometrical shapes, rather than directly out of lines or simple geometrical shapes. Each level 4 process is the **parent**, the controller, of those level 3 nodes that correspond to those complex geometrical shapes which make up the simple object which it represents. And it is the **child**, the controlee, of at least one of the level 5 nodes that corresponds to a complex object of which it is a part (or perhaps even of one of the level 6 nodes describing a scene of which it is a part -- level crossing like this can happen, so long as it is not the rule).

My favorite way of illustrating this multilevel control structure is to mention the three-level "pyramidal" vision processing parallel computer developed by Levitan and his colleages at the University of Massachusetts. The bottom level deals with sensory data and with low-level processing such as segmentation into components. The intermediate level takes care of grouping, shape detection, and so forth; and the top level processes this information "symbolically", constructing an overall interpretation of the scene. The base level is a 512X512 square array of processors each doing exactly the same thing to different parts of the image; and the middle level is composed of a 64X64 square array of relatively powerful processors, each doing exactly the same thing to different parts of the base-level array. Finally, the top level contains 64 very powerful processors, each one operating independently according to LISP programs. The intermediate level may also be augmented by additional connections. This three-level perceptual hierarchy appears be be an extremely effective approach to computer vision.

That orders are passed **down** the perceptual hierarchy was one of the biggest insights of the Gestalt psychologists. Their experiments (Kohler, 1975) showed that we look **for** certain configurations in our visual input. We look for those objects that we expect to see, and we look for those shapes that we are used to seeing. If a level 5 process corresponds to an expected object, then it will tell its children to look for the parts corresponding to that object, and its children will tell **their** children to look for the complex geometrical forms making up the parts to which they refer, et cetera.

3.1.2. Motor Movements

In its **motor control** aspect, this multilevel control network serves to send actions from the abstract level to the concrete level. Again extremely roughly, say level 1 represents muscle movements, level 2 represents simple combinations of muscle movements, level 3 represents medium-complexity combinations of muscle movements, and level 4 represents complex combinations of movements such as raising an arm or kicking a ball. Then when a level 4 process gives an instruction to raise an arm, it gives instructions to its subservient level 3 processes, which then give instructions to their subservient level 2 processes, which given instructions to level 1 processes, which finally instruct the muscles on what to do in order to kick the ball. This sort of control moves **down** the network, but of course all complex motions involve feedback, so that level k processes are monitoring how well their level k-1 processes are doing their jobs and adjusting their instructions accordingly. Feedback corresponds to control moving **up** the network.

In a less abstract, more practically-oriented language, Bernstein (see Whiting, 1984) has given a closely related analysis of motor control. And a very similar hierarchical model of perception and motor control has been given by Jason Brown (1988), under the name of "microgenesis." His idea is that lower levels of the hierarchy correspond to **older**, evolutionarily prior forms of perception and control.

Let us sum up. The multilevel control methodology, in itself, has nothing to do with patterns. It is a verysimple and general way of structuring perception and action: subprocesses within subprocesses within subprocesses, each subprocess giving orders to and receiving feedback from its subsidiaries. In this general sense, the idea that the mind contains a multilevel control hierarchy is extremely noncontroversial. But **psychological** multilevel control networks have one important additional property. They are postulated to deal with questions of **pattern**. As in the visual system, the processes on level N are hypothesized to **recognize patterns** in the output of the processes on level N-1, and to instruct these processes in certain **patterns of behavior**. It is pattern which is passed between the different levels of the hierarchy.

3.1.3. Genetic Programming

Finally, there is the question of how an effective multilevel control network could ever come about. As there is no "master programmer" determining which control networks will work better for which tasks, the only way for a control network to emerge is via **directed trial and error**. And in this context, the only **natural** method of trial and error is the one known as **genetic optimization** or **genetic programming**. These fancy words mean simply that

1) subnetworks of the control network which seem to be working ineffectively are randomly varied

2) subnetworks of the control network which seem to be working ineffectively are a) swapped with one another, or b) replaced with other subnetworks.

This substitution may perhaps be subject to a kind of "speciation," in which the probability of substituting subnetwork A for subnetwork B is roughly proportional to the distance between A and B in the network.

Preliminary computer simulations indicate that, under appropriate conditions, this sort of process can indeed **converge** on efficient programs for executing various perceptual and motor tasks. However, a complete empirical study of this sort of process remains to be undertaken.

3.2. STRUCTURALLY ASSOCIATIVE MEMORY

So much for the multilevel control network. Let us now turn to **long-term memory**. What I call "structurally associative memory" is nothing but a long-term memory model which the connections between processes aredetermined not by control structures, nor by any arbitrary classification system, but by **patterned relations**.

The idea of **associative** memory has a long psychological history. Hundreds, perhaps thousands of experiments on priming indicate that verbal, visual and other types of memory display **associativity of access**. For instance, if one has just heard the word "cat," and one is shown the picture of a dog, one will identify it as a "dog" very quickly. If, on the other hand, one has just heard the word "car" and one is shown the picture of a dog, identification of the dog as a "dog" will take a little bit longer.

Associative memory has also proved very useful in AI. What could be more natural than to suppose that the brain stores related entities near to each other? There are dozens of different associative memory designs in the engineering and computer science literature. Kohonen's (1984) associative memory model was one of the landmark achievements of early neural network theory; and Kanerva's (1988) sparse distributed memory, based on the peculiar statistics of the Hamming distance, has yielded many striking insights into the nature of recall.

Psychological studies of associative memory tend to deal with words or images, where the notion of "association" is intuitively obvious. Engineering associative memories use specialized mathematical definitions of association, based on inner products, bit string comparisons, etc. Neither of these paradigms seems to have a reasonably **general** method of defining association, or "relatedness."

The idea at the core of the **structurally associative** memory is that relatedness should be defined in terms of **pattern**. In the structurally associative memroy, an entity y is connected to another entity x if x is a pattern in y. Thus, if w and x have common patterns, there will be many nodes connected to both w and x. In general, if there are many short paths from w to x in the structurally associative memory, that means that w and x are closely related; that their structures probably intersect.

On the other hand, if y is a pattern **emergent** between w and x, y will not necessarily connect to w or x, but it **will** connect to the node z = w U x, if there is such a node. One might expect that, as a rough rule, z would be higher on the multilevel control network than w or z, thus interconnecting the two networks in a very fundamental way.

The memory of a real person (or computer) can never be **truly** associative -- sometimes two dissimilar things will be stored right next to each other, just by mistake. But it can be approximately structurally associative, and it can continually reorganize itself so as to maintain a high degree of structural associativity despite a continual influx of new information.

In *The Evolving Mind* this reorganization is shown to imply that structurally associative memories **evolve by natural selection** -- an entity stored in structurally associative memory is likely to "survive" (not be moved) if it **fits in well** with (has patterns in common with, generates emergent pattern cooperatively with, etc.) its environment, with those entities that immediately surround it.

3.2.1. The Dynamics of Memory

More specifically, this reorganization must be understood to take place on many different levels. There is no "memory supervisor" ruling over the entire long term memory store, mathematically determining the optimal "location" for each entity. So, logically, the only form which reorganization can take is that of **directed, locally governed trial and error**.

How might this trial and error work? The most plausible hypothesis, as pointed out in *The Structure of Intelligence*, is as follows: one subnetwork is **swapped** with another; or else subnetwork A is merely **copied** into the place of subnetwork B. All else equal, substitution will tend to take place in those regions where associativity is **worse**; but there may also be certain subnetworks that are **protected** against having their sub-subnetworks removed or replaced.

If the substitution(s) obtained by swapping or copying are successful, in the sense of improving associativity, then the new networks formed will tend not to be broken up. If the substitutions are unsuccessful, then more swapping or copying will be done.

Finally, these substitutions may take place in a **multilevel** manner: large networks may be moved around, and at the same time the small networks which make them up may be **internally** rearranged. The multilevel process will work best if, after a large network is moved, a reasonable time period is left for its subnetworks to rearrange among themselves and arrive at a "locally optimal" configuration. This same "waiting" procedure may be applied recursively: after a subnetwork is moved,it should not be moved again until its sub-subnetworks have had a chance to adequately rearrange themselves. Note that this reorganization scheme relies on the existence of certain "barriers." For instance, suppose network A contains network B, which contains network C. C should have more chance of being moved to a given position inside B than to a given position out of B. It should have more chance of moving to a given position inside A-B, than to a given position outside A (here A-B means the portion of A that is not in B). And so on -- if A is contained in Z, C should have more chance of being moved to a position in Z-A than outside Z.

In some cases these restrictions may be so strong as to prohibit any rearrangement at all: in later chapters, this sort of comprehensive rearrangement protection will be identified with the more familiar concept of **reality**. In other cases the restrictions may be very weak, allowing the memory to spontaneously direct itself through a free-floating, never-ending search for perfect associativity.

In this context, I will discuss the psychological classification of people into **thin-boundaried** and **thick-boundaried** personality types. These types would seem to tie in naturally with the notion of rearrangement barriers in the structurally associative memory. A thick-boundaried person tends to have generally stronger rearrangement barriers, and hence tends to **reify** things more, to be more resistant to mental change. A thin-boundaried person, on the other hand, has generally weaker rearrangement barriers, and thus tends to permit even fixed ideas to shift, to display a weaker grasp on "reality."

The strength and placement of these "rearrangement barriers" might seem to be a sticky issue. But the conceptual difficulty is greatly reduced if one assumes that the **memory network is "fractally" structured** -- structured in clusters within clusters ... within clusters, or equivalently networks within networks ... within networks. If this is the case, then one may simply assume that a certain "degree of restriction" comes along with each cluster, each network of networks of ... networks. Larger clusters, larger networks, have larger degrees of restriction.

The only real question remaining is **who** assigns this degree. Are there perhaps mental processes which exist mainly to **adjust** the degrees of restriction imposed by other processes? This is a large question, and a complete resolution will have to wait till later. Partof the answer, however, will be found in the following section, in the concept of the **dual network**.

3.3. THE DUAL NETWORK

Neither a structurally associative memory nor a multilevel control network can, in itself, lead to intelligence. What is necessary is to put the two together: to take a single set of entities/processes, and by drawing a single set of collections between them, structure them **both** according to structural associativity **and** according to multilevel control. This does not mean just drawing two different graphs on the same set of edges: it means that the same connections must serve as part of a structurally associative memory and part of a multilevel control network. Entities which are connected via multilevel control must, on the whole, **also** be connected via structural associativity, and vice versa.

A moment's reflection shows that it is not possible to superpose an arbitrary associative memory structure with a multilevel control hierarchy in this way. In fact, such superposition is only possible if the entities stored in the associative memory are distributed in an approximately "fractal" way (Barnsley, 1988; Edgar, 1990).

In a fractally distributed structurally associative memory, on the "smallest" scale, each process is contained in a densely connected subgraph of "neighbors," each of which is very closely related to it. On the next highest scale, each such neighborhood is connected to a collection of "neighboring neighborhoods," so that the elements of a neighboring neighborhood are fairly closely related to its elements. Such a neighborhood of neighborhoods may be called a 2'nd-level neighborhood, and in an analogous manner one may define k'th-level neighborhoods. Of course, this structure need not be strict: may be breaks in it on every level, and each process may appear at several different vertices.

A good way to understand the fractal structure of the heterarchical network is to think about the distribution of subjects in a large library. One has disciplines, sub-disciplines, sub-sub-disciplines, and so forth -- clusters within clusters within clusters, rather than a uniformly distributed field of subjects. And a good way to visualize the superposition of a hierarchical network on this structure is to postulate a head librarian dealing with each discipline, an assistant librarian dealing with each sub-sub-discipline, anassistant assistant librarian dealing with each sub-sub-sub-discipline, and so on. If one imagines that each librarian, assistant librarian, etc., gives her subsidiaries general **goals** and lets them work out their own strategies, then one has a control hierarchy that works approximately according to the multilevel methodology. The hierarchy of control is lined up perfectly with the fractal heterarchy of conceptual commonality.

A **dual network**, then, is a collection of processes which are arranged simultaneously in an hierarchical network and an heterarchical network. Those processes with close parents in the hierarchical network are, on the whole, correspondingly closely related in the heterarchical network.

This brings us back to the problem of **rearrangement barriers**. The rearrangement barriers of the associative memory network may be set up by the heterarchical network, the multilevel control network. And, strikingly, in the dual network architecture, **substituting** of subnetworks of the memory network is equivalent to **genetic optimization** of the control network. The same operation serves two different functions; the quest for associativity and the quest for efficient control are carried out in exactly the same way. This synergy between structure and dynamics is immensely satisfying.

But, important and elegant as this is, this is not the only significant interaction between the two networks. A structurally associative memory is specifically configured so as to support analogical reasoning. Roughly speaking, analogy works by relating one entity to another entity with which it shares common patterns, and the structurally associative memory stores an entity near those entities with which it shares common patterns. And the hierarchical network, the perceptual-motor hierarchy, **requires** analogical reasoning in order to do its job. The purpose of each **cluster** in the dual network is to instruct its subservient clusters in the way that it estimates will best fulfill the task given to it by its master cluster -- and this estimation is **based on reasoning analogically with respect to the information stored in its memory bank**.

Let's get a little more concrete. The brain is modeled as a dual network of neural networks. It is considered to consist of "level k clusters" of autonomous neural networks, each one of which consists of 1) a number of level k-1 clusters, all related to each other, 2) some networks that monitor and control these level k-1clusters. The degree of control involved here may be highly variable. However, the neurological evidence shows that entire knowledge bases may be outright moved from one part of the brain to another (Blakeslee, 1991), so that in some cases the degree of control is very high.

For example, a level 2 cluster might consist of processes that recognize **shapes** of various sorts in visual inputs, together with a network regulating these processes. This cluster of shape recognition processes would be organized according to the principle of structurally associative memory, so that e.g. the circle process and the ellipse process would be closer to each other than to the square process. This organization would permit the **regulating process** to execute systematic analogical search for a given shape: if in a given situation the circle process were seen to be fairly successful, but the square process not at all successful, then the next step would be to try out those processes near to the circle process.

3.3.1. Precursors of the Dual Network Model

After hinting at the dual network model in *The Structure of Intelligence*, and presenting it fully in *The Evolving Mind*, I came across two other models of mind which mirror many of its aspects. First of all, I learned that many cognitive scientists are interested in analyzing thought as a network of interconnected "**schema**" (Arbib and Hesse, 1986). This term is not always well defined -- often a "schema" is nothing more than a process, an algorithm. But Arbib equates "schema" with Charles S. Peirce's "habit," bringing it very close to the concept of pattern. The global architecture of this network of schema is not discussed, but the connection is there nonetheless.

Also, I encountered the paper "Outline for a Theory of Intelligence" by James S. Albus (1991), Chief of the Robot Systems Division of the National Institute of Standards and Technology. I was pleased to find therein a model of mind strikingly similar to the dual network, complete with diagrams such as Figure 6. Albus's focus is rather different than mine -- he is concerned with the differential equations of control theory rather than the algorithmic structure of reasoning and memory processes. But the connection between the fractal structure of memory and the hierarchical structure of control, which is perhaps the most essential component in the dual network, is virtually implicit in his theory.

Putting the schema theory developed by cognitive scientists together with the global structure identified by Albus through his robotics work, one comes rather close to a crude version of the dual network model. This is not how the dual network model was conceived, but it is a rather satisfying connection. For the dual network **structure** is, after all, a rather straightforward idea. What is less obvious, and what has not emerged from cognitive science or engineering, is the **dynamics** of the dual network. The way the dual network unifies memory reorganization with genetic optimization has not previously been discussed; nor has the dynamics of barrier formation and its relationship with consciousness, language and perception (to be explored in Chapter Six).

3.4 PREDICTION

The dual network model, as just presented, dismisses the problem of predicting the future rather cursorily. But this is not entirely justified. Prediction of the behavior of a complex system is an incredibly very difficult task, and one which lies at the very foundation of intelligence. The dual network model has no problem incorporating this sort of prediction, but something should be said about **how** its prediction processes work, rather than just about how they are interconnected.

One way to predict the future of a system, given certain assumptions about its present and about the laws governing its behavior, is to simply **simulate** the system. But this is inefficient, for a very simple physicalistic reason. Unlike most contemporary digital computers, the brain works in parallel -- there are a hundred billion neurons working at once, plus an unknown multitude of chemical reactions interacting with and underlying this neural behavior. And each neuron is a fairly complicated biochemical system, a far cry from the on-off switch in a digital computer. But when one simulates a system, one goes **one step at a time**. To a certain extent, this wastes the massive parallelism of the brain.

So, the question is, is simulation the best a mind can do, or are there short-cuts? This question ties in with some pressing problems of modern mathematics and theoretical computer science. One of the biggest trends in modern **practical** computer science is the development of parallel-processing computers, and it is of great interest to know when these computers can outperform conventional serial computers, and by what margin.

3.4.1. Discrete Logarithms (*)

For a simple mathematical example, let us look to the theory of finite fields. A finite field is a way of doing arithmetic on a bounded set of integers. For instance, suppose one takes the field of size 13 (the size must be a prime or a prime raised to some power). Then, in this field the largest number is 12. One has, for example, 12 + 1 = 0, 10 + 5 = 2, 3 x 5 = 2, and 8 x 3 = 12. One can do division in a finite field as well, although the results are often counterintuitive -- for instance, 12/8 = 3, and 2/3 = 5 (to see why, just multiply both sides by the denominator).

In finite field theory there is something called the "discrete logarithm" of a number, written **dlog_b(n)**. The discrete logarithm is defined just like the ordinary logarithm, as the inverse of exponentiation. But in a finite field, exponentiation must be defined in terms of the "wrap-around" arithmetic illustrated in the previous paragraph. For instance, in the field of size 7, 3*4* = 4. Thus one has $\mathrm{dlog}_3(4) = 4$. But how could one **compute** the log base 3 of 4, without knowing what it was? The powers of 3 can wrap around the value 7 again and again -- they could wrap around many times before hitting on the correct value, 4.

The problem of finding the discrete logarithm of a number is theoretically easy, in the sense that there are only finitely many possibilities. In our simple example, all one has to do is take 3 to higher and higher powers, until all possibilities are covered. But in practice, if the size of the field is not 7 but some much larger number, this finite number of possibilities can become prohibitively large.

So, what if one defines the dynamical system $n_k = \mathrm{dlog}_b(n_{k-1})$? Suppose one is given n_1, then how can one predict n_{1000}? So far as we know today, there is better way than to proceed in order: first get n_2, then n_3, then n_4, and so on up to n_{999} and n_{1000}. Working on n_3 before one knows n_2 is essentially useless, because a slight change in the answer for n_2 can totally chagne the answer for n_3. The only way to do all 1000 steps in parallel, it seems, would be to first compute a table of **all possible powers** that one might possibly need to know in the course of calculation. But this would require an immense number of processors; at least the square of the size of the field.

This example is, incidentally, of more than academic interest. Many cryptosystems in current use are reliant on discrete logarithms. If one could devise a quickmethod for computing them, one could crack all manner of codes; and the coding theorists would have to come up with something better.

3.4.2. Chaos and Prediction

More physicalistic dynamical systems appear to have the same behavior. The classic example is the "logistic" iteration $x_k = cx_{k-1}(1-x_{k-1})$, where c=4 or c assumes certain values between 3.8

and 4, and the x_k are discrete approximations of real numbers. This equation models the dynamics of certain biological populations, and it also approximates the equations of fluid dynamics under certain conditions.

It seems very, very likely that there is no way to compute x_n from x_1 on an ordinary serial computer, except to proceed one step at a time. Even if one adds a dozen or a thousand or a million processors, the same conclusion seems to hold. Only if one adds a number of processors roughly proportional to **2*n*** can one obtain a significant advantage from parallelism.

In general, all systems of equations called **chaotic** possess similar properties. These include equations modeling the weather, the flow of blood through the body, the motions of planets in solar systems, and the flow of electricity in the brain. The mathematics of these systems is still in a phase of rapid development. But the intuitive picture is clear. To figure out what the weather will be ninety days from now, one must run an incredibly accurate day-by-day simulation -- even with highly parallel processing, there is no viable alternate strategy.

3.4.3. Chaos, Prediction and Intelligence

A mind is the structure of an intelligent system, and intelligence relies on prediction, memory and optimization. Given the assumption that **some** past patterns will persist, a mind must always explore several different hypotheses as to **which** ones will persist. It must explore several different possible futures, by a process of predictive extrapolation. Therefore, intelligence requires the prediction of the future behavior of partially unpredictable systems.

If these systems were as chaotic as $x_k = 4x_k(1-x_k)$, all hope would be lost. But the weather system is a better example. It is chaotic in its particular details -- there is no practical way, today in 1992, to determine the temperature on July 4 1999 in Las Vegas. But there are certain persistent patterns that allow one to predict its behavior in a **qualitative** way. After all, the temperature on July 4 1999 in Las Vegas will **probably** be around 95-110 Fahrenheit. One can make probabilistic, approximate predictions -- one can recognize patterns in the past and hope/assume that they will continue.

Our definition of intelligence conceals the presupposition that **most** of the prediction which the mind has to do is analogous to this trivial weather prediction example. No step-by-step simulation is required, only inductive/analogical reasoning, supported by memory search. However, the fact remains that **sometimes** the mind will run across obstinate situations -- prediction problemss that are not effectively tackled using intuitive memory or using parallel-processing shortcuts. In these cases, the mind has no choice but to resort to direct simulation (on some level of abstraction).

The brain is a massively parallel processor. But when it runs a direct simulation of some process, it is acting like a **serial processor**. In computerese, it is running a **virtual serial machine**. The idea that the parallel brain runs virtual serial machines is not a new one -- in *Consciousness Explained* Daniel Dennett proposes that **consciousness** is a virtual serial machine run on the parallel processor of the brain. As will be seen in Chapter Six, although I cannot accept Dennett's reductionist analysis of consciousness, I find a great deal of merit in this idea.

3.5. STRUCTURED TRANSFORMATION SYSTEMS

To proceed further with my formal theory of intelligence, I must now introduce some slightly technical definitions. The concept of a **structured transformation system** will be absolutely essential to the theory of language and belief to be given in later chapters. But before I can say what a structured transformation system is, I must define a plain old **transformation system**.

In words, a transformation system consists of a set I of **initials**, combined with a set of T **transformation rules**. The initials are the "given information"; the transformation rules are methods for combining andaltering the initials into new statements. The deductive system itself, I will call D(I,T).

For instance, in elementary algebra one has transformation rules such as

$X = Y$ implies $X+Z = Y+Z$, and $XZ = YZ$

$(X + Y) + Z = X + (Y+Z)$

$X - X = 0$

$X + 0 = X$

$X + Y = Y + X$

If one is given the initial

$2q - r = 1$

one can use these transformation rules to obtain

$q = (1 + r)/2.$

The latter formula has the same content as the initial, but its form is different.

If one had a table of numbers, say

r	q
1	1
2	3/2
3	2
4	5/2

5 3

...

99 50

then the "q=(1+r)/2" would be a slightly more intense pattern in one's table than "2q+r=1." For the work involved in computing the table from "2q+r=1" is a little greater -- one must solve for q each time r is plugged in, or else transform the equation into "q=(1+r)/2."

Thus, although in a sense transformation systems add no content to their initials, they are capable of producing **new patterns**. For a list of length 100, as given above, both are clearly patterns. But what if the list were of length 4? Then perhaps "2q + r=1" would **not** be a pattern: the trouble involved in using it might be judged to exceed the difficulty of using the list itself. But perhaps q = (1+r)/2 **would still** be a pattern. It all depends on who's doing the judging of complexities -- but for **any** judge there is likely to be some list length for which one formula is a pattern and the other is not.

This is, of course, a trivial example. A better example is Kepler's observation that planets move in ellipses. This is a nice compact statement, which can be logically derived from Newton's Three Laws of Motion. But the derivation is fairly lengthy and time-consuming. So if one has a brief list of data regarding planetary position, it is quite possible that Kepler's observation will be a significant pattern, but Newton's Three Laws will not. What is involved here is the **complexity of producing x from the process y**. If this complexity is too great, then no matter how simple the process y, y will not be a pattern in x.

3.5.1. Transformation Systems (*)

In this section I will give a brief formal treatment of "transformation systems." Let W be any set, let A be a subset of W, called the set of "expressions"; and let I = $\{W_1, W_2, ..., W_n\}$ be a subset of W, called the set of **initials**. Let W* denote the set {W,WxW,WxWxW,...). And let T = $\{F_1,F_2,...,F_n\}$ be a set of **transformations**; that is, a set of functions each of which maps some elements of W* into elements of A. For instance, if W were a set of propositions, one might have F_1(x,y)= x and y, and F_2(x) = not x.

Let us now define the set D(I,T) of all elements of S which are **derivable** from the assumptions I via the transformations T. First of all, it is clear that I should be a subset of D(I,T). Let us call the elements of I the **depth-zero** elements of D(I,T). Next, what about elements of the form x = $F_i(A_1,...,A_m)$, for some i, where each $A_k=I_j$ for some j? Obviously, these elements are simple transformations of the assumptions; they should be elements of D(I,T) as well. Let us call these the **depth-one** elements of D(I,T). Similarly, one may define an element x of S to be a **depth-n** element of D(I,T) if $x=F_i(A_1,...,A_m)$, for some i, where each of the A_k is a depth-p element of D(I,T), for some p<n. Finally, D(I,T) may then be defined as the set of all x which are depth-n elements of D(I,T) for some n.

For example, if the T are rules of logic and the I are some propositions about the world, then D(I,T) is the set of all propositions which are logically equivalent to some subset of I. In this case deduction is a matter of finding the logical consequences of I, which are presumably a small subset of the total set S of all propositions. This is the general form of deduction. Boolean logic consists of a specific choice of T; andpredicate calculus consists of an addition onto the set T provided by Boolean logic.

It is worth noting that, in this approach to deduction, **truth** is inessential. In formal logic it is conventional to assume that one's assumptions are "true" and one's transformations are "truth-preserving." However, this is just an interpretation foisted on the deductive system after the fact.

3.5.2. Analogical Structure

The set (I,T) constructed above might be called a **transformation system**. It may be likened to a workshop. The initials I are the materials at hand, and the transformations T are the tools. D(I,T) is the set of all things that can be built, using the tools, from the materials.

What is lacking? First of all, blueprints. In order to apply a transformation system to real problem, one must have some idea of which transformations should be applied in which situations.

But if an **intelligence** is going to apply a transformation system, it will need to apply it in a variety of different contexts. It will not know exactly which contexts are going to arise in future. It cannot retain a stack of blueprints for every possible contingency. What it needs is not merely a stack of blueprints, but a **mechanism** for generating blueprints to fit situations.

But, of course, it already **has** such a mechanism -- its innate intelligence, its ability to induce, to reason by analogy, to search through its associative memory. What intelligence needs is a transformation system structured in such a way that ordinary mental processes can serve as its blueprint-generating machine.

In *SI* this sort of transformation system is called a "useful deductive system." Here, however, I am thinking more generally, and I will use the phrase **structured transformation system** instead. A structured transformation system is a transformation system with the property that, if a mind wants to make a "blueprint" telling it how to **construct** something from the initials using the transformations, it can often approximately do so by reasoning analogically with respect to the blueprints from **other** construction projects.

Another way to put it is: a **structured transformation system**, or **STS**, is transformation system with the property that the proximity between x and y in an ideal **structurally associative memory** is correlatedwith the similarity between the blueprint sets corresponding to x and y. A transformation system is structured if the analogically reasoning mind can use it, in practice, to construct things to order. This construction need not be infallible -- it is required only that it work approximately, much of the time.

3.5.2.1. (*) A Formal Definition

One formal definition goes as follows. Let x and y be two elements of D(I,T), and let $G_{I,T}(x)$ and $G_{I,T}(y)$ denote the set of all proofs in the system (I,T) of x and y respectively. Let U equal the **minimum** over all functions v of the sum a|v| + B, where B is the **average**, over all pairs (x,y) so that x and y are both in D(I,T), of the **correlation coefficient** between

$d_{\#}$[St(x **union** v)-St(v), St(y **union** v) - St(v)]

and

$d_*[G_{I,T}(x),G_{I,T}(y)]$.

Then (I,T) is **structured** to degree U.

Here $d_{\#}$(A,B) is the structural complexity of the symmetric difference of A and B. And d_* is a metric on the space of "set of blueprints," so that the $d_*[G_{I,T}(x),G_{I,T}(y)]$ denotes of the distance between the set of proofs of x and the set of proofs of y.

If the function v were omitted, then the degree of structuredness of U would be a measure of how true it is that structurally similar constructions have similar blueprint sets. But the inclusion of the function v broadens the definition. It need not be the case that similar x and y have similar blueprint sets. If x and y display similar **emergent patterns** on conjunction with some entity v, and x and y have similar blueprint sets, then this counts as structuredness too.

3.5.3. Transformation, Prediction and Deduction

What do STS's have to do with prediction? To make this connection, it suffices to interpret the **depth** index of an element of D(I,T) as a **time** index. In other words, one may assume that to apply each transformation in T takes some integer number of "time steps," and consider the construction of an element in D(I,T) as a process of actual temporal construction. This is a natural extension of the "materials, tools and blueprints" metaphor introduced above.

A simulation of some process, then, begins with an **initial condition** (an element of I) and proceeds to apply **dynamical rules** (elements of T), one after the other. In the case of a simple iteration like $x_k = cx_{k-1}(1-x_{k-1})$, the initial condition is an approximation of a real number, and there is only one transformation involved, namely the function f(x) = cx(1-x) or some approximation thereof. But in more complex simulations there may be a variety of different transformations.

For instance, a numerical iteration of the form $x_k = f(k,x_{k-1})$ rather than $x_k = f(x_{k-1})$ requires a different iteration at each time step. This is precisely the kind of iteration used to generate fractals by the iterated function system method (Barnsley, 1988). In this context, oddly enough, a random or chaotic choice of k leads to a **more** intricately structured trajectory than an orderly choice of k.

So, the process of simulating a dynamical system and the process of making a logical deduction are, on the broadest level, the same. They both involve transformation systems. But what about the **structured** part? What would it mean for a family of simulations to be executed according to a **structured** transformation system?

It would mean, quite simply, that the **class** of dynamical rule sequences that lead up to a situation is correlated with the **structure** of the situation. With logical deduction, one often knows what one wants to prove, and has to find out how to prove it -- so it is useful to know what worked to prove similar results. But with simulation, it is exactly the reverse. One often wants to know what the steps in one's transformation sequence will lead to, because one would like to avoid **running** the whole transformation sequence through, one step at a time. So it is useful to know what has **resulted** from running through **similar** transformation sequences. The same correlation is useful for simulation as for deduction -- but for a different reason.

Actually, this is an overstatement. Simulation makes **some** use of reasoning from similarity of results to similarity transformation sequences -- because one may be able to **guess** what the results of a certain transformation sequence will be, and then one will want to know what **similar** transformation sequences have led to, in order to assess the plausibility of one's guess. And deduction makes **some** use of reasoning from similarity of transformation sequences to similarity of results -- one may have an idea for a "proof strategy," and use analogical reasoning to make a guess at whether this strategy will lead to anything interesting. There is adistinction between the two processes, but it is not precisely drawn.

In conclusion, I propose that **most** psychological simulation and deduction is done by structured transformation systems. Some short simulations and deductions may be done without the aid of structure -- but this is the exception that proves the rule. **Long** chains of deductive transformations cannot randomly produce useful results. And long chains of dynamical iterations, if unmonitored by "common sense", are likely to produce errors -- this is true even of digital computer simulations, which are much more meticulous than any program the human brain has ever been known to run.

Psychologically, structured transformation systems are only effective if run in parallel. Running one transformation after another is very slow. **Some** simulations, and **some** logical deductions, will require this. But the mind will do its utmost to avoid it. One demonstration of this is the extreme difficulty of doing long mathematical proofs in one's head. Even the greatest mathematicians used pencil and paper, to record the details of the last five steps while they filled up their minds with the details of the next five.

Chapter Four

PSYCHOLOGY AND LOGIC

I have already talked a little about deduction and its role in the mind. In this chapter, however, I will develop this theme much more fully. The relation between psychology and logic is important, not only because of the central role of deductive logic in human thought, but also because it is a microcosm of the relation between **language and thought** in general. Logic is an example of a linguistic system, and it reveals certain phenomena that are obscured by the sheer complexity of other linguistic systems.

4.1. PSYCHOLOGISM AND LOGISM

Today, as John MacNamara has put it, "logicians and psychologists generally behave like the men and women in an orthodox synagogue. Each group knows about the other, but it is proper form that each should ignore the other" (1986, p.1). But such was not always the case. Until somewhere toward the end of nineteenth century, the two fields of logic and psychology were closely tied together. What changed things was, on the one hand, the emergence of experimental psychology; and, on the other hand, the rediscovery and development of elementary symbolic logic by Boole, deMorgan and others.

The early experimental psychologists purposely avoided explaining intelligence in terms of logic. Mental phenomena were analyzed in terms of images, associations, sensations, and so forth. And on the other hand -- notwithstanding the psychological pretensions of Leibniz's early logical investigations and Boole's *Laws of Thought* -- the early logicians moved further and further each decade toward considering logical operationsas distinct from psychological operations. It was increasingly realized on both sides that the formulas of propositional logic have little connection with emotional, intuitive, ordinary everyday thought.

Of course, no one denies that there is **some** relation between psychology and logic. After all, logical reasoning takes place within the mind. The question is whether mathematical logic is a very special kind of mental process, or whether, on the other hand, it is closely connected with everyday thought processes. And, beginning around a century ago, both logicians and psychologists have overwhelmingly voted for the former answer.

The almost complete dissociation of logic and psychology which one finds today may be partly understood as a reaction against the nineteenth-century doctrines of **psychologism** and **logism**. Both of these doctrines represent extreme views: logism states that psychology is a subset of logic; and psychologism states that logic is a subset of psychology.

Boole's attitude was explicitly logist -- he optimistically suggested that the algebraic equations of his logic corresponded to the structure of human thought. Leibniz, who anticipated many of Boole's discoveries by approximately two centuries, was ambitious beyond the point of logism as I have defined it here: he felt that elementary symbolic logic would ultimately explain not only the mind but the physical world. And logism was also not unknown among psychologists -- it was common, for example, among members of the early Wurzburg school of Denkpsychologie. These theorists felt that human judgements generally followed the forms of rudimentary mathematical logic.

But although logism played a significant part in history, the role of psychologism was by far the greater. Perhaps the most extreme psychologism was that of John Stuart Mill (1843), who in his *System of Logic* argued that

Logic is not a Science distinct from, and coordinate with, Psychology. So far as it is a Science at all, it is a part or branch of Psychology.... Its theoretic grounds are wholly borrowed from Psychology....

Mill understood the axioms of logic as "generalizations from experience." For instance, he gave the following psychological "demonstration" of the Law of ExcludedMiddle (which states that for any p, either p or not-p is always true):

The law on Excluded Middle, then, is simply a generalization of the universal experience that some mental states are destructive of other states. It formulates a certain absolutely constant law, that the appearance of any positive mode of consciousness cannot occur without excluding a correlative negative mode; and that the negative mode cannot occur without excluding the correlative positive mode.... Hence it follows that if consciousness is not in one of the two modes it much be in the other (bk. 2, chap.,7, sec.5)

Even if one accepted psychologism as a general principle, it is hard to see how one could take "demonstrations" of this nature seriously. Of course each "mode of consciousness" or state of mind excludes certain others, but there is no intuitively experienced **exact opposite** to each state of mind. The concept of logical negation is not a "generalization" of but rather a **specialization** and **falsification** of the common psychic experience which Mill describes. The leap from exclusion to exact opposition is far from obvious and was a major step in the development of mathematical logic.

As we will see a little later, Nietzsche (1888/1968) also attempted to trace the rules of logic to their psychological roots. But Nietzsche took a totally different approach: he viewed logic as a special system devised by man for certain purposes, rather than as something wholly deducible from inherent properties of mind. Mill was convinced that logic must follow automatically from "simpler" aspects of mentality, and this belief led him into psychological absurdities.

The early mathematical logicians, particularly Gottlob Frege, attacked Mill with a vengeance. For Frege (1884/1952) the key point was the question: what makes a sentence true? Mill, as an empiricist, believed that all knowledge must be derived from sensory experience. But Frege countered that "this account makes everything subjective, and if we follow it through to the end, does away with truth" (1959, p. vii). He proposed that truth must be given a non-psychological definition, one independent of the dynamics of any particular mind. This Fregean conception of truth received its fullestexpression in Tarski's (1935) and Montague's (1974) work on formal semantics, to be discussed in Chapter Five.

To someone acquainted with formal logic only in its recent manifestations, the very concept of psychologism is likely to seem absurd. But the truth is that, before the work of Boole, Frege, Peano, Russell and so forth transformed logic into an intensely mathematical discipline, the operations of logic **did** have direct psychological relevance. Aristotle's syllogisms made good psychological sense (although we now know that much useful human reasoning relies on inferences which Aristotle deemed incorrect). The simple propositional logic of Leibniz and Boole could be illustrated by means of psychological examples. But the whole development of modern mathematical logic was based on the introduction of patently non-psychological axioms and operations. Today few logicians give psychology a second thought, but for Frege it was a major conceptual battle to free mathematical logic from psychologism.

In sum, psychologists ignored those few voices which insisted on associating everyday mental processes with mathematical logic. And, on the other hand, logicians actively rebelled against the idea that the rules of mathematical logic must relate to rules of mental process. Psychology benefited from avoiding logism, and logic gained greatly from repudiating psychologism.

4.1.1. The Rebirth of Logism

But, of course, that wasn't the end of the story. Although contemporary **psychology** and **logic** have few direct relations with one another, in the century since Frege there has arisen a brand new discipline, one that attempts to bring psychology and logic closer together than they ever have been before. I am speaking, of course, about **artificial intelligence**.

Early AI theorists -- in the sixties and early seventies -- brought back logism with a vengeance. The techniques of early AI were little more than applied Boolean logic and tree search, with a pinch or two of predicate calculus, probability theory and other mathematical tricks thrown in for good measure. But every few years someone optimistically predicted that an intelligent computer was just around the corner. At this stage AI theorists basically ignored psychology -- they felt that deductive logic, and deductive logic alone, was sufficient for understanding mental process.

But by the eighties, AI was humbled by experience. Despite some incredible successes, nothing anywhere neara "thinking machine" has been produced. No longer are AI theorists too proud to look to psychology or even philosophy for assistance. Computer science still relies heavily on formal logic -- not only Boolean logic but more recent innovations such as model theory and non-well-founded sets (Aczel, 1988) -- and AI is no exception. But more and more AI theorists are wondering now if modern logic is adequate for their needs. Many, dissatisfied with logism, are seeking to modify and augment mathematical logic in ways that bring it closer to human reasoning processes. In essence, they are augmenting their vehement logism with small quantities of the psychologism which Frege so abhorred.

4.1.2. The Rebirth of Psychologism

This return to a limited psychologism is at the root of a host of recent developments in several different areas of theoretical AI. Perhaps the best example is nonmonotonic logic, which has received a surprising amount of attention in recent years. But let us dwell, instead, on an area of research with more direct relevance to the present book: automated theorem proving.

Automatic theorem proving -- the science of programming computers to prove mathematical theorems -- was once thought of as a stronghold of pure deductive logic. It seemed so simple: just apply the rules of mathematical logic to the axioms, and you generate theorems. But now many researchers in automated theorem proving have realized that this is only a very small part of what mathematicians do when they prove theorems. Even in this ethereal realm of reasoning, tailor-made for logical deduction, nondeductive, alogical processes are of equal importance.

For example, after many years of productive research on automated theorem proving, Alan Bundy (1991) has come to the conclusion that

> Logic is not enough to understand reasoning. It provides only a low-level, step by step understanding, whereas a high-level, strategic understanding is also required. (p. 178)

Bundy proposes that one can program a computer to demonstrate high-level understanding of mathematical proofs, by supplying it with the ability to manipulate entities called **proof plans**.

A proof plan is defined as a common structure that underlies and helps to generate many differentmathematical proofs. Proof plans are not formulated based on mathematical logic alone, they are rather

> refined to improve their expectancy, generality, prescriptiveness, simplicity, efficiency and parsimony while retaining their correctness. Scientific judgement is used to find a balance between these sometimes opposing criteria. (p.197)

In other words, proof plans, which control and are directed by deductive theorem-proving, are constructed and refined by illogical or alogical means.

Bundy's research programme -- to create a formal, computational theory of proof plans -- is about as blatant as pychologism gets. In fact, Bundy admits that he has ceased to think of himself as a researcher in automated theorem proving, and come to conceive of himself as a sort of abstract psychologist:

> For many years I have regarded myself as a researcher in automatic theorem proving. However, by analyzing the methodology I have pursued in practice, I now realize that my real motivation is the building of a science of reasoning.... Our science of reasoning is normative, empirical and reflective. In these respects it resembles other human sciences like linguistics and Logic. Indeed it includes parts of Logic as a sub-science. (p. 197)

How similar this is, on the surface at least, to Mill's "Logic is ... a part or branch of Psychology"! But the difference, on a deeper level, is quite large. Bundy takes what I would call a Nietzschean rather than a Millean approach. He is not deriving the laws of logic from deeper psychological laws, but rather studying how the powerful, specialized reasoning tool that we call "deductive logic" fits into the general pattern of human reasoning.

4.2. LIMITED BOOLEAN LOGISM

Bundy defends what I would call a "limited Boolean logism." He maintains that Boolean logic and related deductive methods are an important part of mental process, but that they are supplemented by and continually affected by other mental processes. At first sight, this perspective seems completely unproblematic. We think logically when we need to, alogically when we need to; and sometimes the two modes of cognition will interact. Very sensible.

But, as everyone who has taken a semester of university logic is well aware, things are not so simple. Even limited Boolean logism has its troubles. I am speaking about the simple conceptual conundrums of Boolean logic, such as Hempel's paradox of confirmation and the paradoxes of implication. These elementary "paradoxes," though so simple that one could explain them to a child, are obstacles that stand in the way of even the most unambitious Boolean logism. They cast doubt as to whether Boolean logic can ever be of any psychological relevance whatsoever.

4.2.1. Boolean Logic and Modern Logic

One might well wonder, why all this emphasis on **Boolean** logic. After all, from the logician's point of view, Boolean logic -- the logic of "and", "or" and "not" -- is more than a bit out-of-date. It does not even include quantification, which was invented by Peirce before the turn of the century. Computer circuits are based entirely on Boolean logic; however, modern mathematical logic has progressed as far beyond Leibniz, Boole and deMorgan as modern biology has progressed beyond Cuvier, von Baer and Darwin.

But still, it is not as though modern logical systems have **shed** Boolean logic. In one way or another, they are invariably based on Boolean ideas. Mathematically, nearly all logical systems are "Boolean algebras" -- in addition to possessing other, subtler structures. And, until very recently, one would have been hard put to name a logistic model of human reasoning that did not depend on Boolean logic in a very direct way. I have already mentioned two exceptions, nonmonotonic logic and proof plans, but these are recent innovations and still in very early stages of development.

So the paradoxes of Boolean logic are paradoxes of modern mathematical logic in general. They are the most powerful weapon in the arsenal of the contemporary anti-logist. Therefore, the most sensible way to begin our quest to synthesize psychology and logic is to dispense with these paradoxes.

Paradoxes of this nature cannot be "solved." They are too simple for that, too devastatingly fundamental. So my aim here is not to "solve" them, but rather to demonstrate that they are largely **irrelevant** to theproject of limited Boolean logism -- if this project is carried out in the proper way. This demonstration is less logical than psychological. I will assume that the mind works by pattern recognition and multilevel optimization, and show that in this context Boolean logic can control mental processes without succumbing to the troubles predicted by the paradoxes.

4.2.2. The Paradoxes of Boolean Logic

Before going any further, let us be more precise about exactly what these "obstacles" are. I will deal with four classic "paradoxes" of Boolean logic:

1. The first paradox of implication. According to the standard definition of implication one has "a --> (b --> a)" for all a and b. Every true statement is implied by anything whatsoever. For instance, the statement that the moon is made of green cheese implies the statement that one plus one equals two. The statement that Lassie is a dog implies the statement that Ione Skye is an actress. This "paradox" follows naturally from the elegant classical definition of "a --> b" as "either b, or else not a". But it renders the concept of implication inadequate for many purposes.

2. The second paradox of implication. For all a and c, one has "not-c --> (c --> a)". That is, if c is false, then c implies anything whatsoever. From the statement that George Bush has red hair, it follows that psychokinesis is real.

3. Contradiction sensitivity. In the second paradox of implication, set c equal to the conjunction of some proposition and its opposite. Then one has the theorem that, if "A and not-A" is true for any A, **everything else is also true**. This means that Boolean logic is incapable of dealing with sets of data that contain even one contradiction. For instance, assume that "I love my mother", and "I do not love my mother" are both true. Then one may prove that 2+2=5. For surely "I love my mother" implies "I love my mother **or** 2+2=5" (in general, "a --> (a or b)). But, just as surely, "I do not love my mother" and "I love my mother **or** 2+2=5", taken together, imply "2+2=5" (in general, [a and (not-a or b)] --> b). Boolean logic is a model of reasoning in which ambivalence about one's feelings for one's mother leads naturally to the conclusion that 2+2=5.

4. Hempel's confirmation paradox. According to Boolean logic, "all ravens are black" is equivalent to "all nonblack entities are nonravens". That is,schematically, "(raven --> black) --> (not-black --> not-raven)". This is a straightforward consequence of the standard definition of implication. But is it not the case that, if A and B are equivalent hypotheses, evidence in favor of B is evidence in favor of A. It follows that every observation of something which is not black and also not a raven is evidence that ravens are black. This is patently absurd.

4.2.3. The Need for New Fundamental Notions

The standard method for dealing with these paradoxes has to acknowledge them, then dismiss them as irrelevant. In recent years, however, this evasive tactic has grown less common. There have been several attempts to modify standard Boolean-based formal logic in such a way as to avoid these difficulties: relevant logics (Read, 1988), paraconsistent logics (daCosta, 1984), and so forth.

Some of this work is of very high quality. But in a deeper conceptual sense, none of it is really satisfactory. It is, unfortunately, not concrete enough to satisfy even the most logistically inclined psychologist. There is a tremendous difference between a convoluted, abstract system jury-rigged specifically to avoid certain formal problems, and a system with a simple intuitive logic behind it.

An interesting commentary on this issue is provided by the following dialogue, reported by Gian-Carlo Rota (1985). The great mathematician Stanislaw Ulam was preaching to Rota about the importance of subjectivity and context in understanding meaning. Rota begged to differ (at least partly in jest):

"But if what you say is right, what becomes of objectivity, an idea that is so definitively formulated by mathematical logic and the theory of sets, on which you yourself have worked for many years of your youth?"

Ulam answered with "visible emotion":

"Really? What makes you think that mathematical logic corresponds to the way we think? You are suffering from what the French call a **deformation professionelle**. ..."

"Do you then propose that we give up mathematical logic?" said I, in fake amazement.

"Quite the opposite. Logic formalizes only very few of the processes by which we actuallythink. The time has come to enrich formal logic by adding to it some other fundamental notions. ... Do not lose your faith," concluded Stan. "A mighty fortress is mathematics. It will rise to the challenge. It always has."

Ulam speaks of enriching formal logic "by adding to it some other fundamental notions." More specifically, I suggest that we must enrich formal logic by adding to it the fundamental notions of **pattern** and **multilevel control**, as discussed above. The remainder of this chapter is devoted to explaining how, if one views logic in the context of pattern and multilevel control, all four of the "paradoxes" listed above are either resolved or avoided.

This explanation clears the path for a certain form of limited Boolean logism -- a Boolean logism that assigns at least a co-starring role to pattern and multilevel control. And indeed, in the chapters to follow I will develop such a form of Boolean limited logism, by extending the analysis of logic given in this chapter to more complex psychological systems: language and belief systems.

4.3. THE PARADOXES OF IMPLICATION

Let us begin with the first paradox of implication. How is it that a true statement is implied by everything?

This is not our intuitive notion of **consequence**. Suppose one mental process has a dozen subsidiary mental processes, supplies them all withstatement A, and asks each of them to tell it **what follows from A**. What if one of these subsidiary processes responds by outputting true statements at random? Justified, according to Boolean logic -- but **useless**! The process should not survive. What the controlling process needs to know is what one can **use** statement A for -- to know what follows from statement A in the sense that statement A is an integral part of its demonstration.

This is a new interpretation of "implies." In this view, "A implies B" does not mean simply "-B + A", it means that A is an integral part of a natural reasoning process leading towards B. It means that A is helpful in arriving at B. Intuitively, it means that, when one sees that someone has arrived at the conclusion B, it is plausible to assume that they arrived at A **first** andproceeded to B from there. If one looks at implication this way -- structurally, algorithmically, informationally -- then the paradoxes are gone.

In other words, according to the informational definition, A significantly implies B if it is sensible to use A to **get** B. The mathematical properties of this definition have yet to be thoroughly explored. However, it is clear that a true statement is no longer significantly implied by **everything**: the first paradox of implication is gone.

And the second paradox of implication has also disappeared. A false statement no longer implies everything, because the generic proof of B from "A and not-A" makes **no** essential use of A; A could be replaced by anything whatsoever.

4.3.1. Informational Implication (*)

In common argument, when one says that one thing implies another, one means that, by a series of logical reasonings, one can obtain the second thing from the first. But one does **not** mean to include series of logical reasonings which make only inessential use of the first thing. One means that, using the first thing in some substantial way, one may obtain the second through logical reasoning. The question is, then, what does **use** mean?

If one considers only formulas involving --> (implication) and - (negation), it is possible to say something interesting about this in a purely formal way. Let $B_1,...,B_n$ be a proof of B in the deductive system T **union** {A}, where T is some theory. Then, one might define A to be **used** in deriving B_i if either

1) B_i is identical with A, or

2) B_i is obtained, through an application of one of the rules of inference, from B_j's with j<i, and A is used for deriving at least one of these B_j's.

But this simplistic approach becomes hopelessly confused when disjunction or conjunction enters into the picture. And even in this uselessly simple case, it has certain conceptual shortcomings. What if there is a virtually identical proof of A which makes no use of A? Then is it not reasonable to say that the supposed "use" of A is largely, though not entirely, spurious?

It is not inconceivable that a reasonable approximation of the concept of use might be captured by some complex manipulation of connectives. However, I contend that what use **really** has to do with is **structure**. Talking about structure is not so cut-and-dried astalking about logical form -- one always has a lot of loose parameters. But it makes much more intuitive sense.

Let $G_{I,T,v}(B)$ denote the set of all valid proofs of B, relative to some fixed "deductive system" (I,T), of complexity less than v. An element of $G_{I,T,v}$ is a sequence of steps $B_0,B_1,...,B_{n-1}$, where B_n=B, and for k>0 B_k follows from B_{k-1} by one of the transformation rules T. Where Z is an element of $G_{I,T,v}(B)$, let L(Z) = |B|/|Z|. This is a measure of how much it simplifies B to prove it via Z.

Where $G_{I,T,v}(B) = \{Z_1,...,Z_N\}$, and p is a positive integer, let

$A = L(Z_1)*[I(Z_1|Y)]1/p + L(Z_2)*[I(Z_2|Y)]1/p + ... + L(Z_N)*[I(Z_N|Y)]1/p$

$B = I(Z_1|Y)]1/p + I(Z_2|Y)]1/p + ... + [I(Z_N|Y)]1/p$

$Q_{p,v} = A/B$

Note that, since $I(Z_i|Y)$ is always a positive integer, as p tends to infinity, $Q_{p,v}$ tends toward the value L(Z)*I(Z|Y), where Z is the element of $G_{I,T,v}$ that minimizes I(Z|N). The smaller p is, the more fairly the value L(Z) corresponding to **every** element of $G_{I,T,v}$ is counted. The larger p is, the more attention is focused on those proofs that are informationally close to Y. The idea is that those proofs which are closer to Y should count much more than those which are not.

Definition: Let | | be a complexity measure (i.e., a nonnegative-real-valued function). Let (I,T) be a deductive system, let p be a positive integer, and let 0<c<1. Then, relative to | |, (I,T), p and c, we will say **A significantly implies B to degree K**, and write

$A \text{-->}_K B$

if K = cL+(1-c)M is the largest of all numbers such that for some v there exists an element Y of $G_{I,T,v}$ so that

1) $A=B_0$ (in the sequence of deductions described by Y)

2) L = L(Y) = |B|/[|Y|],

3) $M = 1/Q_{p,|Y|}$

According to this definition, A significantly implies B to a high degree if and only if B is an integral part of a "natural" proof of A. The "naturalness" of the proof Y is guaranteed by clause (3), which says that by modifying Y a little bit, it is not so easy to get a simpler proof. Roughly, clause (3) says that Y is an "approximate local minimum" of simplicity, in proof space.

This is the kind of implication that is useful in building up a belief system. For, under ordinaryimplication there can never be any sense in assuming that, since A --> B_i, i=1,2,...,N, and the B_i are true, A might be worth assuming. After all, by contradiction sensitivity a false statement implies everything. But things are not so simple under **relevant** implication. If a statement A significantly **implies** a number of true statements, that means that by appending the statement A to one's assumption set I, one can obtain quality proofs of a number of true statements. If these true statements also happen to be useful, then **from a practical point of view** it may be advisable to append A to I. Deductively such a move is not justified, but **inductively** it is justified. This fits in with the general analysis of deduction given in *SI*, according to which deduction is useful only insofar as induction justifies it.

4.4. CONTRADICTION SENSITIVITY

Having dealt with implication, let us now turn to the paradox of contradiction sensitivity. According to reasoning given above, if one uses propositional or predicate calculus to define the transformation system T, one easily arrives at the following conclusion: if any two of the propositions in I contradict each other, then D(I,T) is the entire set of all propositions. From one contradiction, everything is derivable.

This property appears not to reflect actual human reasoning. A person may contradict herself regarding abortion rights or the honesty of her husband or the ultimate meaning of life. And yet, when she thinks about theoretical physics or parking her car, she may reason deductively to one particular conclusion, finding any contradictory conclusion ridiculous.

In his Ph.D. dissertation, daCosta (1984) conceived the idea of a **paraconsistent** logic, one in which a single contradiction in I does **not** imply everything. Others have extended this idea in various ways. More recently, Avram (1990) has constructed a paraconsistent logic which incorporates the idea of "relevance logic." Propositions are divided into classes and the inference from A to A+B is allowed only when A and B are in the same class. The idea is very simple: according to Avram, although we do use the "contradiction-sensitive" deductive system of standard mathematical logic, we carefully distinguish deductions in one sphere from deductions in another, so that we never, in practice, reason "A implies A orB", unless A and B are in the same "sphere" or "category."

For instance, one might have one class for statements about physics, one for statements about women, et cetera. The formation of A or B is allowed only if A and B belong to the same class. A contradiction regarding one of these classes can therefore destroy only reasoning **within that class**. So if one contradicted oneself when thinking about one's relations with one's wife, then this might give one the ability to deduce **any** statement whatsoever about domestic relations -- but not about physics or car parking or philosophy.

The problem with this approach is its arbitrariness: why not one class for particle physics, one for gravitation, one for solid-state physics, one for brunettes, one for blondes, one for redheads,.... Why not, following Lakoff's (1987) famous analysis of aboriginal classification systems, one category for women, fire and dangerous things?

Of course, it is true that we rarely make statements like "either the Einstein equation has a unique solution under these initial-boundary conditions or that pretty redhead doesn't want anything more to do with me." But still, partitioning is too rigid -- it's not quite right. It yields an elegant formal system, but of course in any categorization there will be borderline cases, and it is unacceptable to simply ignore these away.

The "partitioning" approach is not the only way of defining relevance formally. But it seems to be the only definition with any psychological meaning. Read (1988), for instance, disavows partitioning. But he has nothing of any practical use to put in its place. He mentions the classical notion of variable sharing -- A and B are mutually relevant if they have variables in common. But he admits that this concept is inadequate: for instance, "A" and "-A + B" will in general share variables, but one wishes to forbid their combination in a single expression. He concludes by defining **entailment** in such a way that

[T]he test of whether two propositions are logically relevant is whether either entails the other. Hence, relevance cannot be picked out prior ... to establishing validity or entailment....

But the obvious problem is, this is not really a definition of relevance:

It may of course be objected that this suggested explication of relevance is entirely circular andunilluminating, since it amounts to saying no more than that two propositions are logically relevant if either entails the other....

Read's account of relevance is blatantly circular. Although it is not unilluminating from the formal-logical point of view; it is of no **psychological** value.

4.4.1 Contradiction and the Structure of Mind

There is an an alternate approach: to define relevance not by a partition into classes but rather in terms the theory of structure. It is hypothesized that a mind does not tend to form the disjunction A or B unless the size

%[(St(A **union** v)-St(v)]-[St(B **union** w)-St(w)]%

is small for some (v,w), i.e. unless A and B are in some way **closely related**. In terms of the structurally associative memory model, an entity A will generally be stored **near** those entities to which it is closely related, and it will tend to interact mainly with these entities.

As to the possibility that, by chance, two completely unrelated entities will be combined in some formula, say A or B, it is admitted that this could conceivably pose a danger to thought processes. But the overall structure of mind dictates that a part of the mind which succumbed to self-contradiction and the resulting inefficiency, would soon be ignored and dismantled.

According to the model of mind outlined above, each mental process supervises a number -- say a dozen -- of others. Suppose these dozen are reasoning deductively, and one of them falls prey to an internal self-contradiction, and begins giving out random statements. Then how efficient will that self-contradicting process be? It will be the least efficient of all, and it will shortly be eliminated and replaced. Mind does not work by absolute guarantees, but rather by probabilities, safeguards, redundancy and natural selection.

4.4.2. Contradiction and Implication

We have given one way of explaining why contradiction sensitivity need not be a problem foractual minds. But, as an afterthought, it is worth briefly noting that one may also approach the problem from the point of view of relevant implication. The step from " A and not-A" to B involves the step "not-A --> A or B". What does our definition of significant implication say about this? A moment's reflection reveals that, as noted above, clause (3) kicks in here: A is totally indispensible to this proof of B; A could just as well be replaced by C, D, E or any other proposition. The type of implication involved in contradiction sensitivity is not significant to a very high degree.

4.5. CONFIRMATION

Finally, what of Hempel's confirmation paradox? Why, although "all ravens are black" is equivalent to "all non-black entities are non-ravens," is an observation of a blue chair a lousy piece of evidence for "all ravens are black"?

My resolution is simple, and not conceptually original. Recall the "infon" notation introduced in Section 2. Just because s |-- i //x to degree d, it is not necessarily the case that s |-- j //x to degree d for every j equivalent to i under the rules of Boolean logic. This is, basically, all that needs to be said. Case closed, end of story. Boolean logic is a tool. Only in certain cases does the mind find it useful.

That the Boolean equivalence of i and j does not imply the equality of d(s,i,x) and d(s,j,x) is apparent from the definition of **degree** given above. The degree to which (s,k,x) holds was defined in terms of the intensity with which the elements of k are patterns in s, where complexity is defined by s. Just because i and j are Booleanly equivalent, this does not imply that they will have equal **algorithmic information content**, equal **structure**, equal complexity with respect to some observer s. Setting things up in terms of pattern, one obtains a framework for studying reasoning in which Hempel's paradox does not exist.

3.5.1 A More Psychological View

In case this seems too glib, let us explore the matter from a more psychological perspective. Assume that "All ravens are black" happens to hold with degree d, in my experience, from my perspective. Then to whatdegree does "All non-black entities are non-ravens" hold in my experience, from my perspective?

"All ravens are black" is an **aid** in understanding the nature of the world. It is an aid in identifying ravens. It is a **significant pattern** in my world that those things which are typically referred to with the label "raven," are typically possessors of the color black. When storing in my memory a set of experiences with ravens, I do not have to store with each experience the fact that the raven in question was black -- I just have to store, once, the statement that all ravens are black, and then **connect** this in my memory to the various experiences with ravens.

Now, what about "All non-black entities are non-ravens"? What good does it do me to recognize this? How does it simplify my store of memories? It does not, not hardly at all. When I call up a non-black entity from my memory, I will not need to be reminded that it is not a raven. Why would I have thought that it was a raven in the first place? "Raven-ness?" is not one of the questions which it is generally useful or interesting to ask about entities, whereas on the other hand "color?" **is** one of the questions which it is often interesting to ask about physical objects such as birds.

So, the real question with Hempel's paradox is, what determines the degree assigned to a given proposition s |-- i //x. It is not purely the **logical form** of the proposition, but rather the degree to which the proposition is useful to x, i.e. the emergence between the proposition and the other entities which neighbor it in the memory of x. Degree is determined by psychological dynamics, rather than Boolean logic. Formally, one may say: the logic of memory organization is what determines the subjective complexity measure associated with x.

It is not **always** necessary to worry about where the degrees associated with propositions come from. But when one is confronted with a paradox regarding degrees, then it **is** necessary to worry about it. The real moral of Hempel's paradox, as I see it, is that one should study confirmation in terms of the structure and dynamics of the mind doing the confirming. Studying confirmation otherwise, "in the abstract," borders on meaningless.

In Hempel's paradox one is once again confronted with "what follows what." Boolean logic says that one's belief in "all ravens are black" should be increased following observation of a blue chair. But in fact, observing a blue chair does and should **not** lead to an increase in one's belief in "all ravens are black." Hempel's paradox is a sort of quantitative version of the paradox of implication -- instead of logic saying that B follows from A when it doesn't, one has logic saying that an increase in belief of B follows from an increase in belief in A **when it doesn't**.

4.6. A NIETZSCHEAN VIEW OF LOGIC

At about the same time that Frege, Peano and the rest were laying the foundations of modern mathematical logic, Friedrich Nietszche was creating his own brilliantly ideosyncratic view of the world. This world-view was obscure during Nietzsche's lifetime but, as he predicted, it turned out to be enormously influential throughout the twentieth century.

While the developments of the preceding sections lie squarely within the tradition begun by Frege and Peano, they also fit nicely into the context of Nietszche's thought. In this section I will take a brief detour from our formal considerations, to explore this observation. In Chapter Ten -- after dealing with belief and language -- I will return to Nietzsche's thought, to help us understand the relation between logic, language, consciousness, reality and belief.

4.6.1. The Will to Power

Nietzsche declared consciousness irrelevant and free will illusory. He proposed that hidden structures and processes control virtually everything we feel and do. Although this is a commonplace observation now, at the time it was a radical hypothesis. Nietszche made the first sustained effort to determine the nature of what we now call "the unconscious mind." The unconscious, he suggested, is made up of nothing more or less than "morphology and the will to power." The study of human feelings and behavior is, in Nietszche's view, the study of the various forms of the will to power.

From the start, Nietszche was systematically antisystematic; he would have ridiculed anyone who suggested making a chart of all the possible forms of the will to power. Instead, he concentrated on applying his idea to a variety of phenomena. In *Human, All Too Human* he analyzed hundreds of different human activities in terms of greed, lust, envy and other simple manifestations of the will to power. Substantial parts of *The Genealogy of Morals*, *Beyond Good and Evil*, and *The Twilight of the Idols* were devoted to studying ascetics,philosophers, and other personality types in a similar way. Two entire books -- *The Case of Wagner* and *Nietszche contra*

Wagner -- were devoted to the personality, music and philosophy of Richard Wagner. *The Antichrist* attempted a psychoanalysis of Jesus. And in *Ecce Homo,* he took on perhaps his most difficult subject: himself.

Nietszche was anything but objective. In fact his writings often appear delusional. His most famous book, *Thus Spake Zarathustra*, is written in a bizarrely grandiose mock-Biblical style. And *Ecce Homo* contains chapter titles such as "Why I Am So Wise", "Why I Am So Clever", and "Why I am a Destiny", as well as a lengthy description of his diet. But Nietszche did not mind appearing crazy. He did not believe in an **objective** logic, and he repeatedly stressed that what he wrote down were only **his** personal truths. He encouraged his readers to discover their **own** truths.

He did not, however, believe that everyone's personal truth was equally valuable. According to Nietszche, only a person with the strength to contradict himself continually and ruthlessly can ever arrive at significant insights. A person lacking this strength can only repeat the illusions that make him feel powerful, the illusions that enhance the power of the society which formed him. A person possessing this strength possesses power over himself, and can therefore grope beyond illusion and make a personal truth which is genuinely his own.

4.6.2. Nietzsche on Logic

Logic, according to Nietszche, is simply one particularly fancy manifestation of the will to power. At the core of mathematics and logic is the "will to make things equal" -- the collection of various phenomena into classes, and the assumption that all the phenomena in each class are essentially the same. Nietszche saw this as a lie. It is a necessary lie, because without it generalization and therefore intelligence is impossible. As Nietszche put it in his notebooks [1968a, p. 277],

> **[T]he will to equality is the will to power**.. the consequence of a will that as much as possible **shall be** equal.
>
> Logic is bound to the condition: assume there are identical cases. In fact, to make possible logical thinking and inferences, this conditionmust first be treated fictitiously as fulfilled....
>
> The inventive force that invented categories labored in the service of our needs, namely of our need for security, for quick understanding on the basis of signs and sounds, for means of abbreviation....

So logic is a lie, but a necessary one. It is also a lie which tends to make itself subjectively true: when an intelligence repeatedly assumes that a group of phenomena are the same for purposes of calculation, it eventually comes to believe the phenomena **really** are identical. To quote Nietszche's notebooks again (1968a, p. 275):

> It cannot be doubted that all sense-perceptions are permeated with value judgements.... First **images**..... Then **words**, applied to images. Finally **concepts**, possible only when there are words -- the collecting together of many images in something nonvisible but audible (word). The tiny

amount of emotion to which the "word" gives rise, as we contemplate similar images for which **one** word exists -- this weak emotion is the common element, the basis of the concept. That weak sensations are regarded as alike, sensed **as being the same**, is the fundamental fact. Thus confusion of two sensations that are close neighbors, as we take note of these sensations.... Believing is the primal beginning even in every sense impression....

The **valuation** "I believe that this and that is so" is the essence of truth. In valuations are expressed conditions of preservation and growth. All our organs of knowledge and our senses are developed only with regard to conditions of preservations and growth. Trust in reason and its categories, and dialectic, therefore the valuation of logic, proves only their usefulness for life, proved by experience -- **not** that something is true.

That a great deal of **belief** must be present; that judgements may be ventured; that doubt concerning all essential values is **lacking** -- that is the precondition of every living thing and its life. Therefore, what is needed is that somethingmust be held to be true -- **not** that something **is** true.

"The **real** and the **apparent** world" -- I have traced this antithesis back to **value** relations. We have projected the conditions of **our** preservation as predicates of being in general. Because we have to be stable in our beliefs if we are to prosper, we have made the "real" world a world not of change and becoming, but one of being.

This is what Nietzsche meant when he wrote "there are no facts, only interpretations." A fact is an interpretation which someone has used so often that they have come to depend upon it emotionally and cannot bear to conceive that it might not reflect a "true" reality. As an example of this, he cited the Aristotelian law of contradiction, which states that "A and not-A" is always false, no matter what A is:

We are unable to affirm and to deny one and the same thing: this is a subjective empirical law, not the expression of any 'necessity' but only of an inability.

If, according to Aristotle, the law of contradiction is the most certain of all principles, if it is the ultimate and most basic, upon which every demonstrative proof rests, if the principle of every axiom lies in it; then one should consider all the more rigorously what **presuppositions** already lie at the bottom of it. Either it asserts something about actuality, about being, as if one already knew this from another source; that is, as if opposite attributes **could** not be ascribed to it. Or the proposition means: opposite attributes **should** not be ascribed to it. In that case, logic would be an imperative, not to know the true, but to posit and arrange a world that shall be called true by us.

Note how different this is from Mill's shallow psychologism. In the Introduction I quoted Mill's "derivation" of the Law of Excluded Middle (which is equivalent to the law of contradiction, by an application of deMorgan's identities). Mill sought to **justify** this and other rules of logic by appeal to psychological principles. In Mill's view, the truth of "A or not-A" follows from the fact that each idea has a "negative idea," and whenever an idea is not present, its negative is. This is a very weak argument. One could make a stronger psychological argument for the falsity of "A and not-A" -- namely, one could argue that the mind **cannot**simultaneously entertain two contradictory ideas. But Nietzsche's point is that even this more plausible argument is **false**. As we all know from personal experience, the human mind **can** entertain two contradictory ideas at once. We may try to avoid this state of mind, but it has a habit of coming up over and over again: "I love her/ I don't love her", "I want to study for this test/ I want to listen to the radio instead". The rule of non-contradiction is not, as Mill would have it, correct because it reflects the laws of mental process -- it is, rather, something cleverly **conceived** by human minds, in order to provide for more effective functioning in certain circumstances.

One rather simplistic and stilted way of phrasing Nietszche's view of the world is as follows: **intelligence is impossible without a priori assumptions and rough approximation algorithms, so each intelligent system (each culture, each species) settles on those assumptions and approximations that appear serve its goals best, and accepts them as "true" for the sake of getting on with life**. Logic is simply one of these approximations, based on the false assumption of equality of different entities, and many auxiliary assumptions as well.

This is not all that different from Saint Augustine's maxim "I believe, so that I may understand." Augustine -- like Leibniz, Nietzsche and the existentialists after him and like the Buddhists and Sophists before him -- realized that thought cannot proceed without assuming some dogmatic presupposition as a foundation. But the difference in **attitude** between Augustine and Nietzsche is striking. Augustine wants you to believe in exactly what he does, so that you will understand things the same way he does. Nietzsche, on the other hand, wants you to believe and not believe at the same time; he wants you to **assume** certain approximations, to commit yourself to them, while at the same time continually realizing their tentative nature.

So, what does all this have to do with the mathematical ideas of the preceding sections? Nietzsche saw a universal form underlying the various possible forms of logic -- the will to power. I do not disagree with this diagnosis, but I feel that it is too abstract. The structural logic described above is NIetzschean in spirit, but it is more detailed than anything Nietszche ever said about logic: it makes **explicit** the dependence of logical reasoning processes on the biases, experiences and abilities of the mind that is doing the reasoning. It tries to **capture** this dependence in a precise,mathematical way. The "a priori assumptions and rough approximation algorithms" come into play in the process of **pattern recognition**, of **complexity evaluation**.

Logic is not a corollary of other psychological functions, it is a **special** psychological function of relatively recent invention, one with its own strengths, weaknesses and peculiarities. But it has neither meaning or utility outside of the context of the mind which maintains it and which it helps to maintain. This was Nietzsche's view of logic, and it fits in rather well with the more formal explorations given above.

Chapter Five

LINGUISTIC SYSTEMS

Alfred Tarski, who pioneered the mathematical semantics of **formal** languages (Tarski, 1935), adamantly maintained the impossibility of a mathematical semantics of **natural** language. But nevertheless, his work spawned a minor intellectual industry in the analysis of natural languages using formal semantics. In this chapter I will add a new twist to this research programme -- I will give a mathematical analysis of language and meaning from a **pattern-theoretic** rather than formal-logical angle, with an emphasis on the fundamentally **systemic** nature of language.

The idea that language has to do with pattern is not a new one. It was present in the structuralism of Ferdinand de Saussure. And, more pertinently, it played a central role in the controversial thought of Benjamin Lee Whorf. Although a great deal of attention has been paid to Whorf's linguistic relativity hypothesis (Whorf, 1956), very little has been written about the general philosophy of language underlying his work in comparative grammar and semantics. All of Whorf's thought was grounded in a conviction that language is, in some sense, **made** of pattern and structure:

> Because of the systematic, configurative nature of higher mind, the "patternment" aspect of language always overrides and controls the "lexation" or name-giving aspect.... We are all mistaken in our common belief that any word has an "exact meaning.... [T]he higher mind deals in symbols that have no fixed reference to anything, but are like blank checks, to be filled in as required, that stand for "any value" of a given variable, like ... the x, y, z of algebra....
>
> We should not however make the mistake of thinking that words, even as used by the lower personal mind, represent the opposite pole fromthese variable symbols.... Even the lower mind has caught something of the algebraic nature of language; so that words are in between the variable symbols of pure patternment ... and true fixed quantities. The sentence "I went all the way down there just in order to see Jack" contains only one fixed concrete reference; namely, "Jack." The rest is pattern attached to nothing specifically....

According to Whorf, a language consists of **patterns** which interact according to certain rules, which can somehow take one another as arguments, and which only occasionally make direct "reference" to "real," external objects.

In this chapter I will elaborate on this powerful insight, using the concepts developed in Chapters Two and Three. One result of this exploration will be a general model of language as a special kind of **structured transformation system**. Syntactic rules form a transformation system, and semantics determines the analogical structure of this system. This view of language will allow us to explore the relation between language and thought in a much clearer light than has previously been available. It will aid us in understanding how language relates with deduction, consciousness and belief, and how language aids in the development and maintenance of those constructions which we call **self** and **external reality**.

5.1. SYNTACTIC SYSTEMS

Richard Montague was the first person to make a full-scale effort to prove Tarski wrong, by applying the abstract semantic ideas of mathematical logic to natural languages. Due to his pioneering papers, and the subsequent work of Partee (1975) and others, we can now analyze the semantics of particular sentences in terms of formal logic. This is no small accomplishment. In fact, at the present time, no other theory of semantics, mathematical or no, can boast as effective a track record.

So, in this section, I will begin the discussion of language by reviewing the main points of Montague **grammar** -- the syntactical theory that underlies Montague semantics. Then I will move to the more general notion of **syntactic system**, which will lead toward a deeper understanding of linguistic dynamics.

5.1.1. Montague Syntax

Natural languages are frequently ambiguous: one word or phrase can have more than one meaning. This creates problems for mathematical logic; therefore Montague chose to deal only with **disambiguated** languages. Within the context of the formal approach, this is not a restriction but rather a methodological choice: any formal language can be mapped into a corresponding disambiguated formal language, by one of a number of simple procedures.

For instance, in the "language" of vector algebra, ixjxk is ambiguous, and to disambiguate it one must introduce parentheses, obtaining (ixj)xk, and ix(jxk). One way to disambiguate an English sentence is to draw an "analysis tree" for each interpretation of the sentence, and take these trees to be the elements of the disambiguated language. This is awkward, yes, but it is not a formal obstacle.

So, according to Montague, a **disambiguated language** consists of:

1) a set of syntactic operations, each of which maps sequences of syntactic expressions into single syntactic expressions,

2) a set of syntactic categories, which contain all possible words,

3) syntactic rules, telling which operations may be applied to words in which categories.

The disambiguity of the language is ensured by further axioms stating, in effect, that each syntactic expression can be obtained in accordance with the syntactic rules in exactly one way.

For instance, consider the operation F with three arguments, defined so that F(x,y,z) is the statement "x y z." Consider the categories "noun" and "transitive verb." There is a syntactic rule, in English, saying that this operation can generally be applied if x and z are nouns and y is a transitive verb. Thus yielding, for instance F(I, kill, you) = "I kill you".

5.1.1.1. Montague's Axioms (*)

Formally, in Montague's terminology, a disambiguated language is an ordered quintuple $(A,\{F_l\},\{X_d\},S,d_0)$, defined by the following axioms:

1) $\{F_l, l \text{ in } L\}$ is a set of syntactic operations, where L is an index set. Each operation maps finite ordered sets of syntactic expressions into syntactic expressions.

2) D is a collection of syntactic category names

3) $\{X_d$, d in D$\}$ is a set of sets of basic expressions, each associated with a category name. It is possible for the same basic expression to have two different category names and hence belong to two different X_d

4) S is a set of syntactic rules, each rule having the interpretation "If x_1 belongs to category d_1, and ... and x_n belongs to category d_n, then $F_l(x_1,...,x_n)$ must belong to category d_{n+1}" for some l in L.

5) d_0 is a special category name, to be used for the set of basic expressions denoting truth values.

6) A is the set of all expressions generated by freely applying compositions of elements of the set $\{F_l$, l in L$\}$ to the set {x: x is in X_d for some d in D}.

7) No basic expression can be an output of any syntactic operation

8) No expression in A can be the output of two different syntactic operations

9) No syntactic operation can produce the same output from two different input expressions (i.e. the F_l are one-to-one)

The formalism is obscure and complex, but the ideas are not particularly subtle. The first six axioms define the basic set-up, and the last three axioms ensure disambiguity.

5.1.2. Syntactic Systems

A Montague grammar is a transformation system, in the sense defined above -- the transformation rules are the "syntactic operations," and the initials are the "basic expressions." But it is a very special kind of transformation system. I will need to deal with a somewhat less restrictive transformation-system model of grammatical structure, which I call a **syntactic system**. The "syntactic system" contains the "disambiguated language" as a special case, but it also includes a variety of structures which the Montagovian analysis ignores.

The first step toward syntactic systems is the Sausseurean **structuralist** observation that, syntactically, "John" and "Mike" are the same, as are "cat" and "rat." It is not the meaning of a word that matters, but only **the way it relates to other words**. Therefore, it is natural to **define** a word, for the purposes of syntax, as a **relation** between other words.

More specifically, one may characterize a word as a fuzzy set of **functions**, each of which which takes in a sequence of syntactic expressions and puts out a singlesyntactic expression. And one may characterize a punctuation mark in the same way. The class of syntactic expressions need only be defined, at this point, as a subset of the set of **ordered sets of words and punctuation marks**. From here on I will omit reference to punctuation marks, and speak only of words; but this does not reflect any theoretical difficulty, only a desire to avoid tedious figures of speech.

What makes a collection of functions a **syntax** is a collection of constraints. Constraints tell us which sorts of expressions may be put into which inputs of which words. Thus they embody explicit grammatical rules, as well as "grammatico-semantic" rules such as the rule which tells us that the subject of the word "to walk" must be an **animate** object.

For instance, the word **kiss** is identified, among other functions, with a function f_{kiss} that has three arguments -- a subject, an object and a modifier. f_{kiss}(I,my wife,definitively) = I kiss my wife definitively.

And the word **wife** is identified with, among other functions, a function f_{wife} that has at least five arguments. How one lines them up is arbitrary -- one may write, for instance, $f_{wife}(x_1,x_2,x_3,x_4,x_5,x_6)$, where:

x_1 is constrained to be the subject of a verb of which **wife** is the object,

x_2 is constrained to be a verb phrase of which **wife** is the object,

x_3 is constrained to be an adjectivial phrase modifying **wife**,

x_4 is constrained to be a verb phrase of which **wife** is the subject,

x_5 is constrained to be the object of a verb phrase of which **wife** is the subject.

Arguments that are not filled may simply be left blank. For instance, f_{wife}(, ,my lovely, eats, too much pie) is "my lovely wife eats too much pie." And f_{wife}(I,kiss,my, ,) is "I kiss my wife."

f_I is similar to f_{wife} in its syntactic structure. And, more simply, f_{my} is identified with a function of **two** arguments, one of which is constrained to be a noun phrase, one of which is constrained to be an adverbial phrase.

In these simple examples I have mentioned only strictly grammatical constraints. An example of a less grammatical constraint would be the restriction of the object of "kiss" to entities which are concrete ratherthan abstract. This is an example of a constraint which need not **necessarily** be fulfilled. People may use the word "kiss" metaphorically, as in "When Elwood threw his boss out the window, he kissed his job goodbye." But if something is concrete, it fulfills the constraint **better** than something that isn't animate. Thus a constraint is a fuzzy set -- it tells not only what is allowed, but what is **better**, more easily permissible, than what.

5.1.3. A Formalization (*)

Let's get more precise. Given a set of "concrete" entities X (which may well be the empty set), a **syntactic system** over X may then be defined as:

1) A collection H of subsets of X.

2) A collection of **constraints** -- at least one for each set in H. Each constraint may be written in the form $f(i,x_1,x_2,...)$, and defines a series of fuzzy sets $C_{\mathit{j}}(i)$. Let d(j,i,x) denote the degree to which x belongs to $C_j(i)$. Then the interpretation is that, **in a situation in which infon i obtains**, x can be "plugged into" the x_j position in f with acceptability level d(j,i,x). C(f), the collection of $C_j(i)$ corresponding to f, is the collection of constraints implicit in f.

3) A collection $W_* = \{W,W\#W,W\#W\#W,...,W\#\mathit{n},...\}$, where A#B is defined as the set of all possible entities obtainable by applying the functions in A to the functions in B, and W#*n* denotes W#W#...#W iterated n times. These sets are called "possible syntactic expressions," and are the elements of the fuzzy sets $C_j(i,f)$.

Each element x of W* has a certain "admissibility" A(i,x) defined inductively as follows. The **raw admissibility** RA(x) of the abstract form "$f(i,g_1,g_2,...)$," where the g_i are in W, is the sum over j of the quantities $d(j,i,f,g_j)$. And the **raw admissibility** of an abstract form "$f(i,g_1,g_2,...)$" where the g_i are in W*#(n-1),* is the sum over j of the product $d(j,i,f,g_j) * RA(g_j)$.

Finally, each element of W* is potentially realized by a number of different abstract forms of this nature. The **admissibility** of an element E of W*, relative to a given situation s, is the **maximum** over all abstractforms x that yield E of the product $RA(x) * d_i(s)$. This measures the extent to which the formation of the expression E is grammatical.

5.1.4. A Comparative Analysis

Despite the different mathematical form, my definition of "syntactic system" is not tremendously different from the Montagovian definition of "disambiguated language." The syntactic system represents a generalization of, rather than a radical break from, Montague grammar. There are, however, several important distinctions that may be drawn between the two approaches.

For one thing, loosely following Barwise and Perry (1985), Barwise (1989), and Devlin (1991), the definition of syntactic system incorporates an **infon** i at every juncture. Each real situation supports certain infons to certain degrees. Montague assumes that there is one big situation, in which everything applies; but his axioms could be "situated" without too much difficulty (by modifying 1, 4 and 6). And, correspondingly, my axioms could be de-situated by removing all references to infons.

Next, the Montagovian approach describes only **disambiguated** languages, whereas the concept of syntactic system makes no such restriction. It could easily be restricted to disambiguated languages (by adding on a clause resembling condition 8 of the Montagovian definition -- conditions 7 and 9 are common sense and are automatically fulfilled). But there is no need. Real languages are of course ambiguous in many ways. Montague's possible worlds analysis of meaning requires disambiguity, but the semantical theory to be given below does not.

Finally, the most substantial difference is that the definition of syntactic system defines a word as a set of syntactic operations, and assigns a set of grammatical rules to **each word**. The Montague approach, more manageably, assigns each word to a certain class and then sets up syntactic operations and grammatical rules to work via **classes**.

One way to look at this difference is via algorithmic complexity. A random syntactic system would be useless -- no one could memorize the list of rules. Only a syntactic system that is truly **systematic** -- thatis simple in structure, that has patterns which simplify it a great deal -- can possibly be of any use. **One** way for a syntactic system to be simple in structure is for its constraints to fall into **categories**.

In other words, suppose the class of all words can be divided into categories so that the constraints regarding a word can be predicted from knowledge of what category the word is in. Then the syntactic-system approach reduces to something very similar to the Montague approach (to a situated, ambiguous Montague grammar). But when dealing with syntactic systems in general, not only written and spoken language, it is unnecessarily restrictive to require that all rules operate by categories. It is better and more in line with the pattern-theoretic approach to speak about general syntactic systems, with the understanding that only syntactic systems of low algorithmic complexity are interesting.

5.1.5. What is Language?

Basically, the definition of syntactic system says that each **word** (each fundamental unit) is a certain collection of **functions**, each of which takes in certain kinds of external entities and certain types of functions associated with other words. The kinds of entities and functions that a certain function takes in can depend upon the **situation** in which the associated word is used. There are certain rules by which words can be built up into more complex structures, and by which more complex structures can be built up into yet more complex structures -- each rule applies only to certain types of words or other structures, and the types of structures that it applies to may depend on the situation in which it is being used.

I will give a theory of meaning to go along with the general model of syntax outlined in the previous section. The basic idea of this theory is that the meaning of an entity is **the fuzzy set of patterns related to its occurence**.

Using this characterization of meaning, I will define a **semantic system** as a set of entities which obtain much of their meanings from each other. In this context, it will become clear that the semantical structure of written and spoken language is not at all unique. Written and spoken language may be the most **cohesive** semantic system known to us. But subtle interdefinition, and the intricate interplay of form and content, can be found to various degrees in various domains.

Language will then be defined as requiring a syntactic system, coupled with a semantic system in such a way that a property called **continuous compositionality** holds. Thus, I do not believe that one can give a thoroughly context-independent definition of language. The definition of syntactic system refers frequently to infons, which hold or do not hold in specific situations. At the very foundation of a language is the set of situations in which it evolved, and in which it is used.

5.2. POSSIBLE WORLDS SEMANTICS

The next step after Montague grammar is Montague semantics. Also known as **possible-worlds** semantics, Montague semantics is just as forbiddingly formal as Montague grammar, perhaps more so. However, as I will show, it packs much more of a philosophical punch.

First of all, Montague assumes that there is some set B of meanings. Then he assumes that, to each syntactic operation F, there corresponds a **semantic** operation G taking the same number of arguments and mapping n-tuples of meanings into meanings (rather than n-tuples of expressions into expressions). Finally, he assumes some function f mapping basic expressions into meanings. This setup determines, in an obvious way, the meaning of every element of A -- even those which areconstructed in violation of the syntactic rules. The existence of a correspondence between the F and the G is what Frege called the **principle of compositionality**.

In order to specify B, Montague invokes the notion of **possible worlds**. This notion is used to build up a hierarchy of progressively more complex semantical definitions. First of all, assume that each basic expression contains a certain number of "variables", to be filled in according to context. Suppose that knowing what possible world one is in, at what time, does not necessarily tell one what values these variables must have, although it may perhaps give further information about the meaning of the expression. This does not contradict the very general axioms given above. Then, one may define a **denotation** of an expression as something to which the expression refers in a given possible world, at a specific time, assuming a certain assignment of values to its variables.

And one may define the **sense** of an expression as that to which the expression refers **regardless** of the time and possible world. This is not a precise definition; however, one way to specify it is to define the sense of an expression as the function which assigns to each pair (time, possible world), the denotation which that expression assumes in that possible world at that time.

Finally, one may define the **Fregean meaning** of an expression as the function which maps each triple (possible world, time, assignment of variable values) into the sense which that expression assumes under that assignment. Montague calls this simply the "meaning", but I wish to reserve this word for something different, so "Fregean meaning" it is.

In this scheme, everything reduces to denotations, which may be divided into different "types" and then analyzed in some detail. For instance, one of the most important types of denotation is **truth value**. The truth value of an expression with respect to a given triple (possible world, time, assignment of variable values) is whatever truth-value it denotes in that model. Montague semantics does not, in itself, specify what sorts of expressions may **denote** truth values. It is possible to give a formal definition of the syntactic category "sentence," but not all sentences may take truth values. In English, it seems clear that statements, rather than imperatives or questions, may denote truth values; but this is an empirical observation, not a formal statement, and a great deal of work is required to formulate it mathematically.

In general, a **Fregean meaning of type T** is a function mapping entities of the form (possible world, time, variable assignment) into a denotation of type T. Montague hypothesizes that each type corresponds to an element of D, so that Fregean meaning types and syntactic categories are matched in a one-to-one manner. Compositionality then requires that any rule holding for **syntactic categories** transfers over into a rule for **Fregean meaning types**. For instance, take a syntactic rule like F(x,y,z) = x y z. This maps vectors (noun, noun, transitive verb) into declarative sentences. Then F corresponds to a **semantic** function G which maps meanings of the type corresponding to nouns and verbs, into meanings involving **truth-values** as denotations.

5.2.1. Critique of Montague Semantics

This thumbnail sketch is hardly an adequate portrayal of Montague semantics. The interested reader is urged to look up the original papers. However, I will not require any further development of possible-worlds semantics here. The reason is that I am highly skeptical of the whole project of possible-worlds semantics.

I find it hard to accept that what "1+1=2" means is the same as what "2.718281828...*i*3.1415926535...* = -1" means. However, in the standard implementations of the possible worlds approach, these two assertions both denote truth in every possible world at every time, so they are semantically identical. It is true that each of these assertions can be derived from the other, given the standard mathematical axioms. But they still mean different things.

Possible worlds semantics is **formal** in a very strange sense: it makes no reference to the actual empirical or psychological content of linguistic entities. Montague believed that the most important aspects of semantics could be developed in a purely formal way, and that considerations of content, being somehow more superficial, could be tacked on afterwards. Roughly speaking, he believed that content merely sets the values of the "parameters" provided by the underlying formal structure. But this is at best a debatable working hypothesis, at worst a dogma. The possible-worlds approach has not yet been shown to apply to any but the simplest sentences.

It is remarkable that formal logic which ignores content can deal with semantically troublesome sentences like "John believes that Miss America is bald". But sentences like "Every man who loves a woman loses her"are still troublesome. And it is a long way from these formal puzzles to ordinary discourse, let alone to, say, a fragment of Octavio Paz's poetry (1984):

Salamander

back flame

sunflower

you yourself the sun

the moon

turning for ever around you

pomegranate that bursts itself open each night

fixed star on the brow of the sky

and beat of the sea and the stilled light

open mind above the

to and fro of the sea

The contemporary logicist approach can comprehend this fragment about as well as modern quantum physics can deal with the large-scale dynamics of the brain. There is a tremendous rift between theoretical applicability and practical application.

I am not the only one to sense the fundamental impotence of possible worlds semantics. Many logicians and linguists share my frustration. In the absence of a constructive alternative, however, this frustration is not terribly productive.

One possible alternative is **situation semantics**, a theory of meaning designed to transcend Montague semantics by making reference to **information**. However, the situation semanticists approach information in a very abstract way, starting from set theory. They define an abstract unit of information called an "infon," and attempt to delineate various axioms which infons must obey. While I admire situation semantics very much, I cannot agree with the abstract, set-theoretic approach to information. It seems clear that, just as physics models objects as elements of Euclidean space rather than general sets, a successful semantic theory must come equipped with a **concrete**, particular idea as to what information is. One way to do this is to take the **algorithmic** theory of information. This is the course that will be taken in the following section.

5.3. MEANING AS A FUZZY SET OF PATTERNS

Using the theory of **pattern** and **algorithmic information**, meaning can be defined without even mentioning syntax. Even entities that are not involvedin syntactic systems can have meanings. The meaning of an entity, I suggest, is simply **the set of all patterns related to its occurence**. For instance, the meaning of the concept **cat** is the set of all patterns, the occurence of which is somehow related to the occurence of a cat. Examples would be: the appearance of a dead bird, a litter box, a kitten, a barking dog, a strip dancer in a pussycat outfit, a cartoon cat on TV, a tiger, a tail,....

There are certain technical difficulties in defining "related to" -- these will be dealt with shortly. But it is clear that some things are related to **cat** more strongly than others. Thus the meaning of **cat** is not an ordinary set but a **fuzzy set**. A meaning is a fuzzy set of patterns.

In this view, the meaning of even a simple entity is a very complex construct. In fact, as is shown in the following section, meaning is in general **uncomputable** in the sense of Godel's Theorem. But this does not mean that we cannot approximate meanings, and work with these approximations just as we do other collections of patterns.

This approach to meaning is very easily situated. The meaning of an entity in a given situation is the set of all patterns in that situation which are related to that entity. The meaning of W in situation s will be called the s-meaning of W. The degree to which a certain pattern belongs to the s-meaning of W depends on two things: how intensely the pattern is present in s, and how related the pattern is to W.

These ideas are not difficult to formalize. Let $M_{W,s}(q)$ denote the degree to which q is an element of the s-meaning of W, relative to the situation s. Then one might, for instance, set

$$M_{W,s}(q) = IN[q;s] * corr[W,q]$$

where corr[W,q] denotes the statistical **correlation** between W and q, gauged by the standard "correlation coefficient," and IN[q;s] is the intensity of q as a pattern in s. The correlation must be taken over some past history of situations that are **similar in type** to s; and it may possibly be weighted to give preference to situation which are more strongly similar to s. The determination of similarity between situations, of course, is a function of the mind in which the meanings exist.

Like all pattern-theoretic definitions, this characterization of meaning is unpleasantly messy. Thereare all sorts of loose ends and free parameters; things are not nearly so cut-and-dried as in, for example, the Montagovian possible-worlds approach. But unfortunately, this is the price which one must pay for being **psychologically reasonable**. Meaning exists only relative to a given mind, a given brain; and minds and brains are notorious for not adhering to conventional standards of mathematical nicety.

5.3.1. Meaning and Undecidability (*)

It is clear that, according to the above definitions, determining the s-meaning of any entity W is an extremely difficult computational problem. In fact, if one considers sufficiently complex situations, the problem becomes so hard as to be **undecidable**, in the sense of Godel's Theorem.

Godel showed that truth is not contained in any one formal system; and his results apply directly to the standard model-theoretic approach to semantics. But it is interesting that even a subjective, pragmatic approach to meaning cannot escape undecidability.

Chaitin (1975, 1978, 1987) has given an incisive information-theoretic proof of Godel's Incompleteness Theorem. He has proved that, for any formal system G, there is some integer N so that

1) for all n>N, there exists some binary sequence x so that the statement "H(x) > n" is undecidable in G (it can neither be proved true, nor proved false, in G).

2) "H(x) > N" is not provably true (in G) for any x

Where S is some subset of St(s), let us consider the statement "|S| > N" in this light.

First, arrange the elements of S in a specified order (y_1,...,y_N), and set $x_S = L(y_1)L(y_2)...L(y_N)$, where L maps patterns into binary sequences. Then, as |S| becomes arbitrarily large, so does H(x_S). That is, for any N, there is some M so that when |S| >M, one has H(x_S) > N. But for large enough N, the statement "H(x_S) > N" is undecidable. Consequently, so is "|S| > M". Finally, let $M_{W,s;K}$ denote the set of all Boolean q so that $M_{W,s}$(q) > K. Then I have shown

Theorem: For any formal system G, and any situation s of infinite structural complexity, there is some M so that the statement "|$M_{W,s;K}$| > M" is undecidable in G.

Godel showed that truth cannot be encapsulated in any formal system. According to this theorem, if semantics is defined in terms of information, complexity and pattern, Godel's proof applies equally well to meaning. This is philosophically interesting, becausethe informational approach to meaning makes no reference whatsoever to **truth**. But it is **not** surprising, not since Chaitin has already shown us that Godel's Theorem has as much to do with information as with truth.

5.3.2. Meaning and Possible Worlds

Let us briefly return to Montague semantics. What does the present definition of s-meaning have to do with the Montagovian approach? Montague semantics speaks of denotations, senses and Fregean meanings. Where does meaning, as I have defined it, fit in?

The present approach determines for each expression, given each situation s, a definite s-meaning. But each particular situation is, surely, a subset of the set of pairs (possible worlds, times). It may sometimes be useful to consider an entire possible world, from the beginning of time to the end, as one big situation; or perhaps to consider a "situation" as a class of events intersecting **every** possible world.

The possible-worlds approach begins with denotations: the denotation of an expression is what it expresses regardless of the possible world and time. According to the informational approach, however, there is no reason to believe that denotations such as this exist. In each possible world over each interval of time, and more generally in each situation, each entity has a certain meaning. But since the meaning of an entity is defined relative to the structure of the situation it is used in, there is no reason to believe the meaning of any entity will be constant over all possible situations. Indeed, given most **any** entity, and most any pattern, one could cook up a situation in which that pattern was **not** relevant to that entity, and hence not a part of the meaning of that entity.

This is related to a point made in Barwise (1989). Barwise argues, in effect, that the concept of what an expression expresses **regardless** of possible world and time is not meaningful, because the collection of all pairs (possible world, time) is not a set but a proper class. In order to make the possible-worlds approach set-theoretically meaningful, one must restrict consideration to some particular set of worlds.

In reality, no entity experiences or envisions every mathematically possible world, nor even a reasonably large subset thereof. And it **does** mean something to talk about the meaning of W relative to some particular fuzzy set S of situations. Formally, where $d_S(x)$ denotes thedegree of membership of x in S, $M_{W,S}(q)$ may be defined as the sum over all x in S of $M_{W,x}(q)d_S(x)$. If S is taken to be the collection of all situations in a given mind's memory, then one may omit the subscript S and simply write M_W.

This, finally, is what I mean by the "meaning" of a word or other entity W. In most practical cases, M_W is actually not all that far off from the possible-worlds definition of meaning. Let's take the word "dog," for example. To an ordinary, intelligent, English-speaking person, the concept of "dog" is not that fuzzy: certain things **are** dogs (they belong to M_{dog} with degree 1) and most things **aren't** (they belong to M_{dog} with degree 0). Some things, like wolves or wolf-dog half-breeds, might belong to M_{dog} with intermediate degrees (say .25 or .75), but these are definitely the exception. In a vast majority of the situations in which the word "dog" is used, those things which **are** dogs, or various memories involving them, take part in patterns associated with the word "dog." In Montagovian terms, the elements of M_W are very good candidates for the **sense** of the word "dog." They are, approximately, what "dog" refers to regardless of possible world and time. And for a simple expression like "dog," with no explicit variables, the sense is essentially (though not set-theoretically) the same as the Fregean meaning.

In general, for more complex expressions which may have variables in them (say, "John eats ____'s pet ____"), M_W may be computed either for the expression as an abstract formula, or for the expression **given** some particular assignment of variable values. The latter quantity will often be similar to the **sense** of the expression, given the same particular assignment of variable values. And the former quantity will often be similar to the Fregean meaning of the expression, since the Fregean meaning contains **all** senses for all possible worlds and times, and $M_{\text{John eats \_\_\_'s pet \_\_\_}}$ contains, with some nonzero degree, all elements of $M_{\text{John eats x's pet y}}$ for **every x and y**.

So, in many cases, the present situation-oriented, pattern-based definition of meaning coincides with the possible worlds definition (as well as with the situation-theoretic approach of Barwise or Devlin). This is because, to a large extent, the different approaches are getting at the same underlying intuition. However, it seems to me that the informational definition is **psychologically** a lot more sensible than the possible-worlds approach, and also a lot more sensible than the more abstract situation-theoretic analyses.

5.3. LINGUISTIC SYSTEMS

Now, at long last, we are entering the final stretch of our quest to tie syntax and semantics together. Let me begin with Frege's "principle of compositionality." This axiom, if you recall, states that the meaning of a complex syntactic construct can be determined from a knowledge of: 1) the syntactic operations involved, and 2) the meanings of the simpler syntactic constructs of which the complex syntactic construct is formed.

Mathematically, in the present formalism, compositionality says that for each syntactic operation F there is a "semantic operation" G so that

$$M_{F(x,y)}(q) = G(M_x(q),M_y(q)).$$

Clearly, this principle is not **implied** by the informational approach to meaning. But it is not forbidden either.

For starters, let us consider a rule F(x,y,z) which takes in a noun x, a transitive verb y, and a noun z, and puts out the sentence xyz: F(Sandy, kisses, Andy) = Sandy kisses Andy. The question is, is there some G so that $M_{\text{Sandy kisses Andy}}$ =

$G(M_{\text{Sandy}},M_{\text{kisses}},M_{\text{Andy}})$, and, furthermore, $M_{F(x,y,z)} = G(M_x,M_y,M_z)$ for **any** x,y,z? In other words, is the meaning of the whole determined by the meaning of the "component parts"? Knowing the set of patterns related to "Sandy", "kisses" and "Andy", and the standard grammatical rules, can one predict the set of patterns related to "Sandy kisses Andy"?

Or, to take an absurd example, what if English contained a rule F'(x,y,z), taking arguments x and z human beings, and y a transitive verb, defined so that

F'(x,y,z) =

F(the father of x, the last transitive verb in the Standard High School Dictionary before y, the mother of z).

Montague's restrictions on semantic rules forbid this sort of construction, but the general definition of semantic system places no such restrictions. Then F'(Sandy, kisses, Andy) might equal, say, "Jack kings Jill". There would be **no way** to predict the meaning of "Jack kings Jill" from the meaning of "Sandy", "kisses" and "Andy". The point is that real written and spoken languages do not have crazy rules like this -- and the main reason they do not is compositionality.

Finally, consider an example discussed in Barwise (1989), the opening sentence of Hoban's novel *Riddley Walker*:

On my naming day when I come 12 I gone front spear and kilt a wyld boar he parbly ben the las wyld pig on the Bundel downs any how there hadnt ben none for a long time befor him nor I aint looking to see none agen.

Barwise asks how a compositional account of meaning could possibly explain the meaning of the phrase "gone front spear" -- let alone the whole sentence. The same question could of course be asked in regard to much modern poetry and literature. The point is that we automatically assign meaning even to expressions that are formed in **violation** of the rules of grammar. If an expression is formed in violation of the rules of grammar, there is no way to compute its meaning by going from a function F to a function G as compositionality suggests.

Barwise's "gone front spear" argument is fatal for strict Montague semantics. But it certainly does not imply that compositionality is totally absent from natural languages. I suggest that compositionality is a **tool** for estimating meanings, and a very powerful one. Without this tool, it would be hard to estimate the meaning of a sentence that one had never heard before. However, like all real tools, compositionality is not a complete solution for every problem.

A language would be basically **useless** if it did not possess approximate compositionality for **most** words, most syntactic operations, F. *Riddley Walker* and *Naked Lunch* are more difficult to read than *Huckleberry Finn* and *Catch-22*, and this is precisely because when assigning meaning to the sentences of the former books, one must depend less on compositionality, and more on subtle structural clues internal to the semantics.

One final note is in order. I have been talking about language in a very general way, but the examples I have given have been either Boolean logic or common English. These are good sources for examples, but they may also be misleading. In these cases, compositionality takes a particularly simple form: the deductive **predecessors** of an expression are also its **components**. For instance, where F(Sandy, kisses, Andy) = "Sandy kisses Andy", the arguments "Sandy," "kisses," and "Andy" are **parts** of the sentence "Sandy kisses Andy." Here,compositionality requires that the meaning of a whole predictable from the meaning of its parts.

Fodor (1987), among others, thinks this is essential to the concept of compositionaliy. But on the other hand, nothing in the present theory of language **requires** that the relation between the output of a function and its arguments be a whole/part relationship. This point is particularly relevant in the context of the recent work of Tim van Gelder (1990), which suggests that certain neural network models of thought possess compositionality **without** displaying any sort of whole/part relationship between expressions and their deductive predecessors.

5.4.1. Formal Meaning and Dictionary Meaning

What do all these abstract mathematical definitions of "meaning" have to do with meaning in the dictionary sense? When one looks up a word in a dictionary, one certainly does not find a huge fuzzy set of regularities spanning different situations: one finds a phrase, or a sentence, or a number of sentences!

The answer to this most natural question is as follows. When one looks up a word like "high-falutin'" or "cosmological," one finds a sentence consisting hopefully of simpler words. Using compositionality (as well as of course all our knowledge of grammar and semantics), one construes the meaning of that sentence from the meanings of the simple words, and thus infers the meaning of the word in question. One never learns **any** word as well from the dictionary as from hearing it in practice, but for some words the dictionary can yield a good approximation.

For other words, however, such as "the" or "a," the dictionary is totally useless for imparting meaning -- it can impart technical niceties to someone who already basically **knows** the meaning, but that's about it. And for words like "in" or "out" the dictionary almost as useless -- "in" refers you to "inside," which refers you back to "in," et cetera. The words that are possible to learn from the dictionary are those words which could reasonably be **replaced** in conversation by complex phrases, which could then be understood by appeal to compositionality.

5.4.2. Semantic Systems

We have defined the **syntactic** system, identified the relation between syntax and semantics, and given a newtheory of meaning. What remains is to crystallize this theory of meaning into a definition of the **semantic** system.

Intuitively, a **semantic system** V is a collection of entities whose meanings consist primarily of patterns involving other elements of the system. The **systematicity** of a collection of patterns is the extent to which that collection is a semantic system.

More formally, let D(x/V) denote the percentage of the meaning of x which involves elements of V. This is an intuitively simple idea, but its rigorous definition is a little involved, and will be postponed to the end of the section.

A systematic collection of entities is characterized by a high average D(x/V) for x in V.

Written and spoken languages are examples of collections with **very high** systematicity. The meaning of the word "dog" has a lot to do with physical entities. It also has to do with other linguistic entities: certain aspects of its occurence can be predicted from the fact that it is a noun, that it is animate, etc. But it is among the less dependent words; D(dog/English) is not all that large. On the other hand, the meaning of the word "the" has virtually nothing to do with nonlinguistic entities, and therefore "the" contributes a great deal to the systematicity of English. D(the/English) is certainly very large.

The Sapir-Whorf hypothesis rests upon the assumption that languages are highly systematic. That is its **starting-point**. If the meanings of words and sentences had to do primarily with extra-linguistic phenomena, then how could language have the power Whorf ascribes to it? It is only **after** one realizes the extent to which linguistic entities **depend on each other** for their significance, that one can conceive of language as a coherent, acting entity.

But written and spoken languages are almost certainly not the only systematic meaning systems. It seems that each sensory modality probably has its own semantic system. For instance, the set of patterns involving the visual entity "box" has a lot to do with other visual forms and not that much to do with anything else. And the same goes for most visual forms. Hence, intuitively, one might guess that the collection of visual forms is highly systematic.

5.4.2.1. Formal Definition of D(x/V) (*)

Finally, before moving on, let us deal with the problem of defining D(x/V). Although I will not be using this definition for any specific computations or theoretical developments, it is important to have a precise definition in mind when one speaks about a concept. Otherwise, one does not know what one is talking about.

One approach to defining D(x/V) is as follows. First, for each q and each s, define the degree D(x/V;s,q) to which $M_{W,s}(q)$ involves V as the **maximum**, over all elements v of V, of the expression

$M_{W,s}(v) * corr[v,q] * |St[v]\ St[q]| / |St(q)|$.

This product can never exceed 1; it is close to 1 only if v:

1) is an element of the meaning of q in the situation s,

2) is statistically correlated with q

3) contains much of the same structure that q does

Next, define D(x\V;s) to be the average of D(x/V;s,q) over all q. And, where S is a set of situations, define D(x/S) to be the average of D(x/V;s) over all s in S. Where S is taken to be all situations in the memory of a given mind, one may omit reference to it, and simply speak of D(x/V).

5.4.3 The Definition of a Linguistic System

Now all the hard work is mercifully past. I am prepared to give a pattern-theoretic, "informational" definition of a **linguistic system**. First of all, let us state some minimal requirements. Whatever else it may be, every linguistic system must consist of

1) a **syntactic system**, together with

2) a **collection of situations**, so that this syntactic system, in this collection of situations, gives rise to

3) a **semantic system**, in which the meanings of most expressions may be approximately determined by compositionality.

This is quite a mouthful. But it is not quite enough to constitute an adequate definition of "linguistic system." To see what else is needed, let us recall the concept of **structured transformation system**, defined in Chapter Four. Now, a syntactic system is a transformation system -- this follows immediately from acomparison of the two definitions. But what about the "structured" part?

Does semantics, combined with compositionality, have the capacity to induce a structure on the transformation system that is syntax? What is needed is that **grammatically similar** linguistic constructions (sentences) also tend to be **structurally similar** (where the complexity measure implicit in the phrase "structurally similar" is defined relative to the environment in which the sentences are used). But, if one knew that syntactically similar sentences tended to have similar **meanings**, this would follow as a consequence. One could form a sentence with meaning X by analogy to how one has formed sentences with meanings **close** to X.

The principle of compositionality, under my loose interpretation, implies that for most syntactic operations F there is a "semantic operation" G so that $M_{F(x,y)}$ is close to $G(M_x,M_y)$. But this does not imply that sentences formed by similar rules will tend to have similar meanings. I need an additional hypothesis: namely, that **small changes in F correspond to small changes in G**. It is not enough that each syntactic rule corresponds to a semantic rule -- this correspondence must be "stable with respect to small perturbations."

This property may be called **continuous compositionality**. A little more formally, suppose that F(x,y) and F'(x,y) are close. Compositionality guarantees that there are G and G' so that:

1) $M_{F(x,y)}$ is close to $G(M_x,M_y)$, and

2) $M_{F'(x,y)}$ is close to $G'(M_x,M_y)$.

Continuity of compositionality requires that G and G' be close. But relations (1) and (2) render this "continuity requirement" equivalent to $\mathbf{M_{F(x,y)}}$ and $\mathbf{M_{F'(x,y)}}$ being close.

So, all formalities aside, one may define a linguistic system as

1) a **syntactic** system, together with

2) a collection of **situations**,

3) so that relative to these situations the expressions of the syntactic system form a **semantic system**

4) which is related to the syntactic system according to **continuous compositionality**.

From this definition, one has the immediate result that a linguistic system is a structured transformation system.

Boolean logic, as analyzed in Chapter Four, is a specific example of a linguistic system; in fact it is a subset of natural languages. I have pointed out somerelations between analogical structure and deduction in the context of Boolean logic: these may now be understood as examples of the behavior of linguistic systems, and special cases of the complex dynamics of natural language.

5.3.4. Communication

What is the **purpose** of language? The straightforward answer to this question is "communication." But what exactly does this elusive term denote? The so-called "mathematical theory of communication," founded by Claude Shannon, deals with the **surprise value** of a message relative to a given ensemble of messages. But although this is marvelous mathematics and engineering, it has little to do with **meaning**. The communication of **patterns** is different from the communication of statistical information.

Let us consider the five "illocutionary categories" into which Searle (1983) claims all speech acts may be categorized:

Assertives, which commit the speaker to the truth of an expression

Directives, which attempt to get the speaker to do something. This category is inclusive of both commands and questions.

Commissives, which commit the speaker to do something -- say to join the Navy, or to tell the truth in a court proceeding.

Expressives, which express a psychological state on the part of the speaker

Declaratives, which, by virtue of being uttered, bring about the content of the utterance. For instance, "I pronounce you man and wife."

One could modify this list in various ways. For instance, what Searle calls "assertive" is sometimes called "declarative." And I am not sure about the boundary between assertives and expressives: it is not a crisp distinction. Many utterances combine both of these types in a complicated way -- for example, "My head hurts worse than yours." But these quibbles are irrelevant to what I want to do here.

All of these categories have one obvious thing in common. They say that the speaker, by using a speech act, is trying to **cause** some infon to obtain. In the case of expressives and assertives, one is mainly trying to cause an infon (the content of one's statement) to obtain in the mind of the listener. In particular, among other things, one is telling the listener **the situationin question/ speaker |-- this content**. In the case of assertives, one may also be trying to cause **the situation in question/ listener |-- this content** to appear -- that is, one may be trying to convince the listener to agree with you. But at any rate, the most **basic** thing you are doing is trying to cause a record of what **you** think or feel to occur in **her** mind.

In the case of directives, one is trying to cause the listener to respond either with an assertive statement of her own (in the case of a question) or with some other sort of action. One is trying to make a certain infon appear in one's present physical situation, or in some future situation.

Finally, in the case of commissives and declaratives, things are even more direct. One is swearing oneself into the Navy, or declaring two people married. Within the network of beliefs that makes up one's subjective world, one is actually **causing** certain infons to obtain.

So what **communication** really comes down to, is **molding the world** in a certain way. How does it differ from **other** means of molding the world, such as building something? Only, I suggest, in that it partakes of the deductive and analogical system associated with a given **language**. Rather than defining language as that which communicates, I propose to define communication as the process of **doing something** with language.

In the context of the model of mind outlined in Chapter Three, the definition of language given above might be reformulated as follows: **a linguistic system is a syntactic system coupled with a semantic system in such a way that the coupled system is useful for molding the world**. After all, a syntactic system is useless for molding the world unless it is appropriately coupled with an analogical, associative-memory-based system. And a semantic system can serve in this role only if the property of continuous compositionality is present.

In Chapter Four I considered a very restrictive linguistic system -- Boolean logic. I showed in detail how the syntactic system of Boolean logic is useless in itself -- but extremely useful when appropriately coupled with a semantic, analogical network. With more general languages, many more issues are involved -- but the basic picture is the same. A linguistic system is a syntactic system coupled with a semantic system so as to make communication possible.

5.5. LANGUAGE IN PERCEPTION AND BEHAVIOR

I have theorized about **general** linguistic systems; but the only linguistic systems I have explicitly discussed are Boolean logic and written/spoken language. I will now briefly consider three other linguistic systems, which at least as essential to the functioning of mind. The treatment of these systems will be extremely sketchy, more of an indication of directions for development than a presentation of results. But it would be unthinkable to completely ignore three linguistic systems as essential as perception, motor control and social behavior.

5.5.1. Perception, Action and Language

Let us begin with Nietzsche's analysis of the "inner experience" of an outer world as a construct of language and consciousness:

> The whole of "inner experience" rests upon the fact that a cause for an excitement of the nerve centers is sought and imagined -- and that only a cause thus discovered enters consciousness; this cause in no way corresponds to the real cause -- it is a groping on the basis of previous "inner experiences," i.e. of memory.... Our "outer world" as we project it every moment is indissolubly tied to ... old error.... "Inner experience" enters our consciousness only after it has found a language the individual understands. (p. 266)

In this view, experience enters consciousness only after it has **found an appropriate language**.

Nietzsche also observed that language and perception are similar, both being based on **making equal** that which is not.

> First **images**.... Then **words**, applied to images. Finally **concepts**, possible only when there are words.... The tiny amount of emotion to which the "word" gives rise, as we contemplate similar images for which **one** word exists -- this weak emotion is the common element, the basis of the concept. That weak sensations are regarded as alike, sensed as **being the same**, is the fundamental fact.... Believing is the primal beginning even in every sense impression.... (p.275)

This penetrating observation implies that, in a sense, language is to the middle levels what systematic perception is to the below-conscious levels. Language is based on the identification of word-concepts, which is the recognition of common patterns among the outputs of lower-level, perceptual-motor processes. Perception, on the other hand, is based on the identification of common patterns among the outputs of "sensory organs" or else lower-level perceptual-motor processes. **Both are systematic, with a grammar and a semantics; both are meaning-generating structured transformation systems**.

In humans, visual perception, at least, has a very complicated grammar. The visual cortex **builds a scene** out of the simple parts which it perceives, and it is this scene rather than the individual stimuli which it feeds up to consciousness. And the aural cortex does the same thing, in a less involved way: we listen to someone talking and hear **words**, but these words are pieced together according to a complex system of rules from particular blurry, superimposed sounds. These sensory-modality-dependent rules for building wholes out of parts are full-fledged, situation-dependent grammars.

And there is no doubt that, in the sense defined above, visual and aural forms constitute very intricate semantic systems. Compositionality is slightly confusing: are the **meanings** of the raw sounds or visual stimuli experienced by low-level processes sufficient to determine the meanings of the complex combinations which the conscious mind experiences? Internally, from the point of view of the conscious perceiving mind, raw sounds and visual stimuli **have** meanings, in the sense of being algorithmically related to other things, only through these complex combinations. Therefore, from the phenomenological point of view, compositionality is only interesting above a certain level. Below that level, it is either obvious or meaningless: the meaning of the parts **is** the meaning of those wholes that the part contributes to; the parts have no independent significance.

However, from the point of view of a real or hypothetical **external** observer, with access even to patterns below the level of consciousness, compositionality is interesting all the way down. It is perfectly sensible to ask whether the patterns associated with certain raw stimuli are sufficient to determine the patterns associated with something constructed out of them. And the **answer** to this question should be "yes" -- if, as proposed in Chapter Three, the perceptualhierarchy does indeed operate on the basis of pattern recognition.

Similar arguments apply to motor control. Motions such as the wave of an arm, the kick of a leg, the fast walk, the jog, the shoulder shrug, the sigh -- none of these are indivisible units; all of them are formed by the systematic assemblage of more basic muscle movements. An excellent description of this process of systematic assemblage was given by Charles S. Peirce (1966):

> [M]ost persons have a difficulty in moving the two hands simultaneously and in opposite directions through two parallel circles nearly in the medial plane of the body. To learn to do this, it is necessary to attend, first, to the different actions in different parts of the motion, when suddenly a general conception of the action springs up and it becomes perfectly easy. We think the motion we are trying to do involves this action, and this and this. Then the general idea comes which unites all these actions, and thereupon the desire to perform the motion calls up the general idea. The same mental process is many times employed whenever we are learning to speak a language or are acquiring any kind of skills.

As Peirce points out, learning a motion is a process much like learning a word or a grammatical form, or learning how to add, or learning to recognize a chair as a chair regardless of lighting and orientation. One combines different things into one -- learns to perceive them as one -- because they all serve a common purpose.

But what Peirce does **not** point out is the systematicity of **all** these processes. There are certain **tricks** to learning complex motions, which may not be easy to formulate in words, but which everyone knows intuitively. Some people know more of these tricks than others, but almost all adults have the body of tricks down better than little children. When learning to throw something new -- say a football, or a frisbee, or a javelin -- one operates by **putting together** various accustomed motions. One combines familiar elementary motions in various ways, based on past experience of combining motions **in similar situations**, and then experiments with the results. What makes the process linguistic is the application of different combinatory rules in different situations, and the automatic,systematic assignment of **meanings** to the different combinations.

So, in summary, I suggest that perceptual and motor systems are STS's and, more specifically, **languages** in the sense described above. Nietszche's perception of a similarity between sensorimotor processes and written/spoken language was right on target. This idea may be fleshed out by reference to the modern empirical literature on perception and control, but that is a major task which would take us too far afield.

5.5.2. The Language of Social Behavior

What does it mean to say that a **behavioral** system is a language? Instead of "words," the fundamental units here are specific behaviors, specific acts. One communicates with acts: one acts in certain ways in order to cause certain infons to obtain in the minds of others or in physical reality.

The system of behaviors used by a human being is **inclusive** of the system of speech acts used by that person, as well as of gestures, tones of voice and "body language." But it also includes less overtly communicative acts, such as walking out of a room, taking a job, getting married, cooking dinner, changing the TV channel, etc. This system is in fact **so** large that one might doubt whether it is really a cohesive system, in either the syntactic or semantic senses.

But it is clear that we build up complex acts out of simpler ones; this is so obvious that it hardly requires comment. And there are of course **rules** for doing so. Thus there is a **syntactic**

system of some sort to be found. The only question, then, is if this syntactic system coordinates with a semantic system in the proper way.

I claim that it does. First of all, the structural definition of meaning is perfectly suitable for characterizing the meaning of an act. The meaning of an act is those regularities that are associated with it. In this context, it is not too hard to see that compositionality holds. For instance, the meaning of a woman kissing her husband, quitting her job, and writing a surrealist poem about her cat is approximately predictable from the meaning of a woman kissing her husband, the meaning of a woman quitting her job, and the meaning of a woman writing a surrealist poem about her cat. Or, less colorfully, the meaning of tapping one's feet while wrinkling one's brow during a lecture isapproximately predictable from the meaning of foot-tapping during a lecture, and the meaning of brow-wrinkling during a lecture.

Compositionality is fairly simple to understand here, since the "syntactic" combination of acts tends to directly involve the component acts, or at least recognizable portions thereof. For instance, the meaning of carrying a gun on a New York City bus is easily predictable from, 1) the meaning of carrying a gun and, 2) the meaning of being on a New York City bus.

Compositionality is not **always** the most useful way to compute meanings. For instance, the meaning of carrying a gun on an airplane is not so easily predictable from, 1) the meaning of carrying a gun and, 2) the meaning of being on an airplane. Carrying a gun on an airplane is highly correlated with **hijacking**; this is an added meaning that is not generally associated with the function F(x,y) = carrying x on y.

Even in this example, **some** degree of compositionality may be present. The airport security check is **part** of the meaning of being on an airplane, so for a frequent airline passenger it may be part of the meaning-function G associated with F(x,y). But the degree $M_{\text{being on airplane}}$(security check) is fairly small, thus making the compositionality weak at best.

The syntactic rules governing the formation of appropriate acts for different situations are extremely complex. It is not clear whether they are as complex as the syntactic rules of written and spoken language -- we know that the latter rules have been charted more thoroughly, and indeed are easier to chart, but that does not tell us much. Just as the rules of spoken language tell us how we should form verbal expressions in order to get across desired meanings, so do the rules of behavior tell us how we should form **complex behaviors** out of simpler components in order to get across desired meanings.

The work of Erving Goffmann (1959, 1961), perhaps more than that of any other single investigator, went a long way toward elucidating the manner in which simple acts are built up into complex socio-cultural systems. In *The Presentation of Self in Everyday Life*, Goffman understood social interaction according to the dramaturgical metaphor. Each person, in each **situation**, has a certain impression which she wants to put across. She puts together a "performance" -- a complex combination of simple acts -- that she judges will best transmit this impression. One might say that the **performance** is the analogue in behavior of the"conversation" or "discourse" in speech -- it is the large-scale construction toward which basic units, and smaller combinations thereof, are combined. Goffman's ideas are particularly appropriate here because of their focus on situations. This is not the place to review them in detail, however -- the only point I want to make here is merely that performances are very complex. We tend not to notice this complexity precisely because performances are so routine to us. But try to explain to a person of working-class background how to make a good impression at an interview for a white-collar job. Or try to explain to a person of upper-class background how to hang out in a ghetto bar for six hours without attracting attention. Experiments of this nature show us how much we take for granted, how complex and interconnected are the arrangements of simple acts that we use in our daily lives.

Since the time of Goffmann's early work, a great number of social psychologists have investigated such phenomena, with often fascinating results. Callero (1991), thinking of the incredible complexity of social roles which this work has revealed, states that "a literal translation of a role into specific behavioral requirements for specific actors in specific situations is simply not possible." But I think this statement must be tempered. What **is** possible, what **must be** possible, is the expression of a role as a **network of processes** which implicitly computes a fairly continuous function mapping certain situations into certain fuzzy sets of behaviors. The network tells what a person playing that role is allowed to **do** in a given situation. But the mathematical function implicit in this network is far too complex to be displayed as a "list" of any manageable size. In this interpretation, Callero's statement is correct and insightful.

Chapter Six

CRUCIAL CONNECTIONS

Everything is related to everything else; in fact, if properly perceived, any one thing can be seen to **contain** everything else. This interpenetration, however, need not act as a hindrance to thinking about the overall nature of the world. One must merely pick some concept as a starting point, arbitrarily, and take it where it leads. The deeper one digs into one's initial concept, the more of the interconnected web of ideas one will uncover.

Our main concerns so far have been logic, language, and their roles in the mental network. In this chapter, the scope of the discussion will broaden, almost to the point of disorganization (but not quite). I will consider language in its connection to deductive thought, consciousness, evolution, and physical reality. But this does not represent a digression or a change of subject: it is merely a matter of delving **deeper** into the nature of language, so deep that one encounters these other issues as well.

The connections drawn in this chapter will be essential to the rest of the book. I will pose the crucial question of how language, logic and consciousness conspire with memory to create self, intuition and reality. The "final" resolution of these question will wait until the final chapter, when ideas regarding belief systems and cognitive dynamics can be drawn into the picture. But with the mere posing of the question, half the work is done.

6.1. THE WHORF CONTROVERSY

I have defined communication as the use of language to mold the world. But I have not yet probed the difficult question of **just how useful language is**. The "Sapir-Whorf hypothesis," also known as the hypothesis of linguistic determinism, suggests that the influence of communication is very great indeed. It claims that language is the main constructive force underlying the world that we see around us.

In this section I will give a new perspective on linguistic determinism. I will argue that, when viewed in a sufficiently abstract way, linguistic determinism is a natural consequence of the structure of mind. This does not imply that spoken language is responsible for every aspect of the world you see in front of you -- but it **does** mean that the maintenance of the belief systems which we call "self" and "external reality" would be impossible without the aid of sophisticated linguistic systems.

As has often been observed, the Sapir-Whorf hypothesis may be divided into two separate parts. **First**, the idea that the structure of language is closely related to the structure of mind and "subjective" reality. **Second**, the idea that the structural differences between the languages of different cultures are sufficiently large to imply that these different cultures have significantly different "subjective" realities.

The first claim is the central one. The second claim **implies** the first. If one demonstrates that cultures think differently because they use language differently, then one has demonstrated **a fortiori** that language determines thought. But, suppose it turned that out cross-cultural differences in language and thought were small or uncorrelated -- this would speak against the second claim, but not the first.

Most of the criticism of Whorf's work, however, has centered on his particular arguments for the second claim, which are less theoretical and more empirical. The statistical work of Lucy (1987), Bloom (1981) and others shows that grammatical patterns **do** influence patterns of attention, memory and classification to a certain extent. However, Whorf seems to have exaggerated this extent somewhat. He may well have underestimated the degree of commonality between the language, logic and world-view of an aborigine and the language, logic and world-view of a New Yorker. For a concrete example of Whorfian thought, consider that, in English, we call words like "lightening, spark, wave, eddy, pulsation, flame, storm, phase, cycle, spasm, noise, emotion" nouns. Even though they refer to temporary phenomena, we tend to think of them as definite **entities**, and this is probably related to the way our language treats them.

In the Hopi language, 'lightning, wave, flame, meteor, puff of smoke, pulsation' are verbs -- events of necessarily brief duration cannot be anything but verbs. 'Cloud' and 'storm' are at about the lower limit of duration for nouns. Hopi, you see, actually has a classification of events (or linguistic isolates) by duration type, something strange to our modes of thought.

Based on this analysis, I would bet that Whorf is correct to hypothesize that a Hopi monolingual will tend to classify events by duration, whereas an English monolingual will only do so to a lesser degree. This is in line with the relatively conservative quasi-Whorfism of Lakoff (1987), Searle (1983), etc.

Thus a Hopi monolingual will be **less** likely than an English monolingual to think about waves by analogy to particles, or to think about meteors as falling **objects**. And some of the analogies and correspondences that come naturally to a Hopi monolingual, will take longer to come to an English monolingual. All this does not mean that there are ideas which are **forbidden** to a person by the "decree" of her language. But, as argued extensively in *The Structure of Intelligence*, analogy guides the mind in its every move. It is the reason for the structure of memory. To influence analogy is to influence cognition, memory and behavior.

6.1.1. The Trouble with Translation

Emily Schultz (1990, p. 25) has suggested that Whorf **intentionally** overestimated the degree of variance between languages, and the degree of control which language exerts over thought processes. Had he not done this, she claims, he would not have been so easily able to convince his audience of the essential dependence of thought on language. Parts of the following analysis of Whorf's ideas are inspired by the excellent discussion given in Schultz (1990).

To fully understand the debate over Whorf's ideas, one should really read his essays, most of which are not at all difficult. But, to get some sense of the problem, let us listen to Au (1983, 182-183), an ardent anti-Whorfian:

Many French teachers have told their English-speaking students that "Comment allez-vous?" which is literally "How go you?" actually means "How are you?" ... I wonder if some day an Apache speaker will tell us that Whorf's English translation, "as water, or springs; whiteness moves downward" actually means "It is a dripping spring"; and if a Shawnee speaker will one day tell us that "direct a hollow moving dry spot by movement of tool" actually means "cleaning a gun with a ramrod."

Au is obviously misleading us here: there is no way that his French example is analogous to his Apache and Shawnee examples. "How go you?" is not that far off from "How's it going?", which American English speakers recognize as being very similar in meaning to "How are you?" So the difference between French and English in the instance which Au gives us is very little indeed. It is unlikely that the difference between Hopi and English in describing a dripping spring is as little as the difference between French and English in this given example -- after all, French and English are closely related, and English and Apache are rather unrelated as languages go.

The "dripping spring" passage in Whorf [p.241] goes as follows:

We might isolate something in nature by saying "it is a dripping spring." Apache erects the statement on a verb **ga**: "be white (including clear, uncolored, and so on)." With a prefix **no-** the meaning of downward motion enters: "whiteness moves downward." Then **to**, meaning both "water" and "spring" is prefixed. The result corresponds to our "dripping spring," but synthetically it is "as water, or springs, whiteness moves downward." How utterly unlike our way of thinking!

Hoijer (1953, p.559) has given a slightly different and very penetrating analysis of this phrase "tonoga" or "tonoogah":

Dripping Springs, a noun phrase, names a spot in New Mexico where the water from a spring flows over a rocky bluff and drips into a small pool below; the English name, it is evident, is descriptive of one part of this scene, themovement of the water. The Apache term is, in contrast, a verbal phrase and accentuates quite a different aspect of the scene. The element **to**, which means "water," precedes the verb "noogah," which means, roughly, "whiteness extends downward." **Tonoogah** as a whole, then, may be translated "water-whiteness extends downward," a reference to the fact that a broad streak of white limestone deposit, laid down by the running water, extends downward on the rock.

Note that Whorf has **moves** where Hoijer has the less active **extends**. Also, note that although Hoijer emphasizes that **tonoogah** refers to limestone, he does not say that it refers **only** to limestone and not at all to water -- if it did not refer to the moving water at all, its classification as a **verbal** phrase would need some explanation.

Hoijer's analysis is actually more interesting than Whorf's: it points out that the Apache and the English are looking at **different** aspects of the **same** physical situation. To use the notation introduced in Chapter Two, **Dripping Springs / average-English-speaker** and **Dripping Springs / average-Apache-speaker** are not the same entity.

Depending on which language she uses, a person will tend to **look at** and to **remember** different aspects of Dripping Springs. Dripping Springs will more likely to be connected to **white** things in the mind of an Apache speaker than in the mind of an English speaker.

In some cases Whorf may indeed have been guilty of exaggerating the differences between Amerindian and Indo-European languages. But the matter is not so simple as Au and the other critics believe. Translation is always problematic, even between similar languages but especially between dissimilar ones. Of the *Tao te Ching*, G. Spencer-Brown (1972) writes

I possess some half-dozen or so of the forty-odd translations into English alone. They differ widely because the Chinese language is so powerful that any 'translation' into a western language provides only one of the many possible interpretations of the original. Chinese is a pictorial language, very poetical and mathematical, with no grammar and no parts of speech.

Whether or not you accept Spencer-Brown's assessment of the "power" of Chinese, it is indisputable that a large number of Chinese scholars, mostly competent and with no particular ax to grind, have produced rather different translations of the same very simple work. Chinese seems to permit an ambiguity that cannot be directly translated into English; when translating, one has to pick **one** of the several possible meanings. Of course, the ambiguity **could** be more accurately transmitted by providing a list of possible interpretations instead of just one, but there is a big psychological difference between a list of statements with varying meanings and a brief statement with a variety of intrinsic meanings. The latter conveys the **interconnectedness** of the various meanings in a direct way that the former cannot match.

This is not to say that the monolingual American reader of the *Tao te Ching* can **never** get a sense for the inter-relatedness of the various meanings contained in the original Chinese. It is just to say that she will have to work a little harder to get such a sense, that such a sense will tend to come more naturally to someone who reads the original Chinese. And the monolingual American reader will have an **easier** time getting this sense if she reads several different translations.

So, translation between disparate languages is a genuine problem. If Whorf made Hopi and Apache sound very different from English, but someone else can provide translations that makes Hopi and Apache sound more similar to English, what does that tell us? That one of them was right, and the other wrong? Who's to say that every Amerindian expression has **one true meaning** that can be formulated in **one** simple English expression? In Chapter Five I presented a semantical theory which indicates that meaning is indeed not this simple: that the meaning of even a simple word can be complex and hard to specify precisely.

So it is hard to say whether Whorf translated "accurately" or not. His translations were never blatantly **in**accurate; they were always within the bounds of plausibility in that they maintained the commonsense meanings of the expressions involved. But what if it **were** true that Whorf overemphasized certain aspects of the meanings of Amerindian expressions -- namely those aspects that he felt would seem most alien to average American readers? From **his** interpretations he judged that Apache-, Hopi- or Shawnee-speaking Amerindians tend to think about things differently than English-speaking Americans. From **other** interpretations one might not conclude this. If **both** interpretations have some degreeof validity, then the proper conclusion is that these Amerindians **do** tend to think about things differently than Americans, but probably rather less so than Whorf believed. For the semantic differences which Whorf pointed out are **there**, they are just not as important as Whorf thought, because they do not **exhaust** the meanings of the Amerindian expressions in question.

Thought is influenced by **all** aspects of the meanings of the words and sentences it uses; it is not controlled by any of them. The view of meaning as a fuzzy set of patterns makes this point particularly clear. Whorf focused on certain subsets of the meaning-sets of Amerindian words, chosen for interest and shock value. Others claim that these subsets are not as important as Whorf thought; they argue, in effect, that the subsets which Whorf identified have **small degrees** of membership in the meaning fuzzy sets of the words and sentences he translated. But unless the degrees involved are truly **negligible**, which seems highly unlikely, this sort of quibble does not have much force against Whorf's general theory of language and mind.

6.1.2. Chinese and Western Modes of Thought

Some of the most intriguing evidence in favor of the Sapir-Whorf hypothesis may be found in a little book by Alfred Bloom, entitled *The Linguistic Shaping of Thought* (1981). This book dispels two illusions at once: first, the idea that the Sapir-Whorf hypothesis is empirically false; second, the idea (which one might get, for example, from Lucy (1987)) that the Sapir-Whorf hypothesis is true, but only in ways that are philosophically and psychologically uninteresting. For example, Bloom reports that

> In 1972-73, while I was in Hong Kong working on the development of a questionnaire designed to measure levels of abstraction in political thinking, I happened to ask Chinese-speaking subjects questions of the form, "If the Hong-Kong government was to pass a law requiring that all citizens born outside of Hong Kong make weekly reports of their activities to the police, how

would you reach".... Rather unexpectedly and consistently, subjects reacted "But the government hasn't," "It can't," or "It won't." I tried to press them a little by explaining, for instance, that "I know the government hasn't and won't, but let us imagine that it does or did...." Yet such attempts to lead the subjects to reason aboutthings that they knew could not be the case only served to frustrate them and to lead to such exclamations as "We don't speak/think that way!," "It's unnatural," "It's unChinese!" Some subjects with substantial exposure to Western languages and culture even branded these questions and the logic they imply as prime examples of "Western thinking." By contrast, American and French subjects, responding to similar questions in their native languages, never seemed to find anything unnatural about them and in fact readily indulged in the counterfactual hypothesizing they were designed to elicit.

The unexpected reactions of the Chinese subjects were intriguing, not only because of the cross-cultural cognitive differences they suggested, but also because the Chinese language does not have structures equivalent to those by which English and other Indo-European languages mark the counterfactual realm.

In giving a routine political questionnaire, Bloom stumbled upon an apparent parallel between **patterns of language** and **patterns of thought**.

Subsequent empirical tests verified Bloom's original intuition. Given the same stories to read, Chinese students were far less likely than American students to place a counterfactual interpretation upon them. For example, given information of the form "The philosopher Bier, if he had come into contact with X, would have done Y," Chinese students were **far** more likely to assume that Bier had done things **related to** Y.

Of course, Bloom is not proposing that Chinese speakers cannot reason counterfactually. He gives examples of counterfactual statements in Chinese. Compared to their Indo-European counterparts, however, these are protracted and awkward. The point is that thinking counterfactually is much easier for us than for the Chinese, because our language provides us with ready-made schema for doing so.

These results are surprising and tremendously important. When I first read of them, my reaction was utter disbelief. After all, every Chinese mathematician uses **reductio ad absurdum**, a theorem-proving strategy which is explicitly counterfactual in nature. Obviously Chinese mathematicians develop a mental "schema" for applying counterfactual reasoning to mathematical statements.

But, after putting variants of Bloom's original survey question to several Chinese mathematicians of my acquaintance, I became a believer. My informal survey indicated that Chinese people, even those who speak reasonable English, are simply not comfortable thinking counterfactually about commonplace situations. Counterfactual reasoning in mathematical proofs would seem to be, psychologically, a different "routine" from counterfactual reasoning regarding politics and everyday life. This is an intriguing example of mental "modularization." Just as a person who reasons logically about chess need not reason logically about her boyfriend's activities, a person who reasons counterfactually in mathematics need not reason counterfactually about commonplace real-world events.

Bloom also studied other, related differences between Chinese and Indo-European languages: for instance, the use of articles, or the tendency to "entify" characteristics or acts into things themselves by adding suffixes like "-ance," "-ity," "-ness," "-tion," "-age". In each case the result is the same: the linguistic difference corresponds to a difference in interpreting events, as measured by responses to simple surveys. Obviously all humans think alike to a large extent. But there are scientifically demonstrable differences, which are not academic but rather closely bound up with the interpretation of everyday events.

6.1.3. Contradictions and Loopholes

This brings us to another invalid argument often made against Whorf's ideas: that the very concept of linguistic relativity is self-contradictory. After all, it is asked, if our thoughts and perceptions are not based on objective reality but only on linguistic structures, then how can we trust those thoughts and perceptions that led us to the concept of linguistic relativity in the first place? Whorf is accused of asserting the objective truth of the impossibility of objective truth.

This argument is wrong for many reasons, the main one being that Whorf never actually made such a strong statement for linguistic determinism. He always left loopholes in his statements -- using "largely" instead of "entirely," and so on.

Statements which at first glance seem very strong become, on closer consideration, somewhat open-ended. For instance, consider Whorf's contention that

the world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds -- and this means largely by the linguistic systems of our minds. [p. 215]

Here there are two loopholes. First, "largely" -- what exactly does this imply? And then, "linguistic systems" -- given the concept of an abstract "language of thought," and the fact that Whorf has elsewhere called mathematics and music "quasilanguages," it is not clear exactly what this phrase is supposed to mean.

Whorf just plain never claimed that language **controls** thought, unilaterally and absolutely. And there is nothing paradoxical in the idea that linguistic structures are a **big influence** on our thoughts and perceptions. Even big influences can potentially be overcome -- with hard work and continual self-consciousness, or occasionally just by chance.

6.1.3.1. Language and Category

The misperception of Whorf as an extremist has caused many current researchers to distance themselves from Whorf, while at the same time applying many of his ideas. Listen, for example, to Searle (1983):

I am not saying that language creates reality. Far from it. Rather I am saying that **what counts as** reality -- what counts as a glass of water or a book or a table, what counts as the same glass or a different book or two tables -- is a matter of the linguistic categories that we impose on the world.... And furthermore, when we experience the world, we experience it **through** categories

> that help shape the experiences themselves. The world doesn't come to us already sliced up into objects and experiences; what counts as an object is already a function of our system or representation, and how we perceive the world in our experiences is influenced by that system of representation. The mistake is to suppose that the application of language to the world consists of attaching labels to objects that are, so to speak, self identifying. On my view, the world divides the way we divide it.... Our concept of reality is a matter of our linguistic categories.

Searle's emphasis on "categories" is reminiscent of Lakoff's (1987) *Women, Fire and Dangerous Things*, the title of which refers to an aboriginal language thatgroups women, fire and dangerous things together under one categorical name. It also reminds of Hilary Putnam's formal-semantic theorem, to the effect that

> 'Objects' do not exist independently of conceptual schemes. **We** cut up the world into objects when we introduce one or another scheme of description....

It has become acceptable in philosophical and anthropological circles to admit that language guides our **categorization** of the world. If Whorf were still around, how would he react to this? I suspect he would observe that categorization is just the **simplest** kind of patternment: that language does guide the way we group things together, but it also guides our perceptions and cognitions in subtler ways.

And Whorf might also be a bit amused to find the claim that "our concept of reality is a matter of our linguistic categories" in the same essay as the statement that "I am not saying language creates reality. Far from it." It would seem that contemporary thinkers like Searle find Whorfian ideas **useful**, but they want to avoid controversy by marking a sharp distinction between "our concept of reality" and "reality." What difference does this phenomenal/noumenal distinction make, in practice?

6.1.3.2. Whorf on Culture

So far I have defended Whorf against his critics. However, I must admit that on some issues Whorf went too far even for me. For instance, Whorf probably would not have agreed with the ideas about language and culture sketched in Section 2.7 above. He supposed that written and spoken languages, along with "quasilanguages" like music and mathematics, had a special power and coherence lacked by belief systems such as those inherent in culture. Regarding the interconnection between linguistic, social and psychological realms, he wrote:

> How does such a network of language, culture and behavior come about historically? Which was first: the language patterns or the cultural norms? In main they have grown up together, constantly influencing each other. But in this partnership the nature of the language is the factor that limits free plasticity and rigidifies channels of development in the more autocratic way. This is so because a language is a system, not just an assemblage of norms. Large systematic outlines can change to something really new onlyvery slowly, while many other cultural innovations are made with comparative quickness. Language thus represents the mass mind; it is affected by

inventions and innovations, but affected little and slowly, whereas TO inventors and innovators it legislates with the decree immediate. (p. 156)

Even the most unsophisticated reader would be unlikely to miss the ambivalence of this passage. In the beginning of the paragraph, "in the main they [language, culture and behavior] have grown up together." But by the end of the paragraph, language is "affected little and slowly," whereas language "legislates [to culture and behavior] with the decree immediate." Which is it? Is it coevolution between two systems of roughly equal complexity, or is it the **adaptation** of a relatively simple system to a much more complex one, with relatively little influence in the opposite direction?

In the end Whorf adopts what I would call a strict Darwinist point of view (see *The Evolving Mind* for a great deal more on strict Darwinism). Many evolutionary biologists believe that one cannot analyze evolution without taking into account the fact that the environment of an organism -- consisting as it does of other evolving organisms -- evolves along with the organism, adapting to the organism at the same time as the organism adapts to it. Some, such as James Lovelock (1988), even believe that the **physical** environment evolves to match the organisms which simultaneously evolve to match it. In contrast to these points of view, the strict Darwinists believe that **each organism** evolves independently, stringently influenced by the systematic structure and dynamics of its environment but having very little influence upon its environment. Whorf looks at cultural and behavioral patterns in the same way that strict Darwinism looks at organisms: helpless in the face of the awesome power of their environment, their only option is effective accomodation.

Unlike Whorf, I do **not** agree that cultural and behavioral systems are "just a collection of norms." Far from it. The whole field of social psychology speaks against this supposition. These systems are indeed a collection of norms, but a collection full of subtle interconnections and interdefinitions.

As to their effect on human existence, compared to the effect of language on human existence, here again I must differ with Whorf. Language's effect may be **subtler** and in some ways **deeper**, but the influence of cultural and behavioral systems is much more **direct**.

Spoken language encodes basic background assumptions that subtly guide our analogies. It thus plays a role throughout the mind -- in the language of Chapter Three, at virtually every level of the dual network, in virtually every cluster of processes (only the very lowest levels are exempt). But systems of other kinds guide our analogies as well, perhaps not quite so subtly or pervasively, but in many cases more powerfully. Belief systems about the nature of social and physical reality, or particular aspects thereof, guide our analogies very strongly.

And, finally, it is worth noting that even behavior systems can sometimes guide our cognitive processes. For when we adopt a certain role, put on a certain "performance," we associate things that we would not associate otherwise; and the mind is very good at recognizing and storing associations. This is a relationship which deserves much more attention than it has received.

6.2. LANGUAGE, CONSCIOUSNESS, SERIALITY

The dual network model, as outlined in Chapter Three, is a high-level "wiring diagram" for intelligent systems. But it sidesteps the question: where does consciousness fit in? In *The Structure of Intelligence*, consciousness is modeled as a process that moves from level to level of the multilevel control hierarchy, but only within a certain restricted range. If the zero level is arbitrarily selected to represent the "average" level of consciousness, then we may say consciousness resides primarily on levels from -L to U. The levels below -L represent perceptions that are generally **below** conscious perception. Consciousness is at a distance from the lowest levels of the hierarchy, which represent "sense data" -- it deals only with constructions of at least moderate complexity. And, on the other hand, the levels above U represent perceptions that are in some sense **beyond** conscious perception: too abstract or general for consciousness to encompass.

This theory of consciousness is similar in some respects to Jackendoff's (1986) "intermediate level" theory of consciousness, which states that consciousness corresponds to mental representations that lie midway between the most peripheral, sensory level and the most "central," thoughtlike level. Jackendoff points out that his idea

> goes against the grain of the prevailing approaches to consciousness, which start with the premise that consciousness is unified and then try to locate a unique source for it. [My theory] claims that consciousness is fundamentally not unified and that one should seek multiple sources. [p.52]

Consciousness is not in one place; it is rather associated with a collection of processes that occur in intermediate levels of the psychological hierarchy.

6.2.1. Dennett's Computationalist "Explanation"

We have **located** consciousness in the dual network. But we have not said **what** it is. What tasks does it accomplish, and what does it depend on? One intriguing hypothesis in this direction is supplied by Daniel Dennett, in his book *Consciousness Explained*.

A "meme" is defined as a sociocultural pattern, passed along from generation to generation. Dennett believes that consciousness is a meme rather than something intrinsic to the structure of the brain. He proposes that

> Human consciousness is **itself** a huge complex of memes (or, more exactly, meme-effects in brains) that can best be understood as the operation of a "**von Neumannesque**" [serial] virtual machine **implemented** in the **parallel architecture** of a brain that was not designed for any such activities. The powers of this **virtual machine** vastly enhance the underlying powers of the organic **hardware** on which it runs, but at the same time many of its most curious features, and especially its limitations, can be explained as the byproducts of the **kludges** that make possible this curious but effective reuse of an existing organ for novel purposes.

What is the intuition underlying this radical hypothesis? Thinking of the **streams of consciousness** that permeate James Joyce's fiction, Dennett gives this "von Neumannesque" serial machine the alternate label "Joycean machine." And, subjectively, in most states of mind at any rate, consciousness **does** seem to flow like a stream rather than an ocean: all in one direction, one thought after another.

"I am sure you want to object," Dennett writes, that "[a]ll this has little to do with consciousness! Afterall, a von Neumann machine is entirely unconscious: why should implementing it ... be any more conscious?" But this objection does not faze him:

> I do have an answer: The von Neumann machine, by being wired up from the outset that way, with maximally efficient informational links, didn't have to become the object of its own elaborate perceptual systems. The workings of the Joycean machine, on the other hand, are just as "visible" and "audible" to it as any of the things in the external world that it is designed to perceive -- for the simple reason that they have much of the same perceptual machinery focused on them.

Now this appears to be a trick with mirrors, I know. And it certainly is counterintuitive, hard-to-swallow, initially outrageous -- just what one would expect of an idea that could break through centuries of mystery, controversy and confusion.

In response to the question of what **good** this complex meme called consciousness does us, Dennett quotes Margolis (1987) to the effect that

> a human being ... cannot easily or ordinarily maintain uninterrupted attention on a single problem for more than a few tens of seconds. Yet we work on problems that require vastly more time. The way we do that ... requires periods of mulling to be followed by periods of recapitulation, describing to ourselves what seems to have gone on during the mulling, leading to whatever intermediate results we have reached.... [B]y rehearsing these interim results ... we commit them to memory, for the immediate contents of the stream of consciousness are very quickly lost unless rehearsed.... Given language, we can describe to ourselves what seemed to occur during the mulling that led to a judgement, produce a rehearsable version of the reaching-a-judgement process, and commit that to long-term memory by in fact rehearsing it.

This is nothing more than good common sense. It is well known that consciousness cannot contain more than around seven entities at one time. Therefore, **most** of the regularities present in the mind cannot enter directly into consciousness. But by use of language, complexphenomena can be encapsulated in simple statements, and thus presented to consciousness. If the unconscious "wishes" to present something to consciousness, it must translate some approximation of this thing into simple terms, let consciousness work with the simplified expression, and then afterwards translate **back**. Language is the number one tool for this kind of translation.

6.2.2. Consciousness, Virtual Seriality, and Language

The dual network is intrinsically parallel, but it is **possible** for a process or group of processes within the dual network to repeatedly feed itself its own output as input, thus creating a miniature virtual serial machine, temporarily ignorant of the massively parallel processing going on all around it. The dual network may in many cases connect A and B, and have A and B repeatedly exchange the results of computations without consulting any other processes -- **this** is virtual seriality, where one's "serial machine" consists of A and B together.

I don't completely buy Dennett's computationalist treatment of consciousness. However, I do agree with him that there is a very close connection between consciousness and virtual serial processing.

In Chapter Three we reviewed two important uses for virtual serial processing: making logical deductions, and predicting complex systems by simulation. A few pages above we discussed another, related use: general **linguistic** deduction. Subjectively, these actions are all closely connected with **consciousness**.

Margolus, in the quote given above, has eloquently presented the phenomenological case for the relevance of consciousness to linguistic deduction. In order to compute high-depth elements of D(I,T) for standard linguistic and logical systems, we need to use a complex combination of serial conscious thought and analogical/associative-memory thought. Introspectively, neither one process alone appears to suffice.

And the phenomenological connection between consciousness and prediction is no less direct. Suppose one wants to determine the likely consequences of a given action. One may **intuit**, in a semi-conscious flash, some guess as to the answer. But in order to be sure, one will reason it out slowly and carefully: what will be the immediate consequences, then the consequences of these consequences, and so forth. Almost all prediction is purely unconscious: but when situations get too uncertain, when they deviate too far from past experience, then consciousness has to intervene to dealwith things **serially**, by approximate simulation. In other words, walking down the street, one chooses a path unconsciously. But leaping through a stream from one rock to the next, one chooses one's path **consciously**, weighing each choice in terms of the array of future choices that it will lead to.

In sum, according to Dennett's "computationalist" vision, consciousness is a phenomenon

1) closely related with,

2) on the same levels as, and

3) dealing largely with the output of

serial, linguistic processing. This conception of consciousness is all that is necessary to fit the Sapir-Whorf hypothesis together with the pattern-theoretic analysis of language and mind. For it leads to the conclusion that **language helps to determine the world we consciously perceive**.

6.3 NIETZSCHE ON CONSCIOUSNESS AND LANGUAGE

Dennett's consciousness-as-meme idea is not a new one, nor is his picture of consciousness as linguistic deduction. His entire theoretical framework is, in fact, very similar to the view of consciousness articulated by Friedrich Nietzsche in 1882:

> ... Man, like every living being, thinks continually without knowing it; the thinking that rises to **consciousness** is only the smallest part of all this -- the most superficial and worst part -- for only this conscious thinking **takes the form of words, which is to say signs of communication**, and this fact uncovers the origin of consciousness.
>
> In brief, the development of language and the development of consciousness (**not** of reason but merely of the way reason enters consciousness) go hand in hand.... The emergence of our sense impressions into our own consciousness, the ability to fix them and, as it were, exhibit them externally, increased proportionately with the need to communicate them to **others** by means of signs...
>
> ... [C]onsequently, given the best will in the world to understand ourselves as individually as possible, "to know ourselves," each of us will always succeed in becoming conscious only of what is not individual but "average"...
>
> This is the essence of phenomenalism and perspectivism as **I** understand them: Owing to the nature of **animal consciousness**, the world of which we can become conscious is only a surface- and sign-world, a world that is made common.... (*The Gay Science*; 1968b)

Nietzsche interpreted the high degree of consciousness which we humans display as a socio-cultural phenomenon, an exaggeration of animal consciousness which evolved together with language -- which evolved, in short, as a **meme**. But his view of the **utility** of consciousness was not quite so rosy as Dennett's. According to Nietzsche, only conscious thinking is forced into the straightjacket of language, and for this precise reason conscious thinking is much less fertile than unconscious thinking. Language is for social interaction, therefore that which can be put in the form of language is precisely that which is **common** rather than that which is individual, unusual, unique.

Yet one cannot conclude that Nietzsche felt linguistic, conscious thought to be **unimportant** or **useless**. His attitude was much more complex than that. In a draft of a preface for his never-written treatise *The Will To Power*, he wrote "This is a book for thinking, nothing else." But in the notes for that very book, he wrote of thinking:

> Language depends on the most naive prejudices....
>
> **We cease to think when we refuse to do so under the constraint of language**; we barely reach the doubt that sees this limitation as a limitation.
>
> **Rational thought is interpretation according to a scheme that we cannot throw off**. (p.283)

This is about as Whorfian a statement as one could ever hope to find. Nietzsche valued linguistic, conscious, rational thought immensely -- for much of his life it was his only solace from physical suffering. But he did not trust it, he did not see it as objective; he refused to treat it as a religion.

6.3.1. Imaginary Subjects

Whorf's work focused on the differences in world-view implied by differences in linguistic structure. Nietzsche, on the other hand, saw certain very simple, very essential elements **in common** to all languages, andperceived that they played an essential role in the construction of the concept of an internal and an external world.

For instance, Whorf wrote of the way English, but not Hopi, refers to lightening as an object. Nietszche saw this objectification of non-objects -- crucial in the construction of the **external** world -- not as a peculiar feature of some languages, but rather as a consequence of the **one central** objectification involved in isolating the "self," the inner actor, as distinct from everything else.

Our bad habit of taking a mnemonic, an abbreviative formula, to be an entity, finally as a cause, e.g., to say of lightening "it flashes." Or the little word "I."

[H]itherto one believed, as ordinary people do, that in "I think" there was something of immediate certainty, and that this "I" was the given **cause** of thought, from which by analogy we understood all other causal relationships. However habitual and indispensible this fiction may have become by now -- that in itself proves nothing against its imaginary origin: a belief can be a condition of life and nonetheless be false. (p.268)

The self, the "I", is understood as the basis of the linguistic concept of **subject**, of **actor**. Thus the construction of a self, and the construction of an external world, are perceived as closely related, as emanating from the same fundamental principles. The concept of **subject**, in Nietszche's view, is a prime example of the subtle inter-connection of language and thought. Our language assigns imaginary subjects to actions, and we correspondingly assign imaginary subjects to actions in our conscious and near-conscious thinking; we construct an external world based largely on subjects. And we postulate an imaginary entity called **I**, and attribute to this subject a host of actions that are actually due to the independent and interactive behavior of a number of different subsystems.

These "imaginary" subjects may be understood as the result of an overextended analogy. First, events are correlated with other temporally prior events -- e.g. smoke is correlated with fire. Then, it is observed that in many cases it is useful, and hence satisfying, to explain a **large number** of different events in terms of **one** temporally prior entity. General concepts like"weather," "hatred," "patriotism," and so forth arise, each one out of the desire to explain a certain collection of effects with one entity. These concepts refer to definite collections of specific phenomena; they are simply tools for thinking and remembering.

But then what happens is that, when something cannot be explained in detail, a general concept is adduced as an "explanation." This is not always a mistake: given limited resources, a mind cannot explain everything in detail. It must learn to recognize which things can be explained in terms of well known ideas, and can be ignored until the pressing need to analyze them arises, and which things are anomalous, requiring special attention so that trouble will not occur when the need to analyze them arises. But it **is** a mistake sometimes: a general concept is adduced as an explanation for a phenomenon to which it simply does not apply. Thus "it flashes" for lightening.

"It bit me" is meaningful, it is a general explanation which could easily be backed up by a detailed explanation. But "it flashes" is not: this is a general explanation which is really unrelated to any detailed explanation. The only possible related detailed explanation would be of the form "this and that combination of atomspheric phenomena flashes" -- but that is severely stretching the concept of **it**, and in any case it is not the sort of explanation that would come naturally to the mind of a non-meteorologist. "I did it" is problematic for the same reason "it flashes" is no good. It is not just a shorthand for some detailed explanation ready at hand, it is an empty abstraction.

P.T. Geach, in *Mental Acts*, has made this point in a particularly eloquent way:

> The word 'I', spoken by P.T.G., serves to draw people's attention to P.T.G.; and if it is not at once clear who is speaking, there is a genuine question 'Who said that?' or 'Who is "I"?' Now, consider Descartes brooding ... saying 'I'm getting into an awful muddle -- but then who is this "I" who is getting into a muddle?' When 'I'm getting into a muddle' is a soliloquy, 'I' certainly does not serve to direct Descartes' attention to Descartes, or to show that it is Descartes, none other, who is getting into a muddle. We are not to argue, though, that since 'I' does not refer to the man Rene Descartes it has some other, more intangible thing to refer to. Rather, in this context the word 'I' is idle,superfluous, it is used only because Descartes is habituated to the use of 'I' in expressing his thoughts and feelings to other people.

According to Whorf, this reification of the subject does not happen in Hopi and other non-Indo-European languages. But on this point I must side with Nietzsche. The grammatical manifestation of reification may vary from language to language, but I very strongly suspect that every language postulates **some** form of imaginary acting entity. This, unlike use of counterfactuals, emphasis on flux versus stasis, and other linguistically varying phenomena, is absolutely essential to the concept of language. It is an instinctive application of analogical reasoning to the act of **naming** on which all communication is based, and no culture can escape from it. Humans cannot help but attach a certain amount of concrete reality to the symbols that they use. We can, as Nietzsche suggested, **fight** this tendency, but this is a battle which no one can ever completely win.

An interesting spin-off of this analysis of imaginary subjects is the theory that free will is an **emotion** inspired by **language**. Nietzsche's analyzed free will as

> the expression for the complex state of delight of the person exercising volition, who commands and at the same time identifies himself with the executor of the order -- who, as such, enjoys also the triumph over obstacles, but thinks within himself that it was really his will itself that overcame them. In this way the person exercising volition adds the feelings of delight of his successful executive instruments, the useful 'underwills' or undersouls -- indeed, our body is but a social structure composed of many souls -- to his feelings of delight as commander. **L'effet c'est moi**: what happens here is what happens in every well-constructed and happy commonwealth; namely, the governing class identifies itself with the successes of the commonwealth. (1968, p.216)

The feeling of free will, according to Nietszche, involves 1) the feeling that there is indeed an entity called a "self", and 2) the assignation to this "self" of "responsibility" for one's acts.

In *The Structure of Intelligence*, "delight" and related emotions are given a pattern-theoretic treatment. Following Paulhan, happiness is analyzed as the feelingof increasing order, increasing interemergence and interconnectedness. Here, let us focus instead on the **nature** of the delight involved. Free will is the special kind of happiness derived from a process attributing the successes of its "servant" processes to **itself** -- in other words, it is an example of the joy of making the postulate of an imaginary subject. And this postulate is **linguistic** in nature, so that the connection between free will and consciousness is precisely as close as the relation between language and consciousness.

6.4. A NEW THEORY OF CONSCIOUSNESS

So far, I have discussed some of the **correlates** of consciousness; but I have not explained consciousness itself. To get at the true nature of consciousness, one must confront the feeling of "raw existence" or "self-presence" that is the essence of what we call **living**.

This is a very difficult task, and I will approach is obliquely, by first looking at consciousness is through the medium of **biology**. The biological approach cannot give us the final answer to what is fundamentally a **psychological** problem. But it will be remarkably useful in setting us on the right path.

6.4.1. Consciousness as Perception

Consciousness is **self-perception**. And self-perception could, theoretically, be achieved in two ways. First, by special "perception" routines used only for perceiving high-level mental activities. Or second, by general "perception" routines that are also used for something else. **Evolutionary** thinking makes the second possibility seem far more attractive.

For, suppose the first alternative holds. These special self-perception routines would have to be quite sophisticated. How would they ever get started, in the natural history of the brain? Clearly, in their initial stages, they could have no adaptive advantage. They would have to arise as the **side-effect** of something else. But what?

The second alternative, on the other hand, requires no mysterious "evolution out of the blue." Lower animals demonstrate progressively more sophisticated neural routines for **perceiving the outer world**. If consciousness uses **these** routines for self-perception, then its evolution is not so much of an enigma. All that the evolution of consciousness required was the additionof some new connections onto a complex, fine-tuned, already existing mechanism.

The most reasonable hypothesis, therefore, is that consciousness is the result of taking neural maps normally used for perceiving the outside world, and applying them, not to the external stimuli for which they were intended, but to the inner workings of the mind. Of course, the **lowest levels** of perceptual processes cannot possibly be applied outside of the context for which they evolved. But for slightly higher levels, this is not true. What about the processes that assemble various pictures together into a **scene**? What about the processes that distinguish meaningful sounds or images from background information that is less relevant or interesting. These are highly developed aspects of the human perceptual mechanism.

What I am suggesting is that consciousness works by **mapping higher-level thought processes into middle-level sensory data**. Consciousness consists of "fooling" the perceptual mechanism into thinking it is working with constructs built up directly from external sense data, when it is actually working with **transformed versions** of patterns from levels above it. This explains what we mean when we say we are "thinking visually" about something, or "thinking in words." We mean that our self-perception uses the **standard perception routines** of the brain, which evolved for perception of data coming in from particular sense organs: eyes, ears, noses, taste buds, skin. Our ideas are mapped into pictures, sounds, perhaps even smells, and **in this disguise** they are grouped into wholes and "perceived." Then the perceptions obtained in this way give rise to higher-level patterns, which may be fed back down to the perceptual mechanisms, repeating the process and giving rise to the familiar **circularity of consciousness**.

This view is fairly closely related to Edelman's (1989) theory of consciousness. According to Edelman, consciousness represents the interaction between

1) the recognition of patterns in "interoceptive input," input from neural maps gauging the state of the **body**. This categorization is mediated by the hypothalamic and endocrine systems, the "reptile brain"

2) the recognition of patterns emergent between "interoceptive input" and "exteroceptive" input. Exteroceptive input, input from outside the body, is mediated by hippocampus, septum and cingulate gyri; the recognition of emergent patterns takes place in the thalamus and cortex.

The interaction between these two processes is a kind of "re-entry" between higher-level **cognitive** emergent-pattern recognition and lower-level "automatic" interoceptive and exteroceptive pattern recognition.

However, while Edelman explores many interesting neurological details, he omits any detailed discussion of the intuitive, **psychological** role of perceptual mechanisms in consciousness. The issue of "fooling," and its relationship to the subjective experience of consciousness, is never drawn into the picture. Thus, on a **psychological** level, Edelman's theory of consciousness is somewhat disappointing, particularly in comparison to his Neural Darwinist theory of learning, which is so suggestive both biologically and psychologically.

Finally, it is worth pointing out that none of this contradicts Dennett's "consciousness-as-meme" idea. I have said that there are neural connections leading from higher-level processes, through transformation processes into middle-level perceptual processes. These connections have **evolved**; they are there in every human brain. But they may be **strengthened** through repeated use, or weakened through disuse. Coming into frequent contact with other conscious persons would seem to be a prerequisite for the **strengthening** of these connections. In this sense, therefore, consciousness may be said to be a "meme." The **presence** of the connections is genetic, but their strength is memetic.

6.4.2. Consciousness and the Making of Reality

Now, finally, I am ready to put all the pieces together: consciousness, language, seriality, thought, and perception. The first step in this unification is to do what neither Dennett, Jackendoff or any other modern cognitive scientist has done: to say **what good** consciousness is. I propose, following Nietzsche, that **the function of consciousness is to manufacture reality**.

Consciousness is a feedback dynamic involving higher-level "cognitive" processes and middle-level perceptual processes. What I am suggesting is that a pattern only acquires the presence, the solidity that we call "reality," if it has repeatedly passed through this feedback loop.

Philip K. Dick defined reality as "That which doesn't go away when you stop believing in it." Reality is a kind of near **imperviousness** to mental dynamics, a refusal to be altered by the natural re-organization processes of the dual network. The dual network constantly readjusts itself, swapping one subnetwork foranother in quest of greater associativity and fortuitous genetic creation. But those subnetworks which are **real** cannot be broken up; their pieces cannot be swapped for other pieces.

To put it metaphorically, elements of reality are like islands in the sea of mind. As with real islands, a sufficiently large storm can maul or bury them: there are degrees of restriction. But normal weather patterns rearrange the sea and leave the islands intact.

Why would passing through the feedback loop from higher-level to middle-level tend to cause relative **imperviousness**? The answer to this lies in the specific middle-level perceptual processes involved. These are, I suggest, primarily

1) those processes which act to combine a group of different sensations from the same sense organ together into a single cohesive entity -- a "scene," "image," "sound," "physical location," etc.

2) those processes which act to combine entities recognized by different senses (hearing, vision, touch, etc.) into a single, united form.

Each time something is passed through these processes, it attains a degree of cohesion, a degree of resistance to being broken up. When something is passed through again and again and again, it achieves a superlative degree of cohesion and resistance -- it becomes **real**.

The process of grouping disparate elements together into a whole is a complex one. However, I suggest that one key **part** of this process is an **increase in the degree of restriction** against rearrangements. A subnetwork which cannot easily be disrupted by rearrangement dynamics is inherently much **wholer** than one which can. And once it is protected against rearrangement, its parts have the leisure to slowly adjust themselves to one another, thus attaining yet more refined wholeness. Finally, passing some X through the restriction-degree-increase routines over and over again would obviously result in the construction of extremely solid barriers around that X.

In this view, consciousness is a **serial** process. And it is very similar to the serial processes of prediction, logical deduction, and syntactic sentence, percept-, or act-formation. All of these processes involve a re-entry from higher to lower. Something is **built up** -- a phrase, say, out of words; or a future, out of the present. And then it is **passed down** to the level where its parts came from: the phrase is plugged into a syntactic operation as if it were a word; the futurescenario is treated conjecturally as a present and the mental routines for "present-world" manipulation are applied to it.

But mere similarity is not the only relation between consciousness and deductive, serial processing. Perhaps more crucial is the fact that **in the context of the dual network, structured transformation systems require the interim asusmption of reality every step of the way**. How could deduction work if one step were altered before the next were complete? How could prediction work if the one-week prediction were rearranged before the two-week prediction was done? How could a complex sentence be formed if, while the sentence was being structured on a global level, the subservient phrases of the sentence were being replaced with phrases of completely different types? The re-entrant processes involved in applying structured transformation systems require **reality** to be introduced at each step. And reality, I have argued, requires consciousness.

This, I suggest, is the true nature of the relationship between consciousness, language and thought. Language structures the memory which guides the structured transformation systems of deductive and predictive thought. But neither sentence formation nor deduction nor prediction could function without consciousness.

6.4.2.1. Consciousness as Catch-22

Nietzsche lamented the "coarseness" of the ideas contained in consciousness. But this is inevitable: it is in the very nature of consciousness to construct ideas that are **rigid**. Unconscious ideas are bound to be more fluid, more adept at intuitive shifting. But most of these unconscious ideas were constructed by **structured transformation systems**, which require local rigidity for their effective operation.

Specifically, **imaginary subjects**, which annoyed Nietzsche so, are precisely the price one pays for having linguistic systems that talk about subjects. Without reifying things, without assuming and **imposing** their reality, there is no way to keep them solid in the midst of the shifting dynamics of the mind; there is no way to keep them in one place long enough to work with them. Sometimes the reification turns out to be a little too much -- "I" or "lightning" are reified for one purpose, and then used for another. But the mind is notoriously error-prone; it is a strict adherent to Murphy's Law. The cost of avoiding this type of error would be great asto make thought impossible. Consciousness, and reification along with it, are necessary components of the unconscious creativity which Nietzsche so extolled.

On the other hand, it would be just as futile to lament the unconsciousness of most of the mind. If **everything** were made conscious, the mind would freeze up, it would grind to a halt. Structured transformation systems, which are the main reason consciousness is **necessary**, also require associative memory, which is maintained only by the **fluidity** of subnetworks that have not been made real through consciousness.

Thus consciousness represents a sort of psychological Catch-22. In order to produce fluidity, the mind must produce rigidity. And in order to produce rigidity, the mind must produce fluidity. The two exist in a careful balance; one cannot abolish one without abolishing the other as well.

6.4.3. Consciousness and Self-Reference

Beginning from considerations loosely biological in nature, I have arrived at a novel **psychological** model of consciousness, expressible solely in terms of the dynamics of the dual network. The feeling of "raw existence," I suggest, is simply the feeling of **subnetworks resisting the natural urge to shift**. It is the feeling of **solidity resisting fluidity**.

And the feeling of "self-presence" is one level up from this; it is the feeling of **solidity which produces solidity**. "I am" means "I, this mental process, make myself solid; I maintain my boundaries against the surrounding flux." This is not merely an egotistical delusion -- one may formally show that a mental process **can** make itself solid, by containing a subroutine directing itself down through the feedback loop of reality-construction. A process can self-referentially direct itself to the grouping, solidifying centers of the mind.

One way to write such a process is:

X = s, and direct X to the nearest solidifying process, please

Here s is any object of observation; one may omit it, and obtain a process which does nothing but direct itself.

Or, less formally, one may write

X = s and look at X,

reducing to

X = look at X

in the simplest case, or e.g.

X = I am hungry and look at X

in a more general situation.

In later chapters I will have much more to say about self-referential formulae of this type, and their validity in psychological modeling. It will be formally demonstrated that such self-referential constructions can be elements of mind. For now, however, it is enough to suggest that there is a fundamental importance attached to the self-propelled movement of such processes through the feedback loop of consciousness. **This motion**, I claim, is self-awareness.

6.4.4. Conclusion

This, finally, completes our roundabout excursion into the murky waters of consciousness theory. The theory presented in this section may be understood on two levels: biological and psychological. Some of the neurological details have been fairly speculative; all of the biological statements I have made, however, are testable scientific hypotheses. Once we finish charting the connections of the brain, we will see exactly what sort of re-entry consciousness involves. If it involves re-entry into some sort of **scene-making** or **cross-modally connecting** perceptual process, then the biological theory of this section will be proved correct. If not the theory will have to be modified, or perhaps discarded.

On the other hand, the dual network model very strongly suggests that, whatever the biological details, the **psychology** of consciousness is one of **iteratively strengthening barriers against reorganization**. This is the only logical role for consciousness in the context of a continually fluctuating network of mental processes. So, from the point of view of the dual network model, the barrier-strengthening would have to be accepted even if it did **not** have interesting implications. But in fact it does have at least one very interesting application: it **explains, from first principles, the dependence of language and reason on consciousness**.

Whorf, Dennett and Nietszche, despite their vastly different theoretical perspectives, have one important thing in common: they essentially **equate** consciousness with language and deductive reason. But this is notsatisfactory; there is a sense in which consciousness is more basic, less complex. These other processes make **use** of the inherent nature of consciousness, but do not **define** it. The view of consciousness as iterative barrier-strengthening lets one **deduce** the close connection between consciousness, language and reason, rather than assuming it.

Recall that, at the start of the chapter, I decomposed the Sapir-Whorf hypothesis into two separate hypotheses: 1) that the structure of language strongly influences the structure of thought; 2) that the differences between existing languages are sufficiently great to cause significant differences in thought patterns. I have said nothing new about the second claim. What I **have** done, however, is to **derive** the first claim from basic properties of the dual network model. Whorf liked to use the word "pattern"; it was essential to his thought. So it is not terribly surprising that, in developing a pattern-theoretic model of mind, I have "rediscovered" an **abstract** version of Whorfian linguistics.

Chapter Seven

SELF-GENERATING SYSTEMS

In his recent book *Self-Modifying Systems in Biology and Cognitive Science* (1991), George Kampis has outlined a new approach to the dynamics of complex systems. The key idea is that the Church-Turing thesis applies only to simple systems. Complex biological and psychological systems, Kampis proposes, must be modeled as nonprogrammable, self-referential systems called "component-systems."

In this chapter I will approach Kampis's component-systems with an appreciative but critical eye. And this critique will be followed by the construction of an **alternative** model of self-referential dynamics which I call "self-generating systems" theory. Self-generating systems were devised independently of component-systems, and the two classes of systems have their differences. But on a philosophical level, both formal notions are getting at the same essential idea. Both concepts are aimed at describing systems that, in some sense, **construct themselves**. As I will show in later chapters, this is an idea of the utmost importance to the study of complex psychological dynamics.

7.1 COMPONENT-SYSTEMS

A component-system, as defined by Kampis, consists of a collection of components, each of which can act on other components to produce new components. More precisely,

An abstract component-system can be defined by the following properties:

a) - there is a finite set of non-dividable and permanent building blocks, drawn from a given pool

b) - there is an open-ended variety of the different types of admissible components, built up from the building blocks according to some composition rule (which may be explicit or implicit)

c) - the components of the system are assembled and disassembled by the processes of the system such that every admissible component is also realizable. (p.199)

For illustrative purposes, Kampis suggests that the reader visualize the "non-dividable building blocks" as LEGO blocks, and the "admissible constructions" as different possible structures buildable out of LEGO blocks. One must merely imagine that each LEGO structure contains some appropriate means for acting on other LEGO structures to produce new LEGO structures.

The main biological example of a component system is a "molecular soup" full of organic molecules acting on one another to form new molecules. Psychologically, on the other hand, one is supposed to think of ideas acting on each other to produce new ideas. The central thesis of *Self-Modifying Systems* is that **biological and psychological systems, being component-systems, are fundamentally uncomputable**. This thesis combines two distinct claims:

Claim 1: Formal component-systems display uncomputable behavior.

Claim 2: Formal component-systems are good models for biological and psychological systems.

The first claim is a **mathematical** result, which Kampis calls his "Main Theorem." The second claim, on the other hand, is obviously a **scientific** hypothesis.

In this section I will explore these two claims in some detail. This exploration will lead us to a new class of systems called **self-generating systems** -- a class of systems which is different from, but overlapping with, Kampis's class of component-systems. The contrast between self-generating systems and component-systems will shed a great deal of light on the fundamental issues of system theory.

7.1.1. Quantum and Stochastic Computation

Before pursuing Kampis's main thesis in any more detail, I will first explore the meaning of the term "computable." My attitude toward computation has been influenced immensely by David Deutsch's (1985) work on quantum computation. Deutsch has demonstratedmathematically that **any system modelable by the equations of quantum physics can be simulated to within arbitrary accuracy by a "quantum computer"**. A quantum computer is different from an ordinary Turing machine. However, it cannot compute any functions besides those which an ordinary Turing machine can compute. Deutsch's "Quantum Church-Turing Thesis" states that every physically realizable algorithm can be represented as a program for a quantum computer. In fact, this is not really a thesis but a theorem. In this respect it is far more impressive than the ordinary Church-Turing Thesis.

There is also another Church-Turing Thesis, intermediate between the standard one and the quantum version. One may define a **stochastic computer** as a Turing machine which is capable of doing "random coin tosses." Then the Stochastic Church-Turing Thesis states that every algorithm can be represented as a program for a stochastic computer. Deutsch has shown that stochastic computation is a less general model than quantum computation; and I will make use of this result now and again in the following.

Kampis's proof of the uncomputability of component-systems -- Claim 1 above -- says nothing about quantum or stochastic computers. It speaks only of Turing machine computation. In the following I will argue that this omission is important -- that Kampis's component-systems, although they are not Turing computable, may sometimes be computable by **stochastic** computers. Because stochastic computation is a less general model than **quantum** computation, this implies that at least some component-systems are explicable in terms of quantum physics.

One necessary requirement of any theory of complex systems is **agreement with microscopic physics**. Those component-systems which are not quantum computable, are in contradiction to the principles of physics. What this means is that, in a physical sense, the class of component-systems is **too broad**.

Actually, there is a hole in this argument -- a tiny hole, but one which must be duly noted. "Agreement with microscopic physics" is not strictly synonymous with "agreement with quantum physics." In his best-seller *The Emperor's New Mind*, Roger Penrose briefly discusses Deutsch's theorem, but he dismisses it on the grounds that quantum mechanics will soon be replaced by a unified theory of **quantum gravity**. The unified theory of quantum gravity, Penrose conjectures, will imply that computable systems are fundamentally uncomputable in the strongsense of being able to compute non-Turing-computable **functions**.

The weak point of Penrose's argument, however, is that none of the existing approaches to quantum gravity show any promise of implying uncomputability. For instance, string theory (Green et al, 1987) is similar to quantum theory in its general mathematical form -- it depends on the "quantization" of a classical domain using the Feynman path summation formula. So, if some form of string theory is correct, then it would seem that there is no hope for Penrose's idea.

7.1.2. Kampis's "Main Theorem"

We have broken down Kampis's central thesis into two claims. The first of these, the "Main Theorem" of *Self-Modifying Systems*, is as follows:

Main Theorem. In a component-system it is not possible to know the names and the encoding (the meaning) of the names before the system produces the respective components.... The behaviour of component-systems is fully uncomputable and unpredictable because the produced new observables are different from the earlier ones.

The basic idea here is that the temporal sequence of states of a component-system is in general an uncomputable sequence. Since no Turing machine program can generate an uncomputable sequence, component-systems must be uncomputable.

It seems to me that the key point is clause (c) of the definition, which says that "every admissible component is realizable." Suppose one assumes that the set of all admissible components is uncomputable, and that the dynamics of a component-system are capable of leading to **any** admissible component. Then it follows logically that the dynamics of a component-system cannot be specified by any program. For, if one assumed the opposite, one would obtain a contradiction -- one would have an uncomputable set of entities obtainable from a computer program.

Let us go back to the LEGO metaphor. It would be easy to build a **computable** LEGO universe following Kampis's instructions. For the set of all LEGO structures is countable, and may therefore be mapped into the set of binary sequences, in a one-to-one manner. And each binary sequence may be represented as a Turingmachine program, i.e. as a map from binary sequences to binary sequences. Therefore, using Turing machines, each LEGO structure could be interpreted as a function acting on other LEGO structures. The only problem with this arrangement is that it does not satisfy clause (c) of the definition of component-system. Not every LEGO structure is realizable by our dynamics. Only some computable subset of LEGO structures is realizable.

But now -- and here is where my thinking differs from Kampis's -- suppose one adds a **random** element to one's Turing machine. Suppose each component of the Turing machine is susceptible to errors! Then, in fact, **every possible LEGO structure becomes realizable!** Structures may have negligibly small probability, but never zero probability! This is an example of a component-system which is computable by a **stochastic Turing machine**.

Deutsch has shown that quantum Turing machines are more general than stochastic Turing machines. So what I have shown is that component-systems are perfectly realizable in terms of the equations of quantum mechanics. This implies that there is absolutely no problem with the statement that "molecular soups" or brains are component-systems. But there **is** a problem with the statement that these systems are "fundamentally uncomputable." The Turing model of computation is not in general physically adequate. But quantum computation, something quite similar to Turing computation, **is** always physically adequate, at least so far as our present knowledge of physics goes. And something that is quantum computable is not, in a philosophical sense, fundamentally uncomputable.

7.1.3. Self-Constructing Robots

To put these ideas in sharper focus, let us now turn to a metaphor which Kampis introduces around the middle of the book: the **self-constructing robot**. This idea is a natural extrapolation of modern industrial technology.

Right now, in Japan, there are robotized factories -- factories in which routine assembly-line tasks are carried out by robots rather than people. These are not humanoid robots like C3PO in Star Wars. They look like what they are: sophisticated factory tools. But their capabilities are astounding -- they combine the spatial common sense of a human worker with the speed and precision of a calculator. In fact, it is not unlikelythat, somewhere in Japan, there are detailed plans for a factory in which robots are used to build more robots.

And, of course, the industrial use of robots is not restricted to manufacturing. It is well within contemporary technology to use robots for **repair**. It is not yet profitable to use robots to repair robots, but this is because of simple technical problems, not fundamental engineering obstacles.

The point of all this is: if a robot can repair other robots, why not itself? And if a robot can repair itself, why not reconstruct itself, even when it is not broken? It is not too far beyond current technology to build a robot that **reconstructs itself**. There is no reason not to build a robot whose software (brain) tells it how to reconstruct its hardware (body, including brain). Such a self-constructing robot would embody an enchanting sort of loop: self constructing new self, which constructs new self, which.....

But finally, suppose that someone builds a self-constructing robotw which is, however, **imperfect**. It sometimes makes slight random errors; its arms don't always move quite exactly the way they are supposed to. Then in classic chaotic form, as time goes by, these slight random errors can be expected to build up into large errors. One has a fundamentally unpredictable sequence of machines. There is no telling exactly what the robot will make of itself, given say fifty years time.

To me, a self-constructing robot with errors seems like a wonderfully creative thing! But Kampis's argument is precisely the opposite. In one particularly striking passage, Kampis characterizes a component-system as "a strange computer in which also the software is identified with the hardware." Elaborating, he declares that a component system is a computer which, when executing its operations (software) builds a new hardware.... [W]e have a computer that re-wires itself in a hardware-software interplay: the hardware defines the software and the software defines new hardware. Then the circle starts again. (p. 223)

To me, this sounds exactly like the self-reconstructing robot which I have just finished talking about. But Kampis has something else up his sleeve. He does not believe in the Church-Turing thesis. He believes that a "component-system" is nonprogrammable, inthe sense that no algorithm, no set of rules, can completely describe its behavior. He follows up the previous quotation with a warning to the computationally-inclined:

[A] sceptical reader could say: that's not a big deal. With current-day industrial robot technology this should be possible. Robots are automata; they are computers. They can assemble other robots, maybe even themselves. They have a complete behavior algorithm. So, by analogy, component-systems, too, can have one.

But this is not as easy a matter as it sounds. In a robot the whole software is ready-made and completely defined from the beginning on, and is stored in an accessible form; in a component-system, according to the above story, the "algorithm" is nowhere stored completely; software and hardware define each other without any of them being complete or independent.

The paradigm case of a component-system, according to Kampis, is a "soup" of organic cells. Each cell acts on each other cell, thus creating other cells, and there is no distinction between software and hardware.

But let us consider, once again, our **imperfect self-constructing robot**. This robot is programmed to modify its own hardware, but it is susceptible to random error. Then it is quite possibly true that **no computer program can predict the behavior of the robot**. For the collection of all possible times and places for random error is very large, and the collection of **sets** of times and places for random error is even larger. To predict the behavior of the robot, a computer program would have to predict what would happen to the robot **given each possible set of random errors**. But, for any program of finite length, there is some set of random errors which cannot be compressed into any program of that length.

The moral of the story is that, in the case of the self-reconstructing robot, stochastic computation does what Turing computation does not. It gives the potential for true flexibility; for self-referential creation of the fundamentally, indisputably new. While component-systems cannot be Turing computable, they **can** be stochastically computable. This observation casts a revealing light on the distinction between component-systems and Turing machines.

7.1.4. Creativity

To put the same point another way, I respectfully accuse Kampis of having an overly mystical notion of **creativity**. He complains that computer programs can never create anything beyond what has been put into them -- a very old argument. This is true in the same sense that mathematical theorems are never original creations -- they are all contained in the basic axioms of mathematics. But, even if one accepts this strict notion of creativity, it **still** does not follow that stochastic computers are noncreative.

In general, a stochastic computer has a range of output that is incredibly wide, often uncomputable. In fact, one may very easily construct a stochastic computer that has the capability to construct **anything whatsoever**, just by chance. So stochastic self-reconstructing computers suffer from no lack of potential creativity. How much of this creativity is actualized depends on the intricate interaction of the deterministic and stochastic components.

This brings us to a basic principle of systems theory: **the essence of creativity is the interplay between rules and randomness**. This ancient concept, which received its modern form in the work of Ross Ashby (1954), is one of the most humanly meaningful implications of the computer revolution. It is humbling to realize that even the most marvelous works of the greatest geniuses -- Einstein's General Theory of Relativity, Goethe's *Faust*, Beethoven's Fifth Symphony -- were produced by a complex combination of random chance with strict, deterministic rules. Kampis does not wish us to accept this. But if one is to accept that modern physical science applies to neural processes, then one must, I suggest, accept the equation of creativity with quantum computation.

As an afterthought, it is worth briefly questioning the role that random chance plays here. From the point of view of any **one** computer -- be it Turing, quantum or stochastic -- there are certain **deterministic** sequences of events that are fundamentally indistinguishable from **random** sequences of events. These are sequences whose algorithmic complexity exceeds the algorithmic complexity of the computer who is doing the distinguishing. Gregory Chaitin (1974, 1987) has shown that this statement is essentially a form of Godel's Incompleteness Theorem.

So, from any one subjective point of view, there is no way of telling if some perceived entity is **stochastically** computed, or just plain **Turing** computed. Now, although component-systems are not Turing computable, we have seen that they can be stochasticallycomputable. It follows that, from any one subjective point of view, a component-system might as well be Turing computable! To any particular entity, "random" just means "too complex for me to understand."

7.2 SELF-GENERATING SYSTEMS

Nowhere in *Self-Modifying Systems* does Kampis give an adequate formal definition of "component-system." To my mind, this is the only sizeable flaw in an otherwise outstanding book. This omission is particularly crucial in that it makes it difficult to mount conceptual attacks against Kampis's **nonprogrammability thesis**.

Kampis gives a fairly good reason for this significant omission:

> [C]oncepts of formal dynamics do not fit well to component-systems.... [W]hen we consider component-systems as systems which produce components from components from components, we may, by the same token, think of transformations producing directly other transformations: $\mathbf{f_t}$**:** $\mathbf{f_t}$ **-->** $\mathbf{f_{t'}}$.

> There is a formal problem with this idea. From a mathematical-logical point of view **no mathematical function can belong to its own domain or range**. However, the functions that describe component-systems try to do exactly this, if we take them literally. (p.212)

But in fact, two hundred pages later, Kampis admits that this complaint is not strictly accurate. He refers to Lofgren's (1968) demonstration that the existence of functions belonging to their own domain or range is **independent** of the ordinary axioms of set theory.

As it happens, Lofgren is not the only researcher to point out the existence of set theories in which functions **can** belong to their own domain or range. From my point of view, Paul Aczel's **AFA** axiom provides a much more elegant approach to constructing such unusual functions. So, let us digress for a few paragraphs to describe the AFA axiom.

7.2.1. Hypersets

In mathematics, one defines complex concepts in terms of simpler concepts. But this process must bottom out somewhere -- there must be something that is **sosimple** that there is nothing simpler in terms of which to define it. In modern mathematics, this elementary concept is usually taken to be the "set." The term "set" may be defined **intuitively**, as a collection of entities -- but of course, this is circular, since what is a "collection" if not a "set"?

Mathematicians take this intuitive definition of a set, and then postulate certain simple rules for dealing with sets. All the complex constructions of modern mathematics can be expressed in terms of sets, and a great many of the theorems of modern mathematics can be proved by using the rules of set theory. However, in the 1930's Kurt Godel showed that, given any particular list of rules for manipulating sets, there are some mathematical theorems that can be expressed in the language of sets, but cannot obtained by using the rules on the list. Essentially this is because each list of rules has a certain finite amount of "algorithmic complexity", and cannot be used to prove theorems that possess an algorithmic complexity in excess of this amount.

At first, mathematicians were very loose about what qualified as a set. They dealt with finite sets like {1,2,3}, infinite sets like (1,2,3,...}, the set of all fractions, or the set of all numbers on the number line; and also with abstract sets far more esoteric than these. But since the concept of set never caused them any trouble, they had no motivation to fiddle with it.

But then, around the turn of the century, Bertrand Russell noted a problem. He said, consider the set containing all sets that do not contain themselves. And he asked: does this set in fact contain itself? The trouble is, what if it **does** contain itself? Then it is not a set that does not contain itself -- so it cannot be an element of itself, it cannot contain itself. But, on the other hand, if it **does not** contain itself then it must be an element of itself, since it is the set of all sets that do not contain themselves. A serious problem!

Incidentally, for years this "Russell Paradox" has been formulated as follows: "There is a town in which the barber shaves all men who do not shave themselves. Who shaves the barber?" However, someone has wittily pointed out that this version is less potent than the original. The solution is: the barber is a woman!

In order to avoid Russell's Paradox, a rule was added onto the original axioms of set theory: no set can contain itself as an element. More generally, there can be no "descending chain of membership" -- no set can Acan contain as an element a set which contains A as an element, and so on.

Many mathematicians were uncomfortable with this rule, since it was not "natural", it was simply appended onto the list of rules in order to avoid contradiction. And this discomfort became especially acute after Godel came out with his Incompleteness Theorem. For Russell's Paradox is essentially a variation on the old Paradox of Epiminides the Cretan: "This sentence is false." But Godel's Theorem is an even cleverer variation on this same ancient paradox. Godel showed that, by implementing a clever scheme of coding, one can use any mathematical system to form the proposition X = "This proposition cannot be proved true or false within this mathematical system." If X can be proved true, then it must be false, in which case it is true, so it cannot be proved true. But if X can be proved false, then it must be false, so it has been proved true, and has therefore not been proved false.

Godel, with his ingenious Incompleteness proof, showed that self-reference cannot be banned from mathematics anyway, no matter how hard you try. This made Russell's elaborate theory of Types seem even more excessive than it had before. But still, since mathematicians never seemed to have any use for sets that contained themselves as elements, they simply accepted the axiom and went on doing mathematics.

However, while working with various models of complex systems such as ecosystems and brains, I found that I **did** have a need for sets that contained themselves as elements. I spent a long time trying to concoct ways of avoiding this problem. But then, while sightseeing in Cambridge and browsing through the MIT Bookstore, I came across a little book by Paul Aczel, entitled *Non-Well-Founded Sets*. This book describes a research programme in mathematical logic, active since the late 1960's, aimed at constructing a consistent set theory involving sets that can contain themselves as elements.

The most easily applicable result of this intriguing research programme is the concept of a **hyperset**. A **labeled graph** is defined as any collection of dots with a symbol drawn next to each dot, and arrows drawn between the dots. Aczel's "AFA Axiom" implies that every finite graph corresponds to some set. For instance the graph

Subjectivism

A Objectivism

Mysticism

corresponds to the set A = {objectivism, subjectivism, mysticism}. And the graph

Subjectivism

A Objectivism

Mysticism

Botulism

Aneurism

corresponds to the set A = {{objectivism, subjectivism, mysticism}, botulism, aneurism}.

But what of the graph

A ?

This graph corresponds to the set A whose only element is A -- the set A = {A}. This is a "non-well-founded" set.

And, similarly, the graph

A botulism

objectivism

subjectivism

corresponds to the set A = { A, botulism, {objectivism, subjectivism, A} }.

To the mathematically indoctrinated mind, all this is incredibly liberating! Just as the paradoxes of quantum physics free the mind from objectivism, so hyperset theory frees the mind from the stifling preconception that, if A contains B, B cannot in turncontain A. Common sense tells us that, if mind is a part of physical reality, then physical reality cannot possibly be a part of mind. And common sense tells us that, if **you** are a part of **my** subjective reality, then **I** cannot possibly be a part of **your** subjective reality. But hyperset theory tells us that in this case common sense is wrong.

According to Godel's Theorem, once can never mathematically prove that a complicated mathematical theory is **consistent**, devoid of self-contradictions. But Aczel has shown that, if there are contradictions in hyperset theory, then there are also contradictions in plain ordinary mathematics, the kind that every scientist uses to make calculations. This is as good a consistency result as one could hope for. One may confidently say: there are mathematical objects that contain one another as elements.

In the following pages, I will not need to use any of the technical mathematics of hyperset theory. However, I **will** find it convenient to talk about sets that contain one another in a "circular" way. Hyperset theory ensures that this is okay, that I am not contradicting myself any more than a physicist is when he deals with algebraic or differential equations.

This section began, if you recall, with Kampis's observation that "from a mathematical-logical point of view **no mathematical function can belong to its own domain or range**. However, the functions that describe component-systems try to do exactly this, if one takes them literally." It is easy to see that with hypersets, mathematical logic has transcended the limitation to which Kampis is referring. One particular type of hyperset is the **hyperfunction** -- the function which is contained in its domain or range.

A hyperfunction maps hyperfunctions and/or other entities into hyperfunctions and/or other entities. Because it is a "function," it is not allowed to map any one thing into two different things. To deal with the more general case, I will introduce the term **hyperrelation**. A hyperrelation maps hyperrelations and/or other entities into hyperrelations and/or other entities; and it may map one thing into as many other things as it likes.

These odd constructions, hyperrelations, are the first step on the path toward a cognitive equation. For they give us a straightforward way to talk about components that truly **transform** one another.

7.2.1.1. A More Formal Treatment (*)

Recall that, in order to avoid Russell's Paradox, a rule called the Axiom of Foundation was added onto the original axioms of set theory: no set can contain itself as an element. More generally, one is not allowed to have an infinite descending sequence

$$\ldots a_{n+1}\ a_n \ldots a_1\ b$$

A set b which contains no such sequence is **well-founded**. All the traditional sets of mathematics -- the sets involved in geometry, calculus, topology, etc. -- are well-founded. But, for example, S = {S} is not well-founded, because it leads to the infinite descending sequence

... S S ... S S

And { 1, { 1, { 1, ... is not well-founded, even though it has the natural "solution"

x = {1,x}

Many mathematicians were uncomfortable with this rule, since it was not "natural", it was simply appended onto the list of rules in order to avoid contradiction. But since they never had any use for sets that contained themselves as elements, they simply accepted the axiom and went on doing mathematics.

Paul Aczel (1988) was one of the few who decided to do something about his discomfort. He constructed the "non-well-founded sets" which, following Jon Barwise and Larry Moss (1991), I have called hypersets. According to Aczel's approach, the path to hypersets begins with graphs. A **digraph** (G,E) consists of a set G of entities called "nodes," and a set E of ordered pairs of nodes, these pairs being called **edges**. The most common examples of graphs are finite graphs, as in Figure 1; however, the concept of an infinite graph presents no difficulties. If (n,m) is an element of E, I will write n-->m, and call m the **child** of n, and n the **parent** of m. Fix a set A of tags. Then a **tagged digraph** (G,E,t) is a digraph together with a function t that assigns a tag drawn from A to each childless node of G.

Next, define an **accessible pointed graph (apg)** (G,E,t,p) to consist of a tagged digraph together with a distinguished node p which has the property that every node can be reached by some finite path from p. And define a **decoration** of an apg as a set-valued function d with domain G, satisfying

$$d(n) = t(n)$$

if n is childless and

$$d(n) = \{d(m):n\text{-->}m\}$$

otherwise. That is, a decoration assigns to each childless node its tag, and to each parent node n those nodes m which are its children.

Finally, let us say that an apg **pictures** a set b if there is a decoration d of the graph so that d(p)=b; that is, so that b is the set which decorates the distinguished node. This permits us to state Aczel's Anti-Foundation Axiom (AFA), which characterizes hypersets:

Every apg pictures a unique set.

According to this definition, all the sets of standard set theory are still sets. But there are other sets too. Anything which is a set according to **this** definition, but not the classical definition is a **hyperset**. For example, consider again the following graph:

A

There is only one node, so it must be the distinguished node. What is the unique set pictured by this apg? It must be the set A = {A}! However, according to the ordinary axioms of set theory, no set can contain itself.

Hypersets can contain themselves. One might at first think that this would lead to contradictions, but Aczel has shown that if ordinary set theory is consistent, so is set theory augmented by AFA. In addition, he has proved a very useful result called the **Solution Lemma**. Roughly speaking, the Solution Lemma states that every system of equations in indeterminates x,y,z,..., say

$x=a(x,y,...)$

$y=b(x,y,...)$

...,

has a unique hyperset solution.

For a deep mathematical treatment of hypersets, the reader is referred to Aczel's (1988) original monograph on the subject. However, the clearest discussion of the fundamentals of hyperset theory which I have found is ina delightful little book by Jon Barwise and John Etchemendy entitled *The Liar* (1988). This book contains a more rigorous statement of the Solution Lemma.

As an example, let us construct a function which maps itself into itself -- a function so that $f = f(f)$. If f takes no other arguments besides itself, then by the standard definition $f = \{ (f,f) \} = \{ \{ f, \{f\} \} \}$. f is then the solution of the system of equations

$w = \{f\}$

$x = \{f,w\}$

$f = \{x\}$

Graphically, one has

w f

x

More generally, if $f=f(y)$ and if is only defined on y, f is the solution of

$w=\{f\}$

$x=\{y,w\}$

$f=\{x\}$

And it is clear that, in a similar manner, one may cast any system of expressions of the form

$f_1(f_1,...,f_n,x_1,x_2,...) = f_{i(1)}$

... **(*)**

$f_n(f_1,...,f_n,x_1,x_2,...) = f_{i(n)}$

in the form required by the Solution Lemma and thus obtain a hyperset solution.

Getting back to *Self-Modifying Systems*, I suggest that component-systems should be conceptualized in terms of systems of hyperfunctional equations of the form **(*)** given above; and, more generally, **hyperrelational** equations defined in a similar way.

7.2.1.2. Fuzzy Hypersets (*)

Although fuzzy sets are now commonplace in artificial intelligence, so far as I know fuzzy **hypersets** have never before been discussed. Fortunately, therewould appear to be no particular problems involved with this useful idea: the basic mathematics of fuzzy hypersets, at least as far as I have worked it out, is completely straightforward.

The simplest example of a fuzzy **hyperset** is the set x defined by:

$d_x(x) = c$,

$d_x(y)=0$ for y not equal to x.

Here, if c=0, one has an ordinary well-founded set, namely the empty set. If c=1, one has the set x={x}. Otherwise, one has something **inbetween** the empty set and x={x}.

Each fuzzy hyperset is characterized by a **fuzzy** apg, which is exactly like an apg except that each link of the graph has a certain number in [0,1] associated with it. The **Fuzzy AFA** then states that each fuzzy apg corresponds to a unique fuzzy set. It is easy to see that the natural analogue of the Solution Lemma holds for fuzzy hypersets. And, of course, the consistency of fuzzy hypersets with the axioms of set theory (besides the axiom of reducibility) follows trivially from the fact that each fuzzy hyperset is, in fact, a hyperset under AFA.

7.2.2. Self-Generating Systems

Kampis's examples of component-systems are both relevant and elegant: mobile interattracting LEGO blocks, enzyme systems, self-modifying robots,.... However, for reasons given above, I do not find his formal **definition** of component-system entirely satisfactory. Thus I think it is worthwhile to define a closely related type of system called a "self-generating system."

A self-generating system, at each time, consists of a collection of components which are modelable as "finitely given hyperrelations" -- meaning that they are defined by their actions on a finite number of different possible components. Each component may be thought of as having a certain **degree** of membership in the system, with the constraint that the total degrees of all the components should be **finite**.

Self-generating dynamics is defined as a two-stage process. First, **universal action**: each component acts on each other component with a certain probability, yielding different new components with different probabilities. Then, **transformation**: these resultant components are transformed in some way, yielding a new collection ofcomponents. The results of transformation are then fed back into the first step, and used as fodder for universal action.

The transformation rule may be stochastic -- for instance, it may make **errors**. It may change f into w by mistake 4% of the time. Or, on the other hand, it may simply make a **random addition or deletion** to the definition of each component a certain percentage of the time. Given f which does not act on g at all, it may **randomly define** the action of f on g, or it may define the action of f on g by some complex formula combining deterministic and stochastic elements. By use of randomness, the transformation rule could generate dynamics that are fundamentally unpredictable, in the sense of being non-Turing-computable.

The connection between self-generating systems and component-systems is quite simple. Not every component-system is a self-generating system; but I propose that every **physically useful** component-system is actually a self-generating system. Note that, since the definition of self-generating systems is phrased in terms of hyper **systems**, it is perfectly natural for the number of components to shift in the course of evolution.

Now, the converse is not true: not every physically useful self-generating system is a component-system. The reason is that the self-generating dynamical equation is capable of describing totally deterministic processes. Component-systems can only be obtained if the function R is allowed to contain random elements, although rather good simulations of component-systems can be obtained using chaotic or "pseudorandom" deterministic functions. Thus the class of component-systems and the class of self-generating systems possess a nontrivial intersection. I propose that real complex systems lie in this intersection.

7.2.2.1. A More Formal Treatment (*)

Define an hyperrelation to be **finitely given** if its associated apg is finite, and the labels of its associated apg are all encodable as finite binary sequences. Line up the set of all finitely given hyperrelations in some arbitrary order: f_1, f_2, f_3,.... Given this ordering, an **hypersystem** may be defined as an infinite vector C = (p_1,...,p_n,...), where the p_i are nonnegative and the sum p_1 + p_2 + ... + p_n + ... is finite. (In functional-analytic lingo, the set of hypersystems is therefore isomorphic to the space l_1+).

The entry p_i is to be interpreted as the **degree** with which f_i belongs to the hypersystem represented by C. For instance, in the context of enzyme systems, p_i would denote the **concentration** of the enzyme f_i in the solution in question. In the **deterministic case**, an important but special situation, all the p_i are assumed to be 1 or 0.

Also, I will use the notation $x_i t$ to denote "inanimate objects": entities which can be acted on, but cannot act. I will not refer to these very often in the following, but they may be useful in certain applications.

To each hypersystem System*t*, associate a set of "action products"

$$A[\text{System}_t] = \{f_{it}(f_{jt}),\ f_{it}(f_{jt},x_{kt})\}$$

Here the indices i and j run over all hyperrelations that have a nonzero probability of membership in System_t.

Next, define a "filtering operation" Y_t which determines, based on the degrees of the elements in System_t, the **degrees** of the elements in $A[\text{System}_t]$. The only restriction is that, if either f and g have degree zero in System_t, then Y_t cannot assign f(g) or f(g,x) a nonzero probability. In the deterministic case, Y_t is only a formality, for it may be assumed that all probabilities are either 1 or 0. But in the most extreme fuzzy case, it may come about that A is the formality, leaving Y_t to do most of the work.

Finally, one may collapse these two operations into one composite operation $R[\text{System}_t] = Y_t(A[\text{System}_t])$ -- the "Raw Potentiality" operation. Then, where T is some stochastically computable function mapping hypersystems to hypersystems, one may define a **self-generating system** as an iteration

$$\text{System}_{t+1} = T(\ R[\text{System}_t]\ ), \qquad \textbf{(**)}$$

Here the hypersystem System_1 is considered to be given a priori, yielding a **dynamic iteration**.

7.2.3. Hypersets and Physical Reality

One would like to think of component-systems and self-generating systems as models of real physical complex systems. At first sight, however, there seems to be a serious obstacle in the way of this interpretation. Hyperrelations are peculiar set-theoretic objects. Formally, in Aczel's construction they are defined asequivalence classes of sets of ordinary sets (this construction is somewhat analogous to the construction of real numbers as equivalence classes of Cauchy sequences of rationals). This places them in a cardinality class far, far above the countable computable sets, and also far, far above the Hilbert- space-defined sets which quantum mechanics associates with physical reality.

However, interpreted properly, a finite system of **finitely given hyperrelations** does not violate the stochastic Church-Turing thesis, since any such system of equations can be **simulated** on a stochastic computer. A stochastic computer can never actually contain hyperrelations, but if they are finitely given, it can simulate their behavior easily enough. After all, manipulations with finitely given hyperrelations are merely manipulations with finite graphs!

Physically, what does this mean? While quantum physics does not permit the existence of physical hypersets, it **does** permit physical events that are effectively modeled as finite labeled graphs. Now, suppose that the interactions of some of these physical events can be modeled as interactions between finite labeled graphs, and that these graph interactions are usefully describable using self-generating systems or component-systems. Then these mathematical systems are **emergent patterns** in physical reality. No contradiction. No problem. Physical reality can simulate component-systems; or, to put it another way, the reality of component-systems can be understood as a "virtual reality" running on the hardware of quantum reality.

In sum, I suggest that Kampis's picture of complex system behavior is fundamentally right. Complex systems consist of components that act on one another to create new components. Thus they effectively violate the hierarchy of logical types; they contain emergent patterns which are usefully modelable in terms of stochastic systems of hyperrelations.

But, on the other hand, as already mentioned, Kampis's theory of creativity is flawed. The creativity of a complex system is due both to the unfolding of the rules implicit in its components, and to the mutation of these rules by random error. The opposition which Kampis has set up, between computation and component-systems, is in my opinion a false one. The difference between simple systems and complex systems is not that the former are computable, but that the latter contain **emergent structures** which are modelable in terms of stochastic hyperfunctional iterations (self-generating systems). The concept of self-generating systems makes this point in a very clear way.

7.2.3.1. A Binary Model of Hypersets (*)

To make this conclusion more concrete, let me construct a specific computational scenario which gives rise to hypersets as a natural model. Suppose one has a finite collection of computable relations f, g, h,..., each of which maps **binary sequences** into **binary sequences**. Then one may represent each relation by a finite code sequence, e.g.

s_f = 010100101111010010...01

s_g = 010111101001001010...10

....

And one may define the action f(g) as the result of the following two-step process:

1) letting the **program** f act upon the **sequence** s_g, producing a new sequence s_{fg}.

2) Decoding s_{fg} into a program, by selecting the **first** (in alphabetic order) from among all programs h for which the Hamming distance d(s_h,s_{fg}) is at its absolute **minimum**. This "first closest" program is taken as f(g).

The relations f, g and h are computable relations -- but there is aboslutely nothing wrong with thinking of them as hyperrelations, acting directly on each other. The whole system may be elegantly modeled as a system of hyperrelations, without ever referring to the underlying bit-string manipulations. This requires no new information, only a shift in point of view. I will refer to this system for deriving hyperrelations from computable relations as the **basic computational model**.

7.3 MAGICIANS AND ANTIMAGICIANS

Coauthored with Harold Bowman

Our treatment of self-generating systems has up to this point been purely formal. However, one may also describe self-generating systems in much a less mathematical way, in the guise of **self-referential sentences** (Hofstadter, 1985). For example, suppose that the complete formal definitions of the hyperrelations f, g and h are given by:

f(f) = g, f(g) = h, f(h) = f, f(g,h) = g

g(g) = g, g(f) = h, g(f,h) = f

h(f,g) = h, h(g,h) = g, h(h) = f

This looks terrible. To make it a little prettier, let us rename "f," "g," and "h" as "Fanny," "Geronimo" and "Hattie." Then one may represent this same collection of definitions as follows:

This sentence, which is named Fanny, turns itself into Geronimo, it turns Geronimo into Hattie, it turns Hattie into itself, and it turns the pair 'Geronimo and Hattie' into Geronimo.

This sentence, which is named Geronimo, turns itself into itself, it turns Fanny into Hattie, and it turns the pair 'Fanny and Hattie' into Fanny.

This sentence, which is named Hattie, turns itself into Fanny, it turns the pair 'Geronimo and Hattie' into Geronimo, and it turns the pair 'Fanny and Geronimo' into Hattie.

This says the same thing as the preceding group of mathematical definitions, but it is a little more colorful. The best way to visualize the situation is to think of Fanny, Geronimo and Hattie as a group of three over-active magicians. Each one has a spell to turn each one into someone. For instance, Fanny has a spell to turn herself into Geronimo; she has a spell to keep Hattie the same as she was, and the has a spell to turn the **combined group** Geronimo/Hattie into Geronimo only.

Recall that a self-generating system is a **stochastically computable rule** for evolving populations of finitely given hyperrelations. This is a very general definition; it leaves a lot of freedom. For starters, therefore, let us consider the simplest possible situation. Given a collection of hyperrelations {f,g,h,...}, one can form a vast variety of "compounds" of the form f(g), f(g,h), h(g), and so forth. One of the very simplest self-generating systems says: "given a collection of hyperrelations, replace it with the collection of all compounds which one can form from it."

For instance, one may think about self-generating systems in the context of our earlier Fanny/Hattie/Geronimo example. Each magician had spells for changing magicians (or group of magicians) into othermagicians. So the **simplest** self-generating system involving these magicians consists of

1) each magician applying all the spells she knows, thus creating a new collection of magicians

2) the new collection of magicians then applying all the spells they know

3) etc.

It is easy to see that, in the case of our simple example, **if one starts with all the magicians present**, then all the magicians are so proficient that the group of three magicians will persist forever. If one starts with just Fanny and Geronimo, they will immediately produce Hattie, and the same will be true. Similarly, if one starts with just Fanny and Hattie, they will immediately create Geronimo. But on the other hand, if one starts with Geronimo and Hattie, the two of them will never be able to produce Fanny. Or if one starts with just Fanny, she will immediately turn herself into Geronimo, who will then perpetuate herself forever....

A group of somewhat less proficient magicians is provided by the following set of rules:

f(f) = g

f(g) = f

g(g) = g

g(f) = h

h(f) = g

If one takes

{f,g} time = 1

this rule produces the collection {f(f),f(g),g(f),g(g)} = {g,f,g,h} = {f,g,h}, or

{f,g,h} time = 2

And iterated once again, the rule produces {f(f),f(g),f(h), g(f),g(g),g(h),h(f)}, or

{f,g,h} time = 3

In this particular case, after two steps, our dynamical rule has reached a fixed point. No matter how many times one keeps iterating, one will keep on obtaining {f,g,h}.

In "magician-language", what one has here is

This sentence, whose name is Fanny, turns itself into Geronimo and turns Geronimo into Fanny.

This sentence, whose name is Geronimo, turns itself into itself and turns Fanny into Hattie

This sentence, whose name is Hattie, turns Fanny into Geronimo

Starting from Fanny and Geronimo, as above, one will immediately get all three magicians.

More generally, a self-generating system is simply a rule which determines a "range" collection of hyperrelations from the **compounds** formed by another "domain" collection of hyperrelations. In these simple examples, I have taken collection of compounds itself as the range collection. But this is not the only way to do things. As an example, one could consider the rule: "given a collection of hyperrelations, replace it with two hyperrelations **randomly drawn** from the collection of all compounds which one can form from it." The reader may determine for herself the possible evolutionary courses which our collection {f,g} may take under this rule.

7.3.1. Antimagicians

Both of the "magician" examples discussed above were of the simplest kind. All possible compounds were generated and kept. However, this type of self-generating system does not appear to be capable of generating particularly complex behaviors. To get the full range of dynamical behaviors, one must provide some way for compounds to **eliminate** one another. For instance, in addition to our three faithful magicians Fanny, Geronimo and Hattie, one may introduce three **antimagicians**, called anti-Fanny, anti-Geronimo and anti-Hattie. And one may modify the rules of our game accordingly. At each time step, the following three processes are executed:

first, all magicians cast all their spells

second, all magicians whose anti-magicians have been created are eliminated

third, all anti-magicians are eliminated

For instance, suppose one has

This sentence, named Fanny, turns itself into Geronimo, turns Geronimo into Hattie, and turns Hattie into anti-Fanny

This sentence, named Geronimo, turns itself into Fanny, turns the pair 'Fanny and Geronimo' into Hattie, and turns Hattie into anti-Hattie

This sentence, named Hattie, turns itself into Hattie, turns Fanny into Hattie, and turns Geronimo into Fanny

Without anti-magicians, Hattie would be self-perpetuating. But the anti-magicians change all that. Suppose one starts with

Geronimo and Hattie **time 1**

Then this evolves into the interim population

Fanny, Hattie, anti-Hattie

The magician and its anti-magician self-destruct, leaving

Fanny **time 2**

Then Fanny creates Geronimo, who creates Fanny, who creates Geronimo, and so on ad infinitum...

Geronimo **time 3**

Fanny **time 4**

Geronimo **time 5**

Fanny **time 6**

Geronimo **time 7**

Fanny **time 8**

... **...**

On the other hand, suppose one starts with

Fanny, Geronimo **time 1**

Then one has

Geronimo, Fanny, Hattie **time 2**

and from this one gets the interim population

Fanny, anti-Fanny, Geronimo,

Hattie, anti-Hattie

resulting in

Geronimo **time 3**

and yielding the same attracting cycle as before:

Fanny **time 4**

Geronimo **time 5**

Fanny	**time 6**
Geronimo	**time 5**
Fanny	**time 6**
...	**...**

This sort of behavior, with Hattie and Fanny appearing and then disappearing, is much much easier to set up with anti-magicians than without then. I will show a little later exactly **how much** more computational power is yielded by the introduction of anti-magicians.

To get even more interesting behavior, one must **stochasticize** the dynamics. For example, one could replace the three processes of our "antimagician" iteration with the following three processes, to be executed at each time step:

first, each spell of each magician is cast with a certain fixed

probability (call it p)

second, all magicians whose anti-magicians have been created are eliminated

third, all anti-magicians are eliminated

This creates an unpredictable iteration: if one runs it several times, one may obtain many different results, because there is no telling which spells will be chosen. The reader is encouraged to explore the consequences of "stochasticizing" our Fanny/Geronimo/ Hattie example.

7.3.2. Self-Generation and Computation

So self-generating systems can be considered as models of real systems. But what kind of behavior can they model? In fact it is not too hard to prove that they can model **any kind of behavior at all**. Harold Bowman and I have constructed a very simple argument which shows that self-generating systems are capable of universal computation. This means that any possible behavior can be mimicked by some self-generating system.

Specifically, as hinted above, it turns out that simple systems of the Fanny/Geronimo/Hattie variety are not enough. One needs to introduce anti-magicians as well. But if one does this, then one very easily obtains a recipe for constructing a self-generating system tosimulate any given computer. The basic idea is that systems with anti-magicians give one the ability to express the two fundamental operations of **conjunction** (AND) and **negation** (NOT). Since all computers can be built of AND and NOT gates, it follows (with a little work) that this type of system is a universal computer.

The AND gate is easy; it can be done without anti-magicians. Let's say one wants Fanny to create Josie only if both Geronimo **and** Hattie are present. Then one needs merely say

This sentence, named Fanny, turns the pair 'Geronimo and Hattie' into Josie

By simply **not specifying** Fanny to turn Geronimo individually or Hattie individually into Josie, one makes Fanny into an AND function.

Of course, in the context of the whole system, it is possible that Geronimo or someone **else** will turn Geronimo or Hattie into Josie -- but if one wants Fanny to be a good AND, one must design one's system to prevent this from happening at the same time that Fanny is operating as an AND.

Incidentally, it is worth noting that

This sentence, named Fanny, turns Fanny into Fanny, turns the pair 'Geronimo and Hattie' into Josie, and turns Hattie into Geronimo

serves the purpose of executing the AND operation as well. Extra spells are allowed, so long as they do not interfere.

To get NOT, on the other hand, one must proceed as follows. Let's say one wants Fanny to create Hattie only if Geronimo is **not** present. Then one needs to specify

This sentence, named Fanny, turns itself into Hattie, and turns Geronimo into anti-Hattie

If Geronimo is not present, then Fanny produces Hattie, so Hattie is introduced into the next population (assuming no one else is out there producing anti-Hatties). But if Geronimo **is** present, then Fanny still acts on itself to produce Hattie, but it also acts on Geronimo to produce anti-Hattie. The two cancel out, and one is left with no Hattie (assuming no one else is out there producing Hatties).

7.3.3. Imperfectly Mixed Computation

The biggest lesson of the computer revolution is that by piecing ANDs and NOTs together one can do just about anything. The reasoning of the past few paragraphs, elaborated appropriately, leads to the consequent conclusion that **self-generating systems** (in particular "antimagician" systems") can do just about anything.

But this is just the tip of the iceberg. The next question is, **what about *stochastic* "antimagician" systems?** What if, at each stage, only a certain percentage of possible compounds are formed? This is the case, for example, in real chemical solutions: not every conceivable compounds forms at every moment. In chemical parlance, deterministic self-generating systems correspond to infinitely "well-mixed" solutions, whereas stochastic self-generating systems correspond to the more realistic case.

It turns out that, even in the stochastic case, it is possible to construct "antimagician" systems which carry out **universal computation** -- to within any specified level of accuracy short of perfection. The trick is an appropriate use of redundancy. For instance, if instead of just doing NOT with one hyperfunction, one does it simultaneously with a sufficiently huge number of hyperrelations, one is bound to get it right an arbitrarily high percentage of the time.

What this result shows is that one can build a viable computer in which each connection between components has only a certain probability of existing. One may build a computer out of "components" that circulate around each other, sometimes combining with one another to produce new components, sometimes not. Computation can **self-organize** from an imperfectly-mixed-up substrate.

Applied to biological and psychological systems, this conclusion would seem to have profound consequences. The dual network model views the mind as a collection of **processes**, interacting with one another and constantly creating new processes. The ideas of this sections suggest that these general "processes" may perhaps be fruitfully modeled as interlocking self-referential statements -- as simple statements about how other processes, and they themselves, are to be transformed. This is an intriguing insight, and an important step on the path to the "cognitive equation" of Chapter Eight.

7.4. ARRAY COMPONENT-SYSTEMS (*)

In Section 7.3 I gave a simple "reductionist" model of hyperset dynamics -- the "basic computational model." In this section I will briefly digress to describe a more interesting elaboration of the same fundamental concept. Instead of mapping functions into sequences in an arbitrary way, I will demonstrate how one might elegantly systematize the coding and decoding of functions and sequences.

7.4.1. Array Operations

Let us begin with the concept of a **rational array**. An n-dimensional rational array may be defined inductively as a finite sequence of (n-1)-dimensional rational arrays, where a 1-dimensional rational array is just a finite sequence of rational numbers. Our eight basic operations will be operations on rational arrays.

The most relevant examples are one, two and three-dimensional arrays; however, it is quite possible to envision psychological uses for arrays of higher dimension. For instance, spacetime is four-dimensional, and a five-dimensional array could therefore represent a scalar field over spacetime, an eight-dimensional array a vector field over spacetime, etc.

Each rational array A comes equipped with a natural **coordinate system**, so that each rational stored in A has a unique coordinate vector $(a_1,...,a_n)$, a_i a nonnegative rational. This coordinate system imposes a natural **alphabetic order** on the elements of A, which one may extend to subsets of A by defining subset B to come **before** subset C if B - C contains a point which comes before any point in C - B in the alphabetic ordering.

Human sensory inputs can be expressed very naturally in terms of rational arrays. For instance, light on the retina forms a two-dimensional array; and sound waves on the eardrum form a one-dimensional array. Muscle movements can also be easily expressed in terms of rational arrays: when one has different amounts of stimulus sent to different points, the different points can be envisioned as elements of a three-dimensional rational array. And, to take a biological example, the interactions between proteins can also be effectively expressed in this way: the **surface** of a protein is just a rational array.

In general, any continuous field can be approximated to within arbitrary accuracy by an rational array. For example, if one wants to approximate a field on the positive hyperoctant of R*n* with 5 digits of accuracy, one may divide the positive hyperoctant of R_n up into alatticework of cubes of side 10_{-6}, and construct an n-dimensional rational array containing one element for each cube.

Of course, given sufficiently complex codings, one can dispense with the whole formalism of rational arrays and consider only binary sequences. But here I am not thinking in terms of abstract algorithmic information, I am rather thinking in terms of concrete information processing systems, for which "sufficiently complex codings" can present formidable practical obstacles.

7.4.1.1. Pointwise Operations

The first four of our nine operations are **addition, multiplication, negation, and maximization**, which are defined pointwise. More explicitly, let A and B be two rational arrays. Then the sum of A and B is A+B, the product of A and B is written AB, the maximum of A and B is written A^B, and the negation of A is written -A. If a has coordinates $(a_1,...,a_n)$ in A, and a' has coordinates $(a_1,...,a_n)$ in B, then,

a+a' has coordinates $(a_1,...,a_n)$ in A + B,

aa' has coordinates $(a_1,...,a_n)$ AB,

max(a,a') has coordinates $(a_1,...,a_n)$ in A^B, and

-a has coordinates $(a_1,...,a_n)$ in -A.

if A and B are of different sizes, then $(a_1,...,a_n)$ is assumed to exist in A + B, AB and A^B only if it exists in both A and B.

It is worth noting that this collection of operations is redundant in two ways. One the one hand, by combining negation and maximization one can generate any Boolean function, and thus any computable function, including addition and multiplication. Secondly, by combining multiplication and addition, one can generate any polynomial, and hence approximate any continuous function, including the maximum function, to within arbitrary accuracy. However, our goal here is not to give a minimal set of operations; it is to give an exhaustive set of **basic** operations.

7.4.1.2. Combinatory Operations

Addition, multiplication, negation and maximization are all **pointwise** operations. Now I will introduce two operations that act on whole arrays rather than on an entry-by-entry basis.

First, the **cut-and-paste** operator is a ternary operation which may be written C(A,B,S), where A and B are general rational arrays and S is a sequence of nonnegative integers. The expression C(A,B,S) is to be read: paste B into A, placing the first entry in B into position S in A.

More explicitly, what this means is as follows. Suppose S = (s_1,...,s_k). Then if a has coordinate (s_1 - r_1,...,s_k - r_k) in A, where the r_j are all nonnegative integers, then a has the same coordinate in C(A,B,S). But if b has coordinate (s_1 + r_1 - 1,...,s_k + r_k -1) in B, where the r_j are all nonnegative integers, then b has the same coordinate in C(A,B,S).

For instance, C((1,2,3,4,5,6,7), (9,9,9,9,9,9,9,9), 5) =

(1,2,3,4,9,9,9,9). The number 5 indicates that the elements 5-8 of the sequence B = (9,9,9,9,9,9,9,9) are pasted onto the elements 1-4 of the sequence A = (1,2,3,4,5,6,7).

Cut-and-paste also permits us to build higher-dimensional arrays out of lower-dimensional ones. For example, one has

C((1,2), (3,4), (2,1)) = 1 2

3 4

The coordinate (2,1) gives the point at which the array (3,4) is "pasted" onto the array (1,2).

In general, as the name suggests, cut-and-paste permits us to form new arrays by combining parts of different old arrays.

Next, the **reduce** operation allows one to take part of an array and consider it as an array in itself. This is of obvious utility as an adjunct to the cut-and-paste operation: it allows one to paste in parts of arrays rather than just whole arrays. The simplest way to define the reduce operation is as R(A,S,T), where A is an arbitrary rational array and S and R are lists of nonnegative integers, with the property that the i'th entry in S never exceeds the i'th entry in S. Write S = (s_1,...,s_n), R = (t_1,...,t_n). Then R(A,S,T) is the array composed of all elements of A whose coordinates lie "between" the arrays S and T. Explicitly, if a has coordinate (a_1,...,a_n) in R(A,S,T), this means that a has coordinate (a_1+s_1-1,...,a_n+s_n-1) in A, and $a_i + s_i < t_i$.

Finally, **substitution** is a ternary operation which may be denoted by S(A,B,C), to be read: substitute A for B, everywhere B appears in C. The meaning of this isobvious in simple cases, for instance S(6,(3,4),(1,2,3,4,3,4,5,3,4)) =

(1,2,6,6,5,6).

In general, there is an ambiguity here: what if two appearances of B overlap in C? However, this can be resolved by the following rule: if there are two appearances of B in C, and it is not possible to substitute A for both of them, substitute A for that appearance of B which occurs **first** in C.

Note that substitution is a special instance of cut-and-paste. However, it is a very important instance. A large percentage of the patterns that we recognize are **repetitions**. For instance, the whole Behaviorist school of psychology is based on the recognition of repeated stimulus-response associations! As we shall see, repetitions can be easily expressed in terms of substitution.

A special case of substitution is "change of notation." For instance, S(2,3,A) is the operation of replacing every 3 in A with a 2. The inclusion of substitution as a basic operation guarantees that no specific notation or "encoding" is essential to human thought.

7.4.1.3. Random Generation

Our eighth operation, **random generation**, is the simplest of all. It may be defined as R(A), where A is a one-dimensional integer array. Its function is to create a random array of dimensions given by A, whose entries each have an equal probability of being either zero or one. This operation could be simulated fairly well in terms of the other operations, by using standard pseudo-randomness techniques; but for theoretical purposes I prefer to introduce true stochasticity.

The choice of a 50% chance of a 0 or 1 in each entry is purely a matter of convention. Using the operation R(A) together with the previous seven operations, one may construct arrays so that each entry a has a different probability p_a, and one may choose the p_a's to be arbitrarily close to any number in [0,1].

7.4.1.4. Decoding

Finally, let us consider the operation of **decoding**. This is in a way the most fundamental operation of all. Let us assign each one of our fundamental operations a **code number**, according to the following arbitrary scheme:

addition = 1

multiplication = 2

negation = 3

maximization = 4

substitution = 5

cut-and-paste = 6

decoding = 7

random generation = 8

Next, let us arbitrarily assign the number 9 to stand for "open parenthese," and the number 10 to stand for "close parenthese." Finally, the integers greater than 10 will be understood to denote **variables**. Since the substitution operator is fundamental, the specifics of the encoding do not matter; they can always be renamed.

Given this encoding, many integers sequences may be understood as **sequences of operations**.

This prepares us to define the decoding operator. Where A is a **sequence** (a one-dimensional integer array), and $B_1,...,B_k$ are arbitrary rational arrays, $D[A,B_1,...,B_k]$ is the array obtained by applying the sequence of operations encoded in A to the arrays $B_1,...,B_k$, where the variable 'j' in A is taken to refer to the array B_{j+10}.

For sake of generality, two notational conventions will be required. Not every sequence A yields a well-defined sequence of operations; but if the operation D is given a sequence A which does not yield a well-defined sequence of operations, it will be understood to give output consisting of the 0-dimensional array "0". Also, if A contains m>k variable names, $D[A,B_1,...,B_k]$ may be defined by $D[A,B_1,...,B_k,0,...,0]$, where each 0 represents an array of appropriate size and dimension containing all zero entries.

Two simple examples are:

D[(1,9,9),(1,2,3)] = (1,2,3) + (1,2,3) = (2,4,6)

and

D[(1,9,2,10,9),(1,2,3),(4,2,1)] = (1,2,3) + (4,2,1)(1,2,3) = (5,6,6)

7.4.2. Array Component Systems

Now, using these nine operations, I will give an example of a new kind of self-generating system -- a type of system called an **array component system**, or **ACS.** Let us begin with a list of arrays V_i of the form

$V_i = (A_i, B_{i1},...,B_{im(i,t)})$,

t = 0,1,2,...; i = 1,2,...,N(t)

where the A_i are code sequences, and the B_{ij} are rational arrays. Each such array V_i may be associated with a hyperfunction

$H(V_i) = f_i$,

defined by the equation

$$f_i(f_j s) = H(D[A_i,B_{i1},...,B_{jm(j,t)},A_j s, B_{j1} s,...,B_{jm(j,t)} s])$$

The f_i are the **components**.

In other words, each component is associated with a code sequence, and a bunch of arrays. Applying one component to another means applying the code sequence of the first component to the list of arrays consisting of the first component's arrays, the second component's code sequence, and the second component's arrays. According to the conventions delineated above, it is possible for each component to act on other components which contain different numbers of arrays. "Missing" arrays are simply treated as zero arrays, and the control sequences A_i may potentially contain conditional expressions indicating how to deal with zero arrays.

ACS's illustrate in a very concrete way how hyperrelations may emerge as natural models of systems which are, in themselves, quite well-founded and computable. There is nothing in any way mysterious about the V_i, nor about the idea that the various V_i can act on one another. But in order to express this idea mathematically, one cannot use ordinary functions; one needs to use hyperrelations. To put it another way: in order to express **patterns** relating to the interaction of V_i, the only efficient course is to use hyperrelations.

7.4.3. Immune Systems as ACS's

Let us briefly consider a simple example, which will be taken up more thoroughly in Chapter Ten: immunodynamics.

The immune system is complicated as well as complex, and it contains many different kinds of cells. But the simplest mathematical models deal only with B-cells, and that is what I will do here. Let us begin with the approach of de Boer et al (1990), in which each **antibody type** in the immune system is associated with an integer sequence of length N. To be realistic, of course,antibodies should be modeled as three-dimensional rational arrays, since they are three-dimensional objects; but for the points I am making here, it is immaterial whether antibodies are associated with 1-D or 3-D arrays.

In the 1-D model, one may think of a B-cell as a pair $V_i = (A,B_i)$, where B_i is an integer sequence, and A is an integer **code sequence**. The code sequence specifies what happens when one forms $f_i(f_j)$, or in other words when one forms

$$A(B_i,A,B_j).$$

Specifically, what happens **most** of the time when $f_i(f_j)$ is formed is nothing. But if the conditions are right, the effects can be drastic. Define the **raw match** between two B-cells V_i and V_j as the maximum number of consecutive bits in which the corresponding sequences B_i and B_j are different. And define the **match** between two sequences as max{0, raw match(B_i,B_j) - T}, where T is some given threshold. In terms of component-systems, then, one may think of the dynamics of the immune system as specifying how a B-cell V_i acts on those B-cells V_j for which match(B_i,B_j) is large, thus causing the creation of **new** antibodies.

There is a great deal of biological subtlety involved here. In the crudest formal model, however, what happens is as follows: if $f_i(f_j)$ is formed and match(B_i,B_j) is large, then with a certain probability, the cell V_j is killed by the cell V_i (in fact, this killing takes place indirectly, via antibodies; but I do not need to consider these details here). But when the proportion of B-cells with shape B_j that are killed falls within a certain critical range, then cells of this shape are stimulated to **reproduce**. New B-cells are created.

Some of these new cells are identical to V_j, and some are new types V_k, which have have shape sequences similar, but not identical to B_j. This is **somatic mutation**. It is the creation of new B-cells by certain old B-cells acting on other old B-cells. There is randomness in the process because there is no deterministic way of telling exactly which new types of B-cells will be created.

This B-cells-only model is an extreme oversimplication. But the more accurate models are similar in spirit. Cells in the immune system act on one another, thus stimulating one another to produce new cells. Sometimes these new cells are copies of old cells, but sometimes they are structurally novel. Thisis a simple example of a component-system; and I have indicated in a rough way how it can be modeled using systems of stochastically computable hypersets.

Chapter Eight

THE COGNITIVE EQUATION

To anyone trained in physical science, the overall impression made by psychology and neuroscience is one of incredible **messiness**. So many different chemical compounds, so many different neural subsystems, so many different psychic dysfunctions, so many different components of intelligence, perception, control.... And no overarching conceptual framework in which all aspects come together to form a unified whole. No underlying equation except those of physics and chemistry, which refer to a level incomprehensibly lower than that of thoughts, emotions and beliefs. No **cognitive law of motion**.

Of course, there is no **a priori** reason to expect such a thing as a "cognitive law of motion" to be possible at all. It is amazing that one can find far-reaching yet precise generalizations such as Newton's laws in **any** field of study. To expect to find such conceptual jewels in every single discipline may be asking more than the world has to offer.

But on the other hand, consider: Newton's laws would have been impossible without calculus, general relativity would have been impossible without differential geometry, and quantum physics would have been impossible without functional analysis. It is quite conceivable that, once we have developed the appropriate mathematical concepts, the goal of a "cognitive law of motion" will cease to appear so unrealistic.

In fact, my contention is that **this time has already come**. As of 1993, I suggest, we have collectively developed precisely the mathematical and conceptual tools required to piece together the rudiments of a "fundamental equation of mind." The most important of these tools, I suggest, are four in number:

1) component systems

2) pattern theory

3) algorithmic information

4) strange attractors

In this chapter I show how these ideas may be used to formulate a new type of equation, which I call a "self-generating pattern dynamic." This is the type of equation, I suggest, which makes one thought, one emotion, drift into the next. It is the general form which a cognitive law of motion must take.

In *The Evolving Mind* the term "self-structuring system" is used to describe a system which, more than just organizing itself, **structures and patterns itself**; a system which studies the patterns in its past, thus determining the patterns in its future. Here I will delineate a class of systems which is a **subset** of the self-structuring systems -- namely, the class of systems that evolve by self-generating pattern dynamics. My hypothesis is that minds, as well as being self-structuring, also fall within this narrower category.

This is at the same time a brand new approach to mind, and a re-interpretation of the dual network model given in Chapter Three. The cognitive equation presents a **dynamical** view of mind, whereas the dual network presents a **static** view; but the two are ultimately getting at the same thing. In the dual network perspective, one begins with a structure, asks what the dynamics must be to retain that structure, and obtains the answer: something like the cognitive equation. In the cognitive equation perspective, on the other hand, one begins with a dynamical iteration, asks what sorts of structures will tend to persist under this iteration, and obtains the answer: something like the dual network. Dynamics lead to statics, statics leads to dynamics, and the simultaneous analysis of the two provides the beginning of an understanding of that mysterious process called **mind**.

8.1. MIND AS A SELF-GENERATING SYSTEM

The systems theory of Chapter Seven gives us a new way of looking at the dual network. The mind, filtered through the component-systems/self-generating-systems view, emerges as a **structured network of components**.

Note that this conclusion refers primarily to the **mind** -- the patterns in the brain -- and not to the brain itself. One **could** model the brain as a component-system, insofar as each neuron is not a fixed "component" but aspace of potential components -- one component for each condition of its synaptic potentials. When neuron A feeds its output to neuron B, thus altering its synaptic potential, it is in effect "creating" a new element of the space corresponding to neuron B. This may be a fruitful way to think about the brain. However, it is much more direct and elegant to view the collection **patterns** in the brain as a self-generating component-system -- recalling that a pattern is first of all a process. In the context of general systems theory, the pattern-theoretic model of mind is not merely useful but conceptually essential.

The mind is vastly different from a soup of molecules -- unlike the immune system, it is not even in a rough approximation well-mixed. (Putting brain tissue in a blender to make "synaptosome soup" is a nifty method for determining the levels of different neurotransmitters in the brain, but it has a definite negative effect on brain function.) But the relatively rigid structure of the brain does not prevent it from being a genuine self-generating system, and a genuine component-system.

There is an overall global structure of mind; and this structure self-organizes itself by a dynamic of **typeless interaction**, in which some mental processes act on others to produce yet others, without respect for any kind of "function/argument" distinction. One **can** model this sort of activity in terms of stochastic computation alone, without mentioning hypersets or component-systems -- this is the contemporary trend, which I have followed in my previous research. However, in many situations this point of view becomes awkward, and the only way to express the reality clearly is to adopt a typeless formalism such as the one developed in Sections 8.2 and 8.4.

Let us take a simple heuristic example -- purely for expository purposes, without any pretense of detailed biological realism. Let us think, in an abstract way, about the relation between a mental process that recognizes simple patterns (say lines), and a mental process that recognizes patterns **among** these simple patterns (say shapes). These shape recognizers may be understood as subservient to yet higher-level processes, say object recognizers. If the shape recognizer has some idea what sort of shape to expect, then it must partially **reprogram** the line recognizer, to tell it what sort of lines to look for. But if the line recognizer perpetually receives instructions to look for lines which are not there, then it must partially **reprogram** the shape recognizer, to cause it to give more appropriateinstructions. Assuming there is a certain amount of error innate in this process, one has an obvious circularity. The collection of two processors may be naturally modeled as a self-generating system.

It seems likely that the specific programs involved in these perceptual processes involve linear array operations. But still, one does not yet have an array component system. To see where component-systems come in, one needs to take a slightly more realistic view of the perceptual process. One must consider that the mapping between line-recognizing processes and shape-recognizing processes is many-to-many. Each shape process makes use of many line-recognizing process, and the typical line-recognizing process is connected to a few different shape-recognizing processes. A shape-recognizing process is involved in **creating** new line-recognizing processes; and a group of line-recognizing processes, by continually registering complaints, can cause the object-recognizing parents of the shape-recognizing processes to create new shape-recognizing processes.

What this means is that the **reprogramming** of processes by one another can be the causative agent behind the **creation** of new processes. So the collection of processes, as a whole, is not only a self-generating system but a component-system as well. By acting on one another, the mental processes cause new mental processes to be created. And, due to the stochastic influence of errors as well as to the inherent chaos of complex dynamics, this process of creation is unpredictable. Certain processes are more **likely** to arise than others, but almost anything is **possible**, within the parameters imposed by the remainder of the network of processes that is the mind.

This example, as already emphasized, is merely a theoretical toy. The actual processes underlying shape and line recognition are still a matter of debate. But the basic concept should be clear. Whenever one has sophisticated multilevel control, combined with heterarchical relationship, one has a situation in which self-referential models are appropriate. The whole network of processes **can** be modeled otherwise, using only stochastic computer programs. But the vocabulary of self-generating and component-systems leads to a novel understanding of the basic phenomena involved.

8.2. SELF-GENERATING PATTERN DYNAMICS (*)

Now let us return to the formal "process iterations" of Chapter Seven. Equation **(**)**, in itself, is much too general to be of any use as a "cognitive law of motion." If $System_t$ and T are chosen appropriately, then **(**)** can describe anything whatsoever. That is, after all, the meaning of universal computation! However, this simple iteration is nevertheless the first stop along the path to the desired equation. What is needed is merely to **specialize the operator T**.

Instead of taking the compounds formed from $System_t$, I suggest, one must take the **patterns in these compounds**. This completes the picture of the mind as a system which recognizes patterns in itself, which forms its own patterns **from** its own patterns. There might seem to be some kind of contradiction lurking here: after all, how can patterns in hyperrelations themselves be hyperrelations? But of course, this is precisely the distinctive quality of hyperrelations: they subvert the hierarchy of logical types by potentially belonging to their own domain and range. And this unusual property does **not** violate the laws of physical reality, because the hyperrelations required for practical modeling can themselves be perfectly well modeled in terms of ordinary Boolean functions.

To make this more precise, define the **relative structure** St^ of a set A = {a, b, c, ...} as the set of all x which are patterns in some subset of A **relative to** some other subset of A.

For instance, in the algorithmic information model, "x is an exact pattern in b relative to a" means

1) b produces x from a

2) $I(x|b,a) < I(x|a)$

More generally, statement (2) must be replaced with a less specific formal notion such as

2') $|x\backslash\{b,a\}| < |x\backslash a|$

The generalization of this notion to encompass patterns that are **approximate** rather than exact is quite straightforward.

In this notation, the simplest self-generating pattern dynamic says that, where System_t is the system at time t,

$$\textbf{System}_{t+1} = \textbf{St\^{}(R[System}_t\textbf{])} \quad \textbf{(***)}$$

I call this iteration the **basic deterministic dynamic**. It will serve as a "demonstration equation" for talking about the properties of more complicated cognitive dynamics.

The idea underlying this equation is encapsulated in the following simple maxim: **in a cognitive system, time is the process of structure becoming substance**. In other words, the entities which make up the system **now** all act on one another, and thus produce a new collection of entities which includes all the products of the interactions of entities currently existent. For lack of a better term, I call this exhaustive collection of products the "Raw Potentiality" of the system. Then, the system **one moment later** consists of the **patterns** in this collection, this Raw Potentiality.

8.2.1. A General Self-Generating Pattern Dynamic (*)

For every type of self-generating system, there is a corresponding type of self-generating pattern dynamic. The basic deterministic dynamic is founded on the type of self-generating system that is so totally "well-mixed" that **everything interacts with everything else at each time step**. But in general, this is only the simplest kind of self-generating system: a self-generating system may use **any stochastically computable rule** to transform what the Raw Potentiality of time t into the reality of time t+1.

Furthermore, the basic deterministic dynamic assumes infinite pattern recognition skill; it is anti-Godelian. In general, a self-generating system may use its Raw Potentiality in an incomplete fashion. It need not select **all possible** patterns in the Raw Potentiality; it may pick and choose which ones to retain, in a state-dependent way.

Formally, this means that one must consider iterations of the following form:

$$\textbf{System}_{t+1} = \textbf{F [Z}_t \textbf{ [St\^{}(G[R[System}_t\textbf{]])]]} \qquad \textbf{(****)}$$

where F and G are any stochastically computable functions, and $Z_t = Z[\text{System}_t]$ is a "filtering operator" which selects certain elements of

$\text{St\^{}(G[R[System}_t\text{]]])}$, based on the elements of System_t.

Note that the function F cannot make any reference to System_t; it must act on the level of structure alone. This is why the function Z_t is necessary. The particular system state System_t can affect the **selection** of which patterns to retain, but not the way these patterns are transformed. If this distinction were destroyed, if F and Z_t were allowed to blur together into a more general F_t = F[System$_t$], then the fundamental **structure-dependence** of the iteration would be significantly weakened. One could even define F_t as a constant function on all values of St^(G[R[System$_t$]]), mapping into a future state depending **only** on System$_t$. Thus, in essence, one would have **(**)** back again.

Equation **(****)**, like the basic deterministic dynamic **(***)**, is merely **(**)** with a special form of the transition operator T. T is now assumed to be a some sequence of operations, **one** of which is a possibly filtered application of the relative structure operator St^. This is indeed a bizarre type of dynamic -- instead of acting on real numbers or vectors, it acts on **collections of hyperrelations**. However, it may still be studied using the basic concepts of dynamical systems theory -- fixed points, limit cycles, attractors and so forth.

To see the profound utility of the filtering operator Z_t, note that it may be defined specifically to ensure that only those elements of St^(G[R[System$_t$]]) which are **actually computed by subsystems of System$_t$** are passed through to F and System$_{t+1}$. In other words, one may set

Z_t(X) = Z[System$_t$](X) = X **intersect** R[System$_t$]

Under this definition, **(****)** says loosely that System$_{t+1}$ consists of the patterns which System$_t$ has recognized in itself (and in the "compounds" formed by the interaction of its subsystems). It may be rewritten as

System$_{t+1}$ = F [R[System$_t$] intersect St^(G[R[System$_t$]])]   (*****)

This specialization brings abstract self-generating pattern dynamics down into the realm of physical reality. For reasons that will be clear a little later, it is this equation that I will refer to as the "cognitive equation" or "cognitive law of motion."

8.2.2. Summary

Self-generating pattern dynamics are dynamical iterations on collections of **processes**, and are thus rather different from the numerical iterations of classical dynamical systems theory and modern "chaos theory." However, it would be silly to think that one could understand mental systems by the **exact same methods** used to analyze physical systems.

The basic modeling ideas of graph-theoretic structure and iterative dynamics are applicable to both the mental and the physical worlds. But whereas in the physical domain one is concerned mainly with **numerical vectors**, in the mental realm one is concerned more centrally with **processes**. The two views are not logically contradictory: vectors may be modeled as processes, and processes may be modeled as vectors. However, there is a huge conceptual difference between the two approaches.

In non-technical language, what a "self-generating pattern dynamic" boils down to is the following sequence of steps:

1) Take a collection of processes, and let **each** process act on **all the other** processes, in whatever combinations it likes. Some of these "interactions" may result in nothing; others may result in the creation of new processes. The totality of processes created in this way is called the Raw Potentiality generated by the original collection of processes.

2) Transform these processes in some standard way. For instance, perhaps one wants to model a situation in which each element of the Raw Potentiality has only a certain **percentage chance** of being formed. Then the "transformation" of the Raw Potentiality takes the form of a selection process: a small part of the Raw Potentiality is selected to be retained, and the rest is discarded.

3) Next, determine all the **patterns** in the collection of processes generated by Step 2. Recall that patterns are themselves processes, so that what one has after this step is simply another collection of processes.

4) "Filter out" some of the processes in the collection produced by Step 3. This filtering may be **system-dependent** -- i.e., the original processes present in Step 1 may have a say in which Step 3-generated **pattern**-processes are retained here. For instance, as will be suggested below, it may often be desirable to retain only those patterns that are actually **recognized** by processes in Step 1.

5) Transform the collection of processes produced by Step 4 in some standard way, analogously to Step 2.

6) Take the set of processes produced by Step 5, and feed it back into Step 1, thus beginning the whole process all over again.

This is a very general sequence of steps, and its actual behavior will depend quite sensitively on the nature of the processes introduced in Step 1 on the firstgo-around, as well as on the nature of the transformation and filtering operations. Modern science and mathematics have rather little to say about this type of complex process dynamics. The general ideas of dynamical systems theory are applicable, but the more specific and powerful tools are not. If one wishes to understand the mind, however, **this** is the type of iteration which one must master.

More specifically, in order to model cognitive systems, a specific instance of the filtering operation is particularly useful: one filters out all but those patterns that are actually recognized by the components of the system. In other words, one takes the **intersection** of the products of the system and the patterns in the system. The self-generating pattern dynamic induced by this particular filtering operation is what I call the "cognitive equation."

Informally and in brief, one may describe the cognitive equation as follows:

1) Let all processes that are "connected" to one another act on one another.

2) Take all patterns that were recognized in other processes during Step (1), let these patterns be the new set of processes, and return to Step (1)

An **attractor** for this dynamic is then a set of processes with the property that each element of the set is a) produced by the set of processes, b) a pattern in the set of entities produced by the set of processes. In the following sections I will argue that complex mental systems are attractors for the cognitive equation.

8.3. STRUCTURAL CONSPIRACY

According to chaos theory, the way to study a dynamical iteration is to look for its **attractors**. What type of collection of processes would be an attractor for a self-generating pattern dynamic?

To begin with, let us restrict attention to the basic deterministic dynamic **(***)**. According to this iteration, come time t+1, the entities existent at time t are replaced by the patterns in the Raw Potentiality generated by these entities. But this does **not** imply that all the entities from time t completely vanish. That would be absurd -- the system would be a totally unpredictable chaos. It is quite possible for some of the current entities to survive into the next moment.

If a certain entity survives, this means that, as well as being an element of the current system System$_t$, it is also a **regularity** in the Raw Potentiality of System$_t$, i.e. an element of R[System$_t$]. While at first glance this might seem like a difficult sort of thing to contrive, slightly more careful consideration reveals that this is not the case at all.

As a simple example, consider two entities f and g, defined informally by

f(x) = the result of executing the command "Repeat x two times"

g(x) = the result of executing the command "Repeat x three times"

Then, when f acts on g, one obtains the "compound"

f(g) = the result of executing the command "Repeat x three times" the result of executing the command "Repeat x three times"

And when g acts on f, one obtains the "compound"

g(f) = the result of executing the command "Repeat x two times" the result of executing the command "Repeat x two times" the result of executing the command "Repeat x two times"

Now, obviously the pair (f,g) is a pattern in f(g), since it is easier to store f and g, and then apply f to g, than it is to store f(g). And, in the same way, the pair (g,f) is a pattern in g(f). So f and g, in a sense, perpetuate one another. According to the basic deterministic dynamic, if f and g are both present in System$_t$, then they will both be present in System$_{t+1}$.

One may rephrase this example a little more formally by defining f(x) = x x, g(x) = x x x. In set-theoretic terms, if one makes the default assumption that all variables are universally quantified, this means that f has the form {x,{x,x x}} while g has the form {x,{x,x x x}}. So, when f acts on g, we have the ugly-looking construction { {x,{x,x x x}}, {{x,{x,x x x}}, {x,{x,x x x}} {x,{x,x x x}} }; and when g acts on f, we have the equally unsightly {{x,{x,x x}}, {{x,{x,x x}}, {x,{x,x x}} {x,{x,x x}} {x,{x,x x}}}. It is easy to see that, given this formalization, the conclusions given in the text hold.

Note that this indefinite survival is fundamentally a synergetic effect between f and g. Suppose that, at time t, one had a system consisting of only two entities, f and h, where

h = "cosmogonicallousockhamsteakomodopefiendoplamicreticulu mpenproleta riatti"

Then the effect of h acting on f would, by default, be

h(f) = empty set

And the effect of f acting on h would be

f(h) = "cosmogonicallousockhamsteakomodopefiendoplasmicreticulum

penproletariatticosmogonicallousockhamsteakomodope fiendoplasmicreticulumpenproletariatti"

Now, (f,h) is certainly a pattern in f(h), so that, according to the basic deterministic dynamic, **f** will be a member of System$_{t+1}$. But **h** will not be a member of System$_{t+1}$ -- it is not a pattern in anything in R[System$_t$]. So there is no guarantee that f will be continued to System$_{t+2}$.

What is special about f and g is that they assist one another in producing entities in which they are patterns. But, clearly, the set {f,g} is not **unique** in possessing this property. In general, one may define a **structural conspiracy** as any collection of entities G so that every element of G is a pattern in the Raw Potentiality of G. It is obvious from the basic deterministic dynamic that **one successful strategy for survival over time is to be part of a structural conspiracy**.

Extending this idea to general **deterministic** equations of the form **(****)**, a **structural conspiracy** may be redefined as any collection P which is preserved by the dynamic involved, i.e. by the mathematical operations R, G, St^ and F applied in sequence.

And finally, extending the concept to **stochastic** equations of form **(****)**, a structural conspiracy may be defined as a collection P which has a **nonzero probability** of being preserved by the dynamic. The **value** of this probability might be called the "solidity" of the conspiracy. Stochastic dynamics are interesting in that they have the potential to break down even solid structural conspiracies.

One phrase which I use in my own thinking about self-generating pattern dynamics is "passing through." For an entity, a pattern, to survive the iteration of the fundamental equation, it must remain intact **as a pattern** after the process of universal interdefinition, universal interaction has taken place. The formation of the Raw Potentiality is a sort of holistic melding of all entities with all other entities. But all that survives from this cosmic muddle, at each instant, is the **relative**

structure. If an entity survives this process of melding and separation, then it has passed through the whole and come out intact. Its integral relationship with the rest of the system is confirmed.

8.3.1. Conspiracy and Dynamics

What I have called a structural conspiracy is, in essence, a **fixed point**. It is therefore the simplest kind of attractor which a self-generating pattern dynamic can have. One may also conceive of self-generating-pattern-dynamic **limit cycles** -- collections P so that the presence of P in System_t implies the presence of P in System_{t+k}, for some specific integer k>1.

Nietzsche's fanciful theory of the "eternal recurrence" may be interpreted as the postulation of a universe-wide limit-cycle. His idea was that the system, with all its variation over time, is inevitably repetitive, so that every moment which one experiences is guaranteed to occur again at some point in the future.

And, pursuing the same line of thought a little farther, one may also consider the concept of a **self-generating-pattern-dynamical strange attractor**. In this context, one may define a "strange attractor" as a group P of entities which are "collectively fixed" under a certain dynamic iteration, even though the iteration does not cycle through the elements of P in any periodic way. Strange attractors may be **approximated** by limit cycles with very long and complicated periodic paths.

In ordinary dynamical systems theory, strange attractors often possess the property of **unpredictability**. That is, neither in theory nor in practice is there any way to tell **which** attractor elements will pop up at which future times. Unpredictable strange attractors are called **chaotic attractors**. But on the other hand, some strange attractors are statistically predictable, as in Freeman's "strange attractor with wings" model of the sense of smell. Here chaos coexists with a modicum of overlying order.

It is to be expected that self-generating pattern dynamical systems possess chaotic attractors, as well as more orderly strange attractors. Furthermore, in ordinary dynamics, strange attractors often contain fixed points; and so, in self-generating pattern dynamics, it seems likely that strange structural conspiracies will contain ordinary structural conspiracies (although these ordinary structural conspiracies may well be so **unstable** as to be irrelevant in practice). However, there is at the present time **no mathematical theory** of direct use in exploring the properties of self-generating pattern dynamical systems or any other kind of nontrivial self-generating system. The tools for exploring these models simply do not exist; we must make them up as we go along.

Fixed points are simple enough that one can locate them by simple calculation, or trained intuition. But in classical dynamical systems theory, most strange attractors have been found **numerically**, by computer simulation or data analysis. Only rarely has it been possible to verify the presence of a strange attractor by formal mathematical means; and even in these cases, the existence of the attractor was determined by computational means **first**. So it is to be expected that the procedure for self-generating dynamics will be the same. By running simulations of various self-generating systems, such as self-generating pattern dynamics, we will **happen upon** significant strange attractors ... and follow them where they may lead.

8.3.2. Immunological Pattern Dynamics

The immune system, as argued at the end of Chapter Seven, is a self-generating component-system. The cognitive equation leads us to the very intuitive notion that, even so, it is not **quite** a cognitive system.

Insofar as the immune system is a self-maintaining **network**, the survival of an antibody type is keyed to the ability of the type to recognize some other antibody type. If A recognizes B, then this is to be viewed as B **creating** instances of A (indirectly, via the whole molecular system of communication and reproduction). So the antibody types that survive are those which are **produced** by other antibody types: the immune network is a self-generating component-system.

The next crucial observation is that the **recognition** involved here is a pattern-based operation. From the fact that one specific antibody type recognizes another, then it follows only that there is a significant amountof pattern **emergent** between the two antibody types; it does not follow that the one antibody type is a pattern in the other. But the ensuing **reproduction** allows us to draw a somewhat stronger conclusion. Consider: if type A attacks type B, thus stimulating the production of more type A -- then what has happened? The **original** amounts of A and B, taken together, have served as a **process** for generating a greater amount of A. Is this process a **pattern** in the new A population? Only if one accepts that the type B destroyed was of "less complexity" than the type A generated. For instance, if **two** A's were generated for each **one** B destroyed, then this would seem clear. Thus, the conclusion: in at least some instances, antibody types can be patterns in other antibody types. But this cannot be considered the rule. Therefore, the immune system is not quite a fully cognitive system; it is a borderline case.

Or, to put it another way: the cognitive equation is an idealization, which may not be completely accurate for any biologically-based system. But it models some systems **better** than others. It models the immune system far better than the human heart or a piece of tree bark -- because the immune system has many "thought-like" properties. But, or so I will argue, it models the brain even **more** adeptly.

8.4. MIND AS A STRUCTURAL CONSPIRACY

I have said that mind is a self-generating system, and I have introduced a particular form of self-generating system called a "self-generating pattern dynamic." Obviously these two ideas are not unrelated. In this section I will make their connection explicit, by arguing that **mind is a structural conspiracy** -- an attractor for a self-generating pattern dynamic.

More specifically, I will argue that a **dual network** is a kind of structural conspiracy. The key to relating self-generating pattern dynamics with the dual network is the filtering operator Z_t.

8.4.1. The Dual Network as a Structural Conspiracy

It is not hard to see that, with this filtering operation, an associative memory is **almost** a structural conspiracy. For nearly everything in an associative memory is a pattern emergent among other things in that associative memory. As in the case of multilevel control, there may be a few odd men out -- "basic facts"being stored which are not patterns in anything. What is required in order to make the whole memory network a structural conspiracy is that these "basic facts" be **generatable as a result of some element in memory acting on some other element**. These elements must exist by virtue of being patterns in **other** things -- but, as a side-effect, they must be able to generate "basic facts" as well.

Next, is the perceptual-motor hierarchy a structural conspiracy? Again, not necessarily. A process on level L may be generally expected to be a pattern in the products obtained by letting processes on level L-1 act on processes from level L-2. After all, this is their purpose: to recognize patterns in these products, and to **create** a pattern of success among these products. But what about the bottom levels, which deal with immediate sense-data? If these are present in System$_t$, what is to guarantee they will continue into System$_{t+1}$. And if these do not continue, then under the force of self-generating pattern dynamics, the whole network will come crashing down....

The only solution is that the lower level processes must not **only** be patterns in sense data, they must **also** be patterns in products formed by higher-level processes. In other words, we can only **see** what we can **make**. This is not a novel idea; it is merely a reformulation of the central insight of the Gestalt psychologists.

Technically, one way to achieve this would be for there to exist processes (say on level 3) which **invert** the actions taken by their subordinates (say on level 2), thus giving back the contents of level 1. This inversion, though, has to be part of a process which is **itself a pattern in level 2** (relative to some other mental process). None of this is inconceivable, but none of it is obvious either. It is, ultimately, a **testable prediction** regarding the nature of the mind, produced by equation **(*****)**.

The bottom line is, it is quite possible to conceive of dual networks which are **not** structural conspiracies. But on the other hand, it is not much more difficult, on a purely abstract level, to envision dual networks which **are**. Equation **(*****)** goes beyond the dual network theory of mind, but in an harmonious way. The prediction to which it leads is sufficiently dramatic to deserve a name: the "producibility hypothesis." **To within a high degree of approximation, every mental process X which is not a pattern in some other mental process, can be produced by applying some mental process Y to some mentalprocess Z, where Y and Z are patterns in some other mental process**.

This is a remarkable kind of "closure," a very strong sense in which the mind is a world all its own. It is actually very similar to what Varela (1978) called "autopoesis" -- the only substantive difference is that Varela believes autopoetic systems to be inherently non-computational in nature. So far, psychology has had very little to say about this sort of self-organization and self-production. However, the advent of modern complex systems science promises to change this situation.

8.4.2. Physical Attractors and Process Attractors

All this is quite unorthodox and ambitious. Let me therefore pause to put it into a more physicalistic perspective. The brain, like other extremely complex systems, is unpredictable on the level of detail but roughly predictable on the level of structure. This means that the dynamics of its physical variables display a strange attractor with a complex structure of "wings" or "compartments." Each compartment represents a certain collection of states which give rise to the same, or similar, patterns. Structural predictability means that each compartment has wider doorways to some compartments than to others.

The complex compartment-structure of the strange attractor of the physical dynamics of the brain determines the macroscopic dynamics of the brain. There would seem to be no way of determining this compartment-structure based on numerical dynamical systems theory. Therefore one must "leap up a level" and look at the dynamics of **mental processes**, perhaps represented by interacting, inter-creating **neural maps**. The dynamics of these processes, it is suggested, possess their **own** strange attractors called "structural conspiracies," representing collections of processes which are closed under the operations of patter-recognition and interaction. Process-level dynamics results in a compartmentalized attractor of states of the network of mental processes.

Each state of the network of mental processes represents a large number of possible underlying physical states. Therefore process-level attractors take the form of **coarser structures**, superimposed on physical-level attractors. If physical-level attractors are drawn in ball-point pen, process-level attractors are drawn in magic marker. On the physical level, a structural conspiracy represents a whole complex of compartments. But only the most densely connected regions of the compartment-network of the physical-level attractor can correspond to structural conspiracies.

Admittedly, this perspective on the mind is somewhat speculative, in the sense that it is not closely tied to the current body of empirical data. However, it is in all branches of science essential to look **ahead** of the data, in order to understand what sort of data is really worth collecting. The ideas given here suggest that, if we wish to understand mind and brain, the most important task ahead is to collect information regarding the compartment-structure of the strange attractor of the brain, both on the physical level and the process level; and

above all to understand the complex relation between the strange attractors on these two different levels.

8.5. LANGUAGE ACQUISITION

I have proposed that the **mind** is an attractor for the cognitive equation. But this does not rule out the possibility that some particular **subsets** of the mind may **also** be attractors for the cognitive equation, in themselves. In particular, I suggest that **linguistic systems** tend to be structural conspiracies.

This idea sheds new light on the very difficult psychological problem of **language acquisition**. For in the context of the cognitive equation, language acquisition may be understood as a process of **iterative convergence toward an attractor**. This perspective does not solve all the micro-level puzzles of language acquisition theory -- no general, abstract theory can do that. But it does give a new overarching framework for approaching the question of "how language could possibly be learned."

8.5.1. The Bootstrapping Problem

The crucial puzzle of language acquisition theory is the "bootstrapping problem." What this catch phrase means is: if all parts of language are defined in terms of **other** parts of language, then where is the mind to start the learning process?

Consider the tremendous gap between the **input** and the **output** of the language learning process. What a child is presented with are sentences heard in context. Gradually, the child's mind learns to detect components and properties of these sentences: such things asindividual words, word order, individual word meanings, intonation, stress, syllabic structure of words, general meanings of sentences, pragmatic cues to interpretation, etc. All this is just a matter of correlating things that occur together, and dividing things into natural groupings: difficult but straightforward pattern recognition.

But what the child's mind eventually arrives at is so much more than this. It arrives at an implicit understanding of grammatical **categories** and the rules for their syntactic interrelation. So the problem is, how can a child determine the relative order of noun and verb without first knowing what "nouns" and "verbs" are? But on the other hand, how can she learn to distinguish nouns and verbs except by using cues from word order? Nouns do not have a unique position, a unique intonation contour, a unique modifier or affix -- there is no way to distinguish them from verbs based on non-syntactic pattern recognition.

The formal model of language given in Chapter Five makes the bootstrapping problem appear even more severe. First of all, in the definition of "syntactic system," each **word** is defined as a fuzzy set of functions acting on other words. How then are words to be learned, if each word involves functions acting on other words? With what word could learning possibly start? Yes, some very simple words can be partially represented as functions with null argument; but most words need other words as arguments if they are to make any sense at all.

And, on a higher level of complexity, I have argued that **syntax** makes no sense without **semantics** to guide it. No mind can **use** syntax to communicate unless it has a good understanding of semantics; otherwise, among other problems, the paradoxes of Boolean logic will emerge to louse things up. But on the other hand, **semantics**, in the pattern-theoretic view, involves determining the set of all patterns associated with a given word or sentence. And the bulk of these patterns involve words and more complex syntactic structures like phrases and clauses: this is the systematicity of language.

No syntax without semantics, no semantics without syntax. One cannot recognize correlations among syntactic patterns until one knows syntax to a fair degree. But until one has recognized these correlations, one does not know semantics, and one consequently cannot use syntax for any purpose. But how can one learn syntax at all, if one cannot use it for any purpose?

Chomsky-inspired **parameter-setting** theories circumvent this chicken-and-egg problem in a way which iseither clever, obvious or absurd, depending on your point of view. They assume that the brain has a genetically-programmed "language center," which contains an abstract version of grammar called **Universal Grammar** or **UG**.

UG is understood to contain certain "switches" -- as a switch which determines whether nouns come before or after verbs, a switch which determines whether plurals are formed by affixes or by suffixes, and so on. The class of possible human syntaxes is the class of possible switch settings for UG; and language learning is a process of determining how to set the switches for the particular linguistic environment into which one has been born.

The parameter-setting approach simplifies the bootstrapping problem by maintaining that syntaxes are **not** actually learned; they are merely selected from a pre-arranged array of possibilities. It leaves only the much more manageable problem of **semantic bootstrapping** -- of explaining how semantic knowledge is acquired by induction, and then **combined** with UG to derive an appropriate syntax. Some theorists, however, consider the whole parameter-setting approach to be a monumental cop-out. They stubbornly maintain that all linguistic knowledge must be **induced** from experience. In other words, to use my earlier example, first the child gets a vague idea of the concept of "noun" and "verb"; then, based on this vague idea, she arrives at a vague idea of the relative positioning of nouns and verb. This inkling about positioning leads to a slightly strengthened idea of "noun" and "verb" -- and so forth.

In general, according to this view, the child begins with very **simple** grammatical rules, specific "substitution frames" with slots that are labeled with abstract object types; say "NOUN VERB" or "NOUN go to NOUN" or "NOUN is very ADJECTIVE". Then, once these simple frames are mastered, the child induces patterns among these substitution frames. "NOUN eats NOUN," "NOUN kills NOUN," "NOUN tickles NOUN," etc., are generalized into NOUN VERB NOUN. Next, more complex sentence structures are built up from simple substitution frames, by **induced** transformational rules.

In the inductivist perspective, bootstrapping is understood as a difficult but not insurmountable problem. It is assumed that the 10*10* - 10*12* neurons of the human brain are up to the task. Parameter-setting theorists have a more pessimistic opinion of human intelligence. But the trouble with the whole debate is that **neither**side has a good overall concept of what kind of learning is taken place.

In other words: if it's inductive learning, what kind of structure does the induction process have? Or if it's parameter setting, what is the logic of the process by which these "parameters" are learned -- how can this mechanistic model be squared with the messiness of human biology and psychology? In short, **what is the structure of linguistic intelligence?** My goal in this section is to suggest that the **cognitive equation** may provide some hints toward the resolution of this conceptual difficulty.

8.5.2. Process-Network Theories of Language Learning

The dual network model suggests that language learning must be explicable on the level of **self-organizing, self-generating process dynamics**. This is something of a radical idea, but on the other hand, it can also be related with some of the "mainstream" research in language acquisition theory. And, I will argue, it provides an elegant way of getting around the bootstrapping problem.

8.5.2.1. Constraint Satisfaction Models

Perhaps the most impressive among all parameter-setting theories is Pinker's (1987) constraint satisfaction model. Initially Pinker wanted to model language learning using a **connectionist** architecture a la Rumelhart and McClelland (1986). But this proved impossible; and indeed, all subsequent attempts to apply simple "neural networks" to symbolic learning problems have been equally fruitless.

So instead, Pinker borrowed from artificial intelligence the idea of a self-adjusting **constraint satisfaction** network. The idea is that language acquisition results from the joint action of a group of constraint satisfaction networks: one for assigning words to categories, one for determining grammatical structures, one for understanding and forming intonations, etc.

Consider, for instance, the network concerned with grammatical structures. Each node of this network consists of a **rule prototype**, a potential grammatical rule, which has **its own opinion** regarding the role of each word in the sentence. The dynamics of the network is **competitive**. If the sentence is "The dog bit the man," then one rule might categorize "The dog" as subjectand "bit the man" as verb phrase; another might categorize "The dog bit" as subject and "the man" as verb phrase. But if a certain rule prototype **disagrees** with the majority of its competitors regarding the categorization of a word, then its "weight" is decreased, and its opinion is counted less in the future.

The behavior of the network gets interesting when rules agree regarding some categorizations and disagree regarding others. The weights of rules may fluctuate up and down wildly before settling on an "equilibrium" level. But eventually, **if** the rule network is sufficiently coherent, an "attractor" state will be reached.

If there were **no** initial knowledge, then this competitive process would be worthless. No stable equilibrium would ever arise. But Pinker's idea is that the abstract rules supplied by UG, combined with rudimentary rules learned by induction, are enough to ensure the convergence of the network. This is a fancy and exciting version of the "parameter-setting" idea: parameters are not being directly set, but rather UG abstractions are being used to **guide** the convergence of a self-organizing process.

8.5.2.2. Competition Models

An interesting counterpoint to Pinker's network model is provided by the **evolutionary** approach of Bates and MacWhinney (1987). They present cross-linguistic data suggesting that language learning is **not** a simple process of parameter-setting. Children learning different languages will often differ in their **early** assumptions about grammar, as well as their ultimate syntactic rule structures. Furthermore, the passage from early grammar to mature grammar may be an **oscillatory** one, involving the apparent competition of conflicting tendencies. And different children may, depending on their particular abilities, learn different aspects of the same language at **different times**: one child may produce long sentences full of grammatical errors at an early stage, while another child may first produce flawless short sentences, only then moving on to long ones.

These observations disprove only the crudest of parameter-setting theories; they do not contradict complex parameter-setting theories such as Pinker's constraint satisfaction network, which integrates UG with inductive rule learning in a self-organizational setting. But they do suggest that even this kind of sophisticatedparameter-setting is not quite sophisticated enough. The single-level iteration of a constraint satisfaction network is a far cry from the flexible multilevel iterations of the brain.

What Bates and MacWhinney propose is a sort of "two-level network" -- one level for **forms** and another for **functions**. Form nodes may be connected to function nodes; for example, the form of preverbal positioning in English is correlated with the function of expressing the actor role. But there may also be **intra-level** connections: form nodes may be connected to other form nodes, and function nodes to other function nodes.

In their view, mappings of a single form onto a single function are quite rare; much more common is widely branching interconnection. For instance, they argue that

"subject" is neither a single symbol nor a unitary category. Rather, it is a coalition of many-to-many mappings between the level of form (e.g. nominative case marking, preverbal position, agreement with the verb in person and number) and the level of function (e.g. agent of a transitive action, topic of an ongoing discourse, perspective of the speaker)....

Notice that the entries at the level of form include both "obligatory" or "defining" devices such as subject-verb agreement, and "optional" correlates like the tendency for subjects to be marked with definite articles. This is precisely what we mean when we argue that there is no sharp line between obligatory rules and probabilistic tendencies.

Learning is then a process of modifying the weights of connections. Connections that lead to unsatisfactory results have their weights decreased, and when there is a conflict between two different nodes, the one whose connection is weighted highest will tend to prevail.

8.5.2.3. Summary

Bates and MacWhinney, like Pinker, view language learning as largely a process of **adjusting the connections between various "processes" or "nodes."** While this is not currently **known** to be the correct approach to language acquisition, I submit that it is by far the most plausible framework yet proposed. For Neural Darwinism teaches us that the brain is a networkof interconnected processes, and that learning consists largely of the adjustment of the connections between these processes. The process-network view of language acquisition fits quite neatly into what we know about the brain and mind.

And the question "UG or not UG," when seen in this light, becomes rather less essential. What is most important is the **process dynamics** of language learning. Only once this dynamics is understood can we understand just how much initial information is required to yield the construction of effective linguistic neural maps. Perhaps the inductivists are right, and abstract cognitive abilities are sufficient; or perhaps Chomsky was correct about the necessity of pre-arranged grammatical forms. But one's opinion on this issue cannot serve as the **basis** for a theory of language acquisition. The process-network view relegates the innate-vs.-acquired debate to the status of a side issue.

8.5.3. The Cognitive Equation and Language Learning

So, language learning is largely a process of adjusting the weights between different processes. But how are these processes **arrived at** in the first place? Some of them, perhaps, are supplied genetically. But many, probably most, are learned inductively, by **pattern recognition**. This gives rise to the question of whether a **language** is perhaps a **structural conspiracy**.

The above discussion of "bootstrapping" suggests that this may indeed be the case. Parts of speech like "nouns" and "verbs" are **patterns** among sentences; but they are only producible by processes involving **word order**. On the other hand, rules of word ordering are **patterns** among sentences, but they are only producible by processes involving **parts of speech**.

Bootstrapping states precisely that, once one knows **most** of the rules of syntax, it's not hard to induce the rest. Suppose one assumes that the processes bearing the rules of language all

1) possess modest pattern-recognition capacities, and

2) are programmed to recognize patterns in sentences

Given this, it follows from the bootstrapping problem that any portion of a mind's linguistic system is capable of **producing** the rest, according to the dynamics of the cognitive equation. In other words, it follows that language is an **attractor**, a structural conspiracy.

And if one accepts this conclusion, then the next natural step is to view language learning as a process of **convergence to this attractor**. This is merely a new way of conceptualizing the point of view implicit in the work of Pinker, Bates, MacWhinney, and other process-network-oriented acquisition theorists. These theorists have focused on the dynamics of **already-existing** networks of linguistic rules; but as Pinker explicitly states, this focus is for sake of simplicity only (after all, rule-bearing processes must come from **somewhere**). The cognitive equation shifts the focus from connection adjustment to process creation, but it does not alter the underlying process-network philosophy.

The learning process starts with an initial collection of syntactic rules -- either simple substitution rules picked up from experience, or randomly chosen specific cases of abstract UG rules, or a combination of the two. Then each rule-bearing process **recognizes patterns** -- among incoming and outgoing sentences and its companion processes.

This recognition process results in the production and comprehension of sentences, via its interaction with **outside** perceptual and motor processes, and the associative memory network (recall the intimate connection between syntax and semantics, discussed in Chapter Five). But internally, it **also** leads to the creation of new processes ... which aid in the production and comprehension of sentences, and in the creation of processes.

And this process is repeated until eventually nothing new is generated any more -- then an attractor has been reached. Language, a self-sustaining mental system, has been learned.

Chapter Nine

BELIEF SYSTEMS

I believe, so that I may understand

-- Saint Augustine

Believing is the primal beginning

even in every sense impression....

-- Friedrich Nietzsche

Are belief systems attractors? There is something quite intuitive about the idea . Before one settles on a fixed system of beliefs, one's opinions regarding a certain issue may wander all over the spectrum, following no apparent pattern. But once one arrives at a belief system regarding that topic, one's opinions thereon are unlikely to vary from a narrow range.

But of course, if one is to declare that belief systems are attractors, one must specify: attractors of **what** dynamical system? To say "attractors of brain dynamics" is obvious but inadequate: the brain presents us with a system of billions or trillions of coupled nonlinear equations, which current methods are incapable of analyzing even on a qualitative level. If belief systems are to be usefully viewed as attractors, the relevant dynamical iteration must exist on a higher level than that of individual neurons.

In the preceding chapters I have argued that, in order to make headway toward a real understanding of themind, one must shift up from the neural level and consider the structure and dynamics of interacting **mental processes** or **neural maps** (Edelman, 1988). Specifically, I have proposed an equation for the evolution of mental processes, and I have suggested that psychological systems may be viewed as subsets of the dual network which are **strange attractors** of this equation. Now, in this chapter, I will begin the difficult task of relating these formal ideas to real-world psychology -- to discuss the sense in which particular human belief systems may be seen as subsystems of the dual network, and attractors of the cognitive equation.

After giving a simple formalization of the concept of "belief," I will consider the dynamics of belief systems as displayed in the history of science, with an emphasis on Lakatos's structural analysis of research programmes. Then I will turn to a completely different type of belief system: the conspiracy theory of a paranoid personality. By constrasting these different sorts of belief systems in the context of the dual network and the cognitive equation, a new understanding of the nature of **rationality** will be proposed. It will be concluded that irrationality is a kind of abstract **dissociation** -- a welcome conclusion in the light of recent work relating dissociation with various types of mental illness (van der Kolk et al, 1991).

Personalities and their associated belief systems are notoriously vague and complicated. It might seem futile to attempt to describe such phenomena with precise equations. But the Church-Turing Thesis implies that one can model anything in terms of computational formulas -- if one only chooses the right sort of formulas. My claim is that the "cognitive law of motion," applied in the context of the dual network model, is adequate for describing the dynamics of mentality. The theory of belief systems given in this chapter and the next is a partial substantiation of this hypothesis.

9.1 SYSTEMATIC BELIEF

In this section I will give abstract, formal definitions for the concepts of "belief" and "belief system." Though perhaps somewhat tedious, these definitions serve to tie in the idea of "belief" with the formal vocabulary introduced in Chapters Two and Three;and they provide a solid conceptual foundation for the more practical considerations of the following sections.

The basic idea is that a **belief** is a mental process which, in some regard, gives some other mental process the "benefit of the doubt." Recall that, in Chapter Two, I defined an **infon** as a fuzzy set of patterns. Suppose that a certain process X will place the process s in the associative memory **just as if** s displayed infon i -- without even checking to see whether s really does display i. Then I will say that X **embodies the belief** that s displays infon i. X gives s the benefit of the doubt regarding i.

The mental utility of this sort of benefit-giving is obvious: the less processing spent on s, the more available for other tasks. Mental resources are limited and must be efficiently budgeted. But it is equally clear that minds must be very careful where to suspend their doubts.

Next, a **test** of a belief may be defined as a process with the potential to create an infon which, if it were verified to be present, would **decrease** the intensity of the belief. In other words, a test of a belief X regarding s has the potential to create an infon j which caused X to give s **less** benefit of the doubt. Some beliefs are more testable than others; and some very valuable beliefs are surprisingly difficult to test.

Finally, a **belief system** is a group of beliefs which mutually support one another, in the sense that an increased degree of belief in one of the member beliefs will generally lead to increased degrees of belief in most of the other member beliefs. The systematicity of belief makes testing particularly difficult, because in judging the effect of infon j on belief X, one must consider the **indirect** effects of j on X, via the effects of j on the other elements of the belief system. But, unfortunately for hard-line rationalists, systematicity appears to be **necessary** for intelligence. It's a messy world out there!

9.1.1. Formal Definition of Belief (*)

A **belief**, as I understand it, is a proposition of the form

" s |-- i with degree d"

or, in more felicitous notation,

(s,i;d).

In words, it is a proposition of the form "the collection of patterns labeled i is present in the entity s with intensity d." To say that the individual x holds the belief (s,i;d), I will write

"s |-- i //x with degree d",

or, more compactly,

(s,i,x;d).

Mentally, such a proposition will be represented as a collection of processes which, when presented with the entity s, will **place s in the associative memory** exactly as they would place an entity which they had verified to contain patterns i with intensity d. A belief about s is a process which is willing to give s the benefit of the doubt in certain regards. This definition is simple and natural. It does not hint at the full psychological significance of belief; but for the moment, it will serve us well.

Next, what does it mean to **test** a belief? I will say that an infon j is a **test** of a belief (s,i,x) relative to the observer y, with certainty level e, to degree NM, where

N = the degree to which the observer y believes that the determination of the degree in d(s,j,x) will cause a **decrease** in d(s,i,x).

M = the amount of effort which the observer y believes will be required to determine the degree that s |-- j holds to within certainty e

I believe that this formal definition, awkward as it is, captures what one means when one makes a statement like "That would be a test of Jane's belief in so and so." It is not an objective definition, and it is not particularly profound, but neither is it vacuous: it serves its purpose well.

Factor N alone says that j is a test of i if **y** believes that determining whether j holds will affect x's degree of belief that i holds. This is the **essence** of test. But it is not adequate in itself, because j is not a useful test of i unless it is actually **possible** to determine the degree to which j holds. This is the purpose of the factor M: it measures the practicality of executing the test j.

To see the need for M, consider the theory, well known among philosophers, that there is some spot on the Earth's surface which has the property that anyone who stands there will see the devil. The only test of this is to stand on every single spot on the earth's surface, which is either impossible or impractically difficult, depending on the nature of space and time.

Or consider Galileo's belief that what one sees by pointing a telescope toward space is actually "out there". Since at that time there was no other source of detailed information as to what was "out there," there was no way to test this belief. Now we have sent men and probes into space, and we have measured the properties of heavenly bodies with radio telescopy and other methods; all these tests have supported Galileo's belief. But it is not hard to see why most of Galileo's contemporaries thought his belief unreasonable.

The role of the "observer" y is simple enough. If one posits an outside, "impartial" observer with access to all possible futures, then one can have an objective definition of test, which measures the degree to which the presence of a certain infon **really will** alter the strength of a belief. On the other hand, one may also consider the most "partial" observer of all: the belief-holder. It is interesting to observe that, when a certain **human** belief system appears to be strongly resistant to test, the belief-holders will generally acknowledge this fact just as readily as outside observers.

9.1.2. Systematic Belief (*)

The formal definition of "belief system" is a little bit technical, but the basic idea is very simple: a belief system is a collection of beliefs which are **mutually supporting** in that a test for any one of them is a test for many of the others. It is permitted that evidence in favor of some of the beliefs may be evidence **against** some of the others -- that what increases the intensity of belief in A may decrease the intensity of belief in B, where both A and B are in the system. But this must not be the rule -- the positive reinforcement must, on balance, outweigh the negative reinforcement.

To be precise, consider a set of beliefs $\{A_1,...,A_n\}$. Let $c_{ij} = c_{ij}(K;y)$ denote the amount of increase in the degree to which A_j holds that, in the belief of y, will result from an **increase** by an amount ofK in the degree to which A_i holds. Decrease is to be interpreted as negative increase, so that if y believes that a **decrease** in the degree to which A_j holds will result from an increase in the degree to which A_i holds by amount, then $c_{ij}(K;y)$ will be negative. As with tests, unless otherwise specified it should be assumed that y=x.

Then the **coherence** $C(\{A_1,...,A_n\})$ of the set $\{A_1,...,A_n\}$ may be defined as the sum over all i, j and K of the c_{ij}. And the **compatibility** of a belief B with a set of beliefs $\{A_1,...,A_n\}$ may be defined as $C(\{A_1,...,A_n,B\})$ -

$C(\{A_1,...,A_n\})$.

The coherence of a set of beliefs is the degree to which the various member beliefs support each other, on the average, in the course of the mental process of the entity containing the beliefs. It is not the degree to which the various member beliefs "logically support" each other -- it depends on no system of evaluation besides that of the holder of the beliefs. If I think two beliefs contradict each other, but in your mind they strongly reinforce eachother, then according to the above definition the two beliefs may still be a strongly coherent belief system relative to your mind. It follows that the "same" set of beliefs may form a different dynamical system in two different minds.

Additionally, it is not necessary that two beliefs in the same mind always stand in the same relation to each other there. If A_1 contradicts A_2 half the time, but supports A_2 half the time with about equal intensity, then the result will be a c_{12} near zero.

If none of the c_{ij} are negative, then the belief system is "consistent": none of the beliefs work against eachother. Obviously, consistency implies coherence, though not a high degree of coherence; but coherence does not imply consistency. If some of its component beliefs contradict eachother, but others support eachother, then the coherence of a set of beliefs can still be high -- as long as the total amount of support exceeds the total amount of contradiction.

If a set of beliefs has negative coherence it might be said to be "incoherent." Clearly, an incoherent set of beliefs does not deserve the title "belief system." Let us define a belief system as a set of beliefs which has positive coherence.

The compatibility of a belief B with a belief system measures the expected amount by which the addition of Bto the belief system would change the coherence of the belief system. If this change would be positive, then B has positive compatibility; and if this change would be negative, then B has negative compatibility -- it might be said to be incompatible.

Finally, it must be noted that a given **human** mind may contain two mutually **incompatible** belief systems. This possibility reflects the fundamentally "dissociated" (McKellar, 1979) nature of human mentality, whereby the mind can "split" into partially autonomous mental sub-networks. The computation of the coefficients c_{ij} may be done with respect to any system one desires -- be it a person's mind, a society, or one **component** of a person's mind.

9.1.3. Belief and Logic

How does a mind determine how much one belief supports another? In formal terms, how does it determine the "correlation" function c_{ij} between belief i and belief j? Should an analysis of belief merely accept these "intercorrelations" a priori, as given products of the believing mind in question? Or is there some preferred "rational" method of computing the effect of a change in the intensity of one belief on the intensity of another?

To see how very difficult these question are, assume for the sake of argument that all beliefs are propositions in Boolean logic. Consider a significantly cross-referential belief system S -- one in which most beliefs refer to a number of other beliefs. Then, as William Poundstone (1989) has pointed out, the problem of determining whether a new belief is logically consistent with the belief system S is at least as hard as the well-known problem of "Boolean Satisfiability," or SAT.

Not only is there no known algorithm for solving SAT effectively within a reasonable amount of time; it has been proved that SAT is NP-complete, which means (very roughly speaking) that if there is such an algorithm, then there is also a reasonably rapid and effective algorithm for solving any other problem in the class NP. And the class NP includes virtually every difficult computational problem ever confronted in a practical situation.

So the problem of determining the consistency of a belief with a significantly cross-referential belief system is about as difficult as any computational problemyet confronted in any real situation. To get a vague idea of how hard this is, consider the fact that, using the best algorithms known, and a computer the size of the known universe with processing elements ths size of protons, each working for the entire estimated lifetime of the universe, as fast as the laws of physics allow, it would not be possible to determine the logical consistency of a belief with a significantly cross-referential belief system containing six hundred beliefs.

It must be emphasized that the problem of **making a good guess** as to whether or not a belief is logically consistent with a given belief system is an entirely different matter. What is so astoundingly difficult is getting the exact right answer every time. If one allows oneself a certain proportion of errors, one may well be able to arrive at an answer with reasonable rapidity. Obviously, the rapidity decreases with the proportion of error permitted; the **rate** of this decrease, however, is a difficult mathematical question.

So when a mind determines the functions c_{ij} relating its beliefs, it may take logical consistency into account, but it seems extremely unlikely that it can do so with perfect accuracy, for three reasons: 1) based on experience, the human mind does not appear to be terribly logically consistent; 2) the brain is not an exact mechanism like a computer, and it almost certainly works according to rough probabilistic approximation methods; 3) the problem of determining logical consistency is NP-complete and it is hence very unlikely that it has a rapid, accurate solution for any but the smallest belief systems.

Hence it is unreasonable to require that a system of beliefs be "rational" in structure, at least if rationality is defined in terms of propositional logic. And the structural modifications to propositional logic suggested in Chapter Four only serve to make the problem of determining the c_{ij} even more difficult. In order to compute anything using the structural definition of implication, one has to compute the algorithmic information contained in various sequences, which is impossible in general and difficult in most particular cases.

From these considerations one may conclude that the determination of the functions c_{ij} -- of the structure of a belief system -- is so difficult that the mind must confront it with a rough, approximate method. In particular, I propose that the mind confronts it with a combination of deduction, induction and analogy: that itdoes indeed seek to enforce logical consistency, but lacking an effective general means of doing so, it looks for inconsistency wherever experience tells it inconsistency is most likely to lurk.

9.2 BELIEF SYSTEMS IN THE HISTORY OF SCIENCE

No mind consists of fragmentary beliefs, supported or refuted by testing on an individual basis. In reality, belief is almost always systematic. To illustrate this, let us consider some philosophically interesting examples from the history of science.

In his famous book, *The Structure of Scientific Revolutions*, Thomas Kuhn (1962) proposed that science evolves according to a highly discontinuous process consisting of 1) long periods of "normal science," in which the prevailing scientific belief system remains unchanged, and new beliefs are accepted or rejected largely on the basis of their compatibility with this belief system, and 2) rare, sudden "paradigm changes," in which the old belief system is replaced with a new one.

According to this analysis, the historical tendency of scientists has been to conform to the prevailing belief system until there suddenly emerges a common belief that the process of testing has yielded results which cannot possibly be made compatible with the old system. This point of revolution is called a "crisis." Classic examples of scientific revolution are the switch from Newtonian mechanics to relativity and quantum theory, and the switch from Ptolemaic to Copernican cosmology. This phenomenon is clearest in physics, but it is visible everywhere.

Kuhn never said much about how belief systems work; he placed the burden of explanation on sociology. Imre Lakatos (1978) was much more specific. He hypothesized that science is organized into belief systems called "research programmes," each of which consists of a "hard core" of essential beliefs and a "periphery" of beliefs which serves as a medium between the hard core and the context. According to this point of view, if A is a belief on the periphery of a research programme, and a test is done which decreases its intensity significantly, then A is replaced with an alternate belief A' which is, though incompatible with A and perhaps other peripheralbeliefs, still compatible with the hard core of the programme.

Admittedly, the distinction between "hard core" and "periphery" is much clearer in retrospect that at the time a theory is being developed. In reality, the presence of a troublesome piece of data often leads to much debate as to what is peripheral and what is essential. Nonetheless, Lakatosian analysis can be quite penetrating.

For instance, consider the Ptolemaic research programme, the analysis of the motions of heavenly bodies in terms of circular paths. One could argue that the "hard core" here contains the belief that the circle is the basic unit of heavenly motion, and the belief that the earth is the center of the universe; whereas initially the periphery contained, among other things, the belief that the heavenly bodies revolve around the earth in circular paths.

When testing refuted the latter belief, it was rejected and replaced with another belief that was also compatible with the hard core: the belief that the heavenly bodies move in "epicycles," circles around circles around the earth. And when testing refuted this, it was rejected and replaced with the belief that the heavenly bodies move in circles around circles around circles around the earth -- and so on, several more times. Data was accomodated, but the hard core was not touched.

Consider next the Copernican theory, that the planets revolve in circles around the sun. This retains part but not all of the hard core of the Ptolemaic belief system, and it generates a new periphery. In Copernicus's time, it was not clear why, if the earth moved, everything on its surface didn't fly off. There were certain vague theories in this regard, but not until around the time of Newton was there a convincing explanation. These vague, dilemma-ridden theories epitomize Lakatos's concept of periphery.

Philosophers of science have a number of different explanations of the transition from Ptolemaic to Copernican cosmology. It was not that the Copernican belief system explained the data much better than its predecessor; in fact, it has been argued that, when the two are restricted to the same number of parameters, their explanatory power is approximately equal (Feyerabend, 1970). It was not that there was a sociological "crisis" in the scientific community; therewas merely a conceptual crisis, which is visible only in retrospect. Extant documents reveal no awareness of crisis.

Was it that the Copernican theory was "simpler"? True, a single circle for each planet seems far simpler than a hierarchy of circles within circles within circles within circles.... But the complexity of the Ptolemaic epicycles is rivalled by the complexity of contemporaneous explanations as to how the earth can move yet the objects on its surface not be blown away. As Feyerabend has rightly concluded, there is no single explanation for this change of belief system; however, detailed historical analysis can yield insight into the complex processes involved.

9.2.1. Belief Generation

Lakatos's ideas can easily be integrated into the above-given model of belief systems. The first step is a simple one: belief in an element of the hard core strongly encourages belief in the other theories of the system, and belief in a theory of the system almost never discourages belief in an element of the hard core. There are many ways to formalize this intuition; for example, given an integer p and a number a, one might define the **hard core** of a belief system $\{A_1,...,A_n\}$ as the set of A_i for which the p'th-power **average** over all j of c_{ij} exceeds a. This says that the hard core is composed of those beliefs which many other beliefs depend on.

But unfortunately, this sort of characterization of the hard core is not entirely adequate. What it fails to capture is the way the hard core of a research programme not only supports but actually **generates** peripheral theories. For instance, the hard core of Newtonian mechanics -- the three laws of motion, and the machinery of differential and integral calculus -- is astoundingly adept at producing analyses of particular physical phenomena. One need merely make a few incorrect simplifying assumptions -- say, neglect air resistance, assume the bottom of a river is flat, assume the mass within the sun is evenly distributed, etc. -- and then one has a useful peripheral theory. And when theperipheral theory is refuted, this merely indicates that another "plausible" incorrect assumption is needed.

There is an old story about a farmer who hires an applied mathematician to help him optimize his productivity. The mathematician begins "First, let us assume a spherical cow...," and the farmer fires him. The farmer thinks the mathematician is off his rocker, but all the mathematician is doing is applying a common peripheral element of his belief system. This peripheral element, though absurd in the context of the farmer's belief system, is often quite effective when interpreted in terms of the belief system of modern science. The peripheral theory seems ridiculous "in itself", but it was invented by the hard core for a certain purpose and it serves this purpose well.

For a different kind of example, recall what Newtonian mechanics tells us about the solar system: a single planet orbiting the sun, assuming that both are spherical with uniform density, should move in an ellipse. But in fact, the orbit of Mercury deviates from ellipticity by approximately 43 seconds of arc every century.

This fact can be accomodated within the framework of Newtonian mechanics, for instance by changing the plausible simplifying assumption of uniform spherical mass distribution -- a step which leads to all sorts of interesting, peripheral mathematical theories. In fact, when all known data is taken into account, Newtonian mechanics **does** predict a precession, just a smaller precession than is observed. So it is easy to suppose that, with more accurate data, the exact amount of precession could be predicted.

But eventually, General Relativity came along and predicted the exact amount of the precession of Mercury's orbit "from first principles," assuming a uniform, spherical sun. Now the precession of Mercury's orbit is seen as a result of the way mass curves space -- a notion entirely foreign to Newtonian physics. But that's another story. The point, for now, is that the hard core of a theory can **suggest** or **create** peripheral theories as well as supporting them.

And indeed, it is hard to see how a belief system could survive sustained experimental attack unless **some** of its component beliefs came equipped with significant generative power. If a belief system is to defend itself when one of its beliefs is attacked, it must be able to generate compatible new beliefs to take the place of theold. These generative elements will be helpful to the system over the long term only if they are unlikely to be refuted -- and an element is least likely to be refuted if it is strongly supported by other elements of the system. Therefore, systems with generative hard cores are the "hardiest" systems; the most likely to preserve themselves in the face of experimental onslaught.

The idea of a "generative hard core" may be formalized in many different ways; however, the most natural course is to avail ourselves of the theory of self-generating component systems developed in Chapters Seven and Eight. In other words, I am suggesting that **a scientific belief system, like a linguistic system, is a self-generating structured transformation system**. Belief systems combine these two important system-theoretic structures to form something new, something with dazzling synergetic properties not contain in either structures on its own.

Structured transformation systems unite **deduction** and **analogy** in a striking way, via the connection between grammar and semantics which continuous compositionality enforces. Self-generating systems provide an incredible power for unpredictable, self-organizing creativity. Putting the two together, one obtains, at least in the best case, an adaptable, sturdy tool for exploring the world: adaptable because of the STS part, and sturdy because of the self-generation. This is exactly what the difficult task of science requires.

9.2.2. Conclusion

In the history of science one has a record of the dynamics of belief systems -- a record which, to some extent, brings otherwise obscure mental processes out into the open. It is clear that, in the history of science, belief has been tremendously systematic. Consistently, beliefs have been discarded, maintained or created with an eye toward compatibility with the generative "hard cores" of dominant belief systems. I suggest -- and this is hardly a radical contention -- that this process is not specific to scientific belief, but is rather a general property of thought.

I have suggested that scientific belief systems are self-generating structured transformation systems. In the following sections I will make this suggestion yet more specific: I will propose that **all** belief systems are not only self-generating structured transformationsystems but also **attractors for the cognitive equation**.

But in fact, this is almost implicit in what I have said so far. For consider: beliefs in a system support one another, by definition, but how does this support take place on the level of psychological dynamics? By far the easiest way for beliefs to support one another is for them to **produce** one another. But what do the processes in the dual network produce but **patterns**. Thus a belief system emerges as a collection of mental processes which is **closed under generation and pattern recognition** -- an attractor for the cognitive equation.

What Lakatos's model implies is that belief systems are attractors with a special kind of structure: a two-level structure, with hard core separate from periphery. But if one replaces the rigid hard core vs. periphery dichotomy with a **gradation** of importance, from most central to most peripheral, then one obtains nothing besides a **dual network structure** for belief systems. The hard core is the highest-level processes, the outermost periphery are the lowest-level. Processes are grouped hierarchically for effective production and application; and heterarchically for effective associative reference.

In this way, a belief system emerges as a sort of "mini mind," complete in itself both structurally and dynamically. And one arrives at an enchanting conceptual paradox: only by attaining the ability to survive separately from the rest of the mind, can a belief system make itself of significant **use** to the rest of the mind. This conclusion will return in Chapter Twelve, equipped with further bells and whistles.

9.3. A CONSPIRATORIAL BELIEF SYSTEM

I have discussed some of the most outstanding belief systems ever created by the human mind: Newtonian mechanics, Galilean astronomy, general relativity. Let us now consider a less admirable system of beliefs: the **conspiracy theory** of a woman, known to the author, suffering from paranoid delusion. As I am a mathematician and not a clinical psychologist, I am not pretending to offer a "diagnosis" of the woman possessing this belief system. My goal is merely to broaden our conceptual horizons regarding the nature of psychodynamics, by giving a specific example to back upthe theoretical abstractions of the cognitive equation and the dual network.

9.3.1. Jane's Conspiratorial Belief System

"Jane" almost never eats because she believes that "all her food" has been poisoned. She has a history of bulimia, and she has lost twenty-five pounds in the last month and a half; she is now 5'1" and eighty five pounds. She believes that any food she buys in a store or a restaurant, or receives at the home of a friend, has been poisoned; and when asked who is doing the poisoning, she generally either doesn't answer or says, accusingly, "**You** know!" She has recurrent leg pains, which she ascribes to food poisoning.

Furthermore, she believes that the same people who are poisoning her food are following her everywhere she goes, even across distances of thousands of miles. When asked how she can tell that people are following her, she either says "I'm not stupid!" or explains that they give her subtle hints such as wearing the same color clothing as her. When she sees someone wearing the same color clothing as she is, she often assumes the person is a "follower," and sometimes confronts the person angrily. She has recently had a number of serious problems with the administration of the college which she attends, and she believes that this was due to the influence of the same people who are poisoning her food and following her.

To give a partial list, she believes that this conspiracy involves: 1) a self-help group that she joined several years ago, when attending a college in a different part of the country, for help with her eating problems; 2) professors at this school, from which she was suspended, and which she subsequently left; 3) one of her good friends from high school.

Her belief system is impressively resistant to test. If you suggest that perhaps food makes her feel ill because her long-term and short-term eating problems have altered her digestive system for the worse, she concludes that you must be either stupid or part of the conspiracy. If you remind her that five years ago doctors warned her that her leg problem would get worse unless she stopped running and otherwise putting extreme pressure on it, and suggest that perhaps her leg would be better if she stopped working as a dancer, she concludes that you must be either stupid or part of the conspiracy. If yousuggest that her problems at school may have partly been due to the fact that she was convinced that people were conspiring against her, and consequently acted toward them in a hostile manner -- she concludes that you must be either stupid or part of the conspiracy.

9.3.2. Jane and the Cognitive Equation

I have analyzed the structure of Jane's conspiracy theory; now how does this relate to the "cognitive equation of motion" given in Chapter Eight. Recall that this equation, in it simplest incarnation, says roughly the following:

1) Let all processes that are "connected" to one another act on one another.

2) Take all patterns that were recognized in other processes during Step (1), let these patterns be the new set of processes, and return to Step (1).

An **attractor** for this dynamic is then a set of processes X with the property that each element of the set is a) produced by the interaction of some elements of X, b) a pattern in the set of entities produced by the interactions of the elements of X.

In order to show that Jane's belief system is an attractor for this dynamic, it suffices to show that each element of the belief system is a pattern among other elements of the system, and is potentially producible by other elements of the system. Consider, for instance, the seven beliefs

C_0: There is a group conspiring against me

C_1: My food is poisoned by the conspiracy

C_2: My friends and co-workers are part of the conspiracy

C_3: My leg pain is caused by the conspiracy

C_4: My food tastes bad

C_5: My friends and co-workers are being unpleasant to me

C_6: My leg is in extreme pain

In the following discussion, it will be implicitly assumed that each of these beliefs is stored **redundantly** in the brain; that each one is contained in a number of different "neural maps" or "mental processes." Thus, when it is said that C_0, C_1, C_2 and C_6 "combine to produce" C_3, this should be interpreted to mean that **a certain percentage of the time**, when these four belief-processes come together, the belief-process C_3 is the result.

Furthermore, it must be remembered that each of the brief statements listed above next to the labels C_i is only a shorthand way of referring to what is in reality a diverse collection of ideas and events. For instance, the statement "my co-workers are being unpleasant to me" is **shorthand** for a conglomoration of memories of unpleasantness. Different processes encapsulating C_5 may focus on different specific memories.

Without further ado, then, let us begin at the beginning. Obviously, the belief C_0 is a pattern among the three beliefs which follow it. So, suppose that each of the mental processes corresponding to C_1, C_2 and C_3 is equipped with a generalization routine of the form "When encountering enough other beliefs that contain a certain sufficiently large component in common with me, create a process stating that this component often occurs." If this is the case, then C_0 may also be **created** by the cooperative action of C_1, C_2 and C_3, or some binary subset thereof.

One might wonder why the process corresponding to, say, C_1 should contain a generalization routine of this type. The only answer is that such routines are of general utility in intelligent systems, and that they add only negligible complexity to a process such as C_1 which deals with such formidable concepts as "food" and "conspiracy." In a self-organizing model of the mind, one may not assume that recognitive capacity is contained in a single "generalization center"; it must be achieved in a highly distributed way.

9.3.2.1. Production of Particular Conspiracies

Next, what about C_1? Taking C_0, C_2, C_3 and C_4 as given, C_1 is a fairly natural inference. Suppose the process corresponding to C_0 contains a probabilistic generalization routine of the form "The greater the number of events that have been determined to be caused by conspiracy, the more likely it is that event X is caused by conspiracy." Then when C_0 combines with C_2 and C_3, it will have located two events determined to be caused by conspiracy. And when this compound encounters C_4, the generalization capacity of C_0 will be likely to lead to the creation of a belief such as C_1.

So C_1 is **produced** by the cooperative action of these four beliefs. In what sense is it a **pattern** in the other beliefs? It is a pattern because it **simplifies** the long list of events that are summarized in the simplestatement "My food is being poisoned." This statement encapsulates a large number of different instances of apparent food poisoning, each with its own list of plausible explanations. Given that the concept of a conspiracy is already **there**, the attribution of the poisoning to the conspiracy provides a tremendous simplification; instead of a list of hypotheses regarding who did what, there is only the single explanation "**They** did it." Note that for someone without a bent toward conspiracy theories (without a strong C_0), the cost of supplying the concept "conspiracy" would sufficiently great that C_1 would **not** be a pattern in a handful of cases of apparent food poisoning. But for Jane, $I(C_4|C_1,C_0) < I(C_4|C_0)$. Relative to the background information C_0, C_1 simplifies C_4.

Clearly, C_2 and C_3 may be treated in a manner similar to C_1.

9.3.2.2. Production of Actual Events

Now let us turn to the last three belief-processes. What about C_5, the belief that her co-workers are acting unpleasantly toward her? First of all, it is plain that the belief C_2 works to produce the belief C_5. If one believes that one's co-workers are conspiring against one, one is far more likely to interpret their behavior as being unpleasant.

And furthermore, **given** C_2, the more unpleasant her co-workers are, the **simpler** the form C_2 can take. If the co-workers are acting pleasant, then C_2 has the task of explaining how this pleasantry is actually false, and is a form of conspiracy. But if the co-workers are acting unpleasant, then C_2 can be vastly simpler. So, in this sense, it may be said that C_5 is a pattern in C_2.

By similar reasoning, it may be seen that C_4 and C_6 are both **produced by** other beliefs in the list, and **patterns in or among** other beliefs in the list.

9.3.2.3. Jane's Conspiracy as a "Structural Conspiracy"

The arguments of the past few paragraphs are somewhat reminiscent of R.D. Laing's *Knots* (1972), which describes various self-perpetuating interpersonal and intrapersonal dynamics. Some of Laing's "knots" have been cast in mathematical form by Francisco Varela (1978). However, Laing's "knots" rather glibly treat self-referential dynamics in terms of propositionallogic, which as we have seen is of dubious psychological value. The present treatment draws on a far more carefully refined model of the mind.

It follows from the above arguments that Jane's conspiratorial belief system is in fact a **structural conspiracy**. It is approximately a fixed point for the "cognitive law of motion." A more precise statement, however, must take into account the fact that the specific contents of the belief-processes C_i are constantly shifting. So the belief system is not exactly **fixed**: it is subject to change, but only within certain narrow bounds. It is a **strange attractor** for the law of motion.

Whether it is a **chaotic** attractor is not obvious from first principles. However, this question could easily be resolved by computer simulations. One would need to assume particular **probabilities** for the creation of a given belief from the combination of a certain group of beliefs, taking into account the variety of possible belief-processes falling under each general label C_i. Then one could simulate the equation of motion and see what occurred. My strong suspicion is that there is indeed chaos here. The specific beliefs and their strengths most likely fluctuate pseudorandomly, while the overall conspiratorial structure remains the same.

9.3.3. Implication and Conspiracy (*)

As an aside, it is interesting to relate the self-production of Jane's belief system with the notion of **informational implication** introduced in Chapter Four. Recall that A significantly implies B, with respect to a given deductive system, if there is some chain of deductions leading from A to B, which uses A in a fundamental way, and which is at least as simple as other, related chains of deductions. What interests us here is how it is possible for two entities to significantly imply **each other**.

Formally, "A implies B to degree K" was written as $A \dashrightarrow_K B$, where K was defined as the minimum of $cL + (1-c)M$, for any sequence Y of deductions leading from A to B (any sequence of expressions

$A=B_0,B_1,...,B_n=B$, where B_{i+1} follows from B_i according to one of the transformation rules of the deductive system in question). L was the ratio $|B|/|Y|$, and M was a conceptually simple but formally messy measure of how much additional simplicity Y provides over those otherproofs that are very similar to it. Finally, c was some number between 0 and 1, inserted to put the quantities L and M on a comparable "scale."

For sake of simplicity, let us rechristen the beliefs C_1, C_2 and C_3 as "F," "W," and "L" respectively. In other words, L denotes the hypothesis that the leg pain is due to a conspiracy, W denotes the hypothesis that the work and social problems are due to a conspiracy, and F denotes the hypothesis that the food problems are due to a conspiracy.

Phrased in terms of implication, the self-generating dynamics of Jane's belief system would seem to suggest

$$(L \text{ and } W) \dashrightarrow_{K(F)} F$$

$$(F \text{ and } W) \dashrightarrow_{K(L)} L$$

$$(F \text{ and } L) \dashrightarrow_{K(W)} W$$

where the degrees K(F), K(L) and K(W) are all non-negligible. But how is this possible?

Let Y(F) denote the "optimal sequence" used in the computation of K(F); define Y(L) and Y(W) similarly. One need not worry about exactly what form these optimal sequences take; it is enough to state that the "deductive system" involved has to do with Jane's personal belief system. Her belief system clearly includes an analogical transformation rule based on the idea that if one thing is caused by a conspiracy, then it is likely that another thing is too, which transforms statements of the form "A is likely caused by a conspiracy" into other statements of the form "A and ___ are likely caused by a conspiracy."

Then, it is clear that L(Y) cannot be large for all of these Y, perhaps not for any of them. For one has

$$L[Y(F)] = |F|/|Y(F)| < |F|/[|L|+|W|]$$

$$L[Y(W)] = |W|/|Y(W)| < |W|/[|L|+|F|]$$

$$L[Y(L)] = |L|/|Y(L)| < |L|/[|F|+|W|]$$

For example, if each of the conspiracy theories is of equal intuitive simplicity to Jane, then all these L(Y)'s are less than 1/3. Or if, say, the work theory is **twice** as simple than the others, then L[Y(W)] may be close to 1, but L[Y(F)] and L[Y(L)] are less than 1/4. In any case, perhaps sometimes the **most "a priori" plausible** of the beliefs may attain a fairly large K by having a fairly large L, but for the others a large K must be explained in terms of a large M.

So, recall how the constant M involved in determining the degree K in A $\dashrightarrow_K$ B was defined -- as the weighted sum, over all proofs Z of B, of L(Z). The **weight** attached to Z was determined by I(Z|Y), i.e. by how similar Z is to Y. A power p was introduced into the weight functions, in order to control how little the those Z that are extremely similar to Y are counted.

If M[Y(W)] is large, this means that the theory that a conspiracy is responsible for Jane's work problems is much **simpler** than other theories similar to it. This can be taken in two ways. If p is very large, then M basically deals only with proofs that are virtually identical to Y. On the other hand, if p is moderate in size, then M will incorporate a comparison of the simplicity granted by Y(W) with the simplicity of true alternatives, such as the theory that **Jane** herself is responsible for her work problems. Now, to almost any other person, it would be **very simple indeed** to deduce Jane's work problems from Jane's personality. But to **Jane** herself, this deduction is not at all intuitive.

So, formally speaking, Jane's circular implication can be seen to come from two sources. First of all, a very large p, which corresponds to a very lenient definition of what constitutes a "natural proof" of something. Or, alternately, a blanket negative judgement of the simplicity of all alternative theories. Both of these alternatives amount to the same thing: excessive self-trust, non-consideration of alternative hypotheses ... what I will call **conservatism**.

So, in sum, the informational-implication approach has not given us terribly much by way of new insight into Jane's situation. What I **have** shown, on the other hand, is that Jane's real-life delusional thinking fits in very nicely with the **formal** theory of reasoning given in Chapter Four. This sort of correspondence between theory and everyday reality is precisely what the standard Boolean-logic approach to reasoning lacks.

9.4 BELIEF AND RATIONALITY

Jane's belief system is clearly, according to the standards of modern "sane" society, irrational. It is worth asking how this irrationality is tied in with the **dynamical** properties of the belief system, as discussed in the previous section. This investigation will leadtoward a strikingly general dynamical formulation of the concept of rationality.

9.4.1. Conservatism and Irrelevance

The irrationality of Jane's belief system manifests itself in two properties. First of all, Jane is simply **too glib** in her generation of theories. Given any unpleasant situation, her belief system has no problem whatsoever reeling off an explanation: the theory is always "the conspirators did it." New events never require new explanations. No matter how different one event is from another, the explanation never changes. Let us call this property **conservatism**.

To put it abstractly, let E_s denote the collection of beliefs which a belief system generates in order to explain an event s. That is, when situation s arises, E_s is the set of **explanatory** processes which the belief system generates. Then one undesirable property of Jane's belief system is that the rate of change of E_s with respect to s is simply too small.

The second undesirable property of Jane's belief system is, I suggest, that the theories created to explain an event never have much to do with the specific structure of the event. Formally, the collection of patterns which emerge between E_s and s is invariably very small. Her belief system explains an event in a way which has nothing to do with the details of the actual nature of the event. Let us call this property **irrelevance**.

Of course, Jane would reject these criticisms. She might say "I don't need to change my explanation; I've always got the right one!" A dogmatist of this sort is the exact opposite of the prototypical skeptic, who trusts nothing. The skeptic is continually looking for holes in every argument; whereas Jane doesn't bother to look for holes in any argument. She places absolute trust in one postulate, and doesn't even bother to look for holes in arguments purporting to contradict it, for she simply "knows" the holes are there.

This attitude may be most easily understood in the context of the mathematical theory of pattern. The pattern-theoretic approach to intelligence assumes that the environment is chaotic on the level of detailed numerical paramters, but roughly **structurally predictable**. In Charles S. Peirce's phrase, it assumes that the world possesses a "tendency to take habits."

Under this assumption, it is clear that conservatism and irrelevance and reluctance to test are, in any given case, fairly **likely** to be flaws. First of all because if change is likely, if old ideas are not necessarily true for the future, then a belief system which does not change is undesirable. And secondly because if induction is imperfect, and the mind works by induction, then one must always face the fact that one's own conclusions may be incorrect.

9.4.2. The Genesis of Delusion

Why, exactly, is Jane's belief system conservative and irrelevant? To answer this, it is convenient to first ask how Jane's mind ever got **into** the irrational attractor which I have described.

The beginning, it seems, was an instance of C_5 and C_2: a professor at school was asking her questions relating to her Overeaters Anonymous group, and she came to the conclusion that people were talking about her behind her back. Whether or not this initial conspiracy was real is not essential; the point is that it was nowhere nearly as unlikely as the conspiracies imagined by her later.

Even if no real conspiracy was involved, I would not say that this first step was "unjustified". It was only a guess; and there is nothing unjustified about making a wrong guess. After all, the mind works largely by trial and error. What is important is that Jane's initial belief in a conspiracy was not strongly incompatible with the remainder of her sane, commonsensical mind.

After this, all that were needed were a few instances of C_4 or C_6, and a few more instances of C_5. This caused the creation of some C_0 belief-processes; then the feedback dynamics implicit in the analysis of the previous section kicked in. The point is that only a small number of C_i are necessary to start a cybernetic process leading to a vast proliferation. Eventually C_0 became so strong that plausible stories about conspiracies were no longer necessary; an all-purpose "them" was sufficient.

Most of us weather unpleasant experiences without developing extravagant conspiracy theories. In the initial stages of its growth, Jane's conspiratorial belief system depended crucially on certain other aspects of Jane's personality; specifically, on her absolute refusal to accept any responsibility for her misfortunes. But once this early phase was past, the spread of her belief system may have had little to do with the remainder of her mind. It may have been a process of isolated expansion, like the growth of a cancer.

9.4.3. Rationality and Dynamics

The lesson is that irrational belief systems are self-supporting, self-contained, integral units. Considered as attractors, they are just as genuine and stable as the belief systems which we consider "normal." The difference is that they gain **too much** of their support from internal self-generating dynamics -- they do not draw enough on the remainder of the mental process network.

This is perhaps the most **objective** test of rationality one can possibly pose: how much support is internal, and how much is external? Excessive internal support is clearly inclined to cause conservatism and irrelevance. In this way the **irrationality** of a person's mind may be traced back to the existence of **overly autonomous subattractors** of the cognitive equation. The mind itself is an attractor of the cognitive equation; but small portions of the mind may also be attractors for this same equation. When a portion of the mind survives because it is itself an attractor, rather than because of its relations with the rest of the mind, there is a significant danger of irrationality.

Looking ahead to Chapter Twelve, another way to put this is as follows: **irrationality is a consequence of dissociation**. This formulation is particularly attractive since dissociation has been used as an explanation for a variety of mental illnesses and strange psychological phenomena -- schizophrenia, MPD, post-traumatic stress syndrome, cryptomnesia, hypnosis, hysterical seizure, etc. (Van der Kolk et al, 1991). The general concept of dissociation is that of a "split" in the network of processes that makes up the mind. Here I have shown that this sort of split may arise due to the dynamical autonomy of certain collections of processes.

9.5. MONOLOGUE AND DIALOGUE

Consider once again Galileo's belief that what one sees when one points a telescope out into space is actually there. As noted above, this seems quitereasonable from today's perspective. After all, it is easy to check that when one points a telescope toward an earthbound object, what one sees is indeed there. But we are accustomed to the Newtonian insight that the same natural laws apply to the heavens and the earth; and the common intuition of Galileo's time was quite the opposite. Hence Galileo was going against commonsense logic.

Also, it was said at the time that he was making hypotheses which could not possibly be proven, merely dealing in speculation. Now we see that this objection is largely unfounded; we have measured the heavens with radio waves, we have sent men and robotic probes to nearby heavenly bodies, and the results agree with what our telescopes report. But to the common sense of Galileo's time, the idea of sending men into space was no less preposterous than the notion of building a time machine; no less ridiculous than the delusions of a paranoiac.

Furthermore, it is now known that Galileo's maps of the moon were drastically incorrect; so it is not exactly true that what he saw through his primitive telescopes was actually there!

Galileo argued that the telescope gave a correct view of space because it gave a correct view of earth; however, others argued that this analogy was incorrect, saying "when the telescope is pointed toward earth, everyone who looks through it saw the same thing; but when it's pointed toward space, we often see different things."

Now we know enough about lenses and the psychology of perception to make educated guesses as to the possible causes of this phenomenon, reported by many of those who looked through Galileo's telescopes. But at the time, the only arguments Galileo could offer were of the form "There must be something funny going on either in your eye or in this particular lense, because what is seen through the telescope in the absence of extraneous interference is indeed truly, objectively there." In a way, he reasoned dogmatically and ideologically rather than empirically.

How is Galileo's belief system intrinsically different from the paranoid belief system discussed above? Both ignore common sense and the results of tests, and both are founded on "wild" analogies. Was Galileo's train of thought just as crazy a speculation as Jane's, the only difference being that Galileo was lucky enough to be "right"? Or is it more accurate to saythat, whereas both of them blatantly ignored common logic in order to pursue their intuitions, Galileo's intuition was better than Jane's? I find the latter explanation appealing, but it begs the question: was the superiority of Galileo's intuition somehow related to the structure of his belief system?

Whereas Jane's belief system is conservative and irrelevant, Galileo's belief system was **productive**. Once you assume that what you see through the telescope is really out there, you can look at all the different stars and planets and draw detailed maps; you can compare what you see through different telescopes; you can construct detailed theories as to why you see what you see. True, if it's **not** really out there then you're just constructing an elaborate network of theory and experiment about the workings of a particular gadget. But at least the assumption leads to a pursuit of some complexity: it produces **new pattern**. A conspiracy theory, taken to the extreme described above, does no such thing. It gives you access to no new worlds; it merely derides as worthless all attempts to investigate the properties of the everyday world. Why bother, if you already know what the answer will be?

Call a belief system **productive** to the extent that it is correlated with the emergence of new patterns in the mind of the system containing it. I suggest that productivity in this sense is strongly correlated with the "reasonableness" of belief systems. The underlying goal of the next few sections is to pursue this correlation, in the context of the dual network and the cognitive equation.

9.5.1 Stages of Development

One often hears arguments similar to the following: "In the early stages of the development of a theory, anything goes. At this stage, it may be advisable to ignore discouraging test results -- to proceed counterinductively. This can lend insight into flaws in the test results or their standard interpretations, and it can open the way to creative development of more general theories which may incorporate the test results. And it may be advisable to think in bizarre, irrational ways -- so as to generate original hypotheses. But once this stage of discovery is completed and the stage of justification is embarked upon, these procedures are nolonger allowed: then one must merely test one's hypotheses against the data."

Of course, this analysis of the evolution of theories is extremely naive: science does not work by a fragmented logic of hypothesis formation and testing, but rather by a systematic logic of research programmes. But there is obviously some truth to it.

I have suggested that two properties characterize a dogmatic belief system:

1) the variation in the structure of the explanations offered with respect to the events being explained is generally small (formally, $d[St(E_s),St(E_t)]/ d_\#[s,t]$ is generally small, where d and $d_\#$ denote appropriate metrics)

2) the nature of explanations offered has nothing to do with the events being explained (formally, $Em(E_s,s)$ is generally small)

Intuitively, these conditions -- conservatism and irrelevance -- simply mean that the system is not significantly responsive to test. In light of these criteria, I propose the following fundamental normative rule:

During the developmental stage, a belief system may be permitted to be unresponsive to test results (formally, to have consistently small $d[St(E_s)-St(E_t)]/d_\#[s,t]$ and/or $Em(E_s,s)$). However, after this initial stage has passed, this should not be considered justified.

This is a **systemic** rendering of the classical distinction between "context of discovery" and "context of justification."

I will call any belief system fulfilling the conditions of non-conservatism and (sic) non-irrelevance a **dialogical** system. A dialogical system is one which engages in a dialogue with its context. The opposite of a dialogical system is a **monological** system, a belief system which speaks only to itself, ignoring its context in all but the shallowest respects.

A system which is in the stage of development, but will eventually be dialogical, may be called **predialogical**. In its early stage of development, a predialogical system may be indistinguishable from a monological one. Pre-dialogicality, almost by definition, can be established only in retrospect. Human minds and societies deal with the problem of distinguishing monologicality from predialogicality the same way they deal with everything else -- by inductionand analogy, by making educated guesses based on what they've seen in the past. And, of course, these analogies draw on certain belief systems, thus completing the circle and destroying any hope of gleaning a truly objective theory of "justification."

The terms "dialogical" and "monological" are not original; they were used by Mikhail Bakhtin in his analysis of Dostoevsky. The reality of Dostoevsky's novels is called "dialogical," meaning that it is the result of significant interaction between different world-views.

> His path leads not from idea to idea, but from orientation to orientation. To think, for him, means to question and to listen, to try out orientations.... Even **agreement** retains its **dialogic** character ... it never leads to a **merging** of voices and truths in a single **impersonal** truth, as in the monologic world.

Each of Dostoevsky's major novels contains a number of conflicting belief systems -- and the action starts when the belief systems become **dialogical** in the sense defined here. They test each other, and produce creative explanations in response to the phenomena which they provide for each other.

9.5.2. Progressive and Regressive

Lakatos has proposed that good scientific research programmes are "progressive" in that they consistently produce new results which are surprising or dramatic. Bad research programmes are "regressive" in that they do not. This is a valuable analysis, but I don't think it gets to the core of the matter. "Surprising" and "dramatic" are subjective terms; so this criterion doesn't really say much more than "a programme is good if it excites people."

However, I do think that the "monologicity/dialogicity" approach to justification is closely related to Lakatos's notion of progressive and regressive research programmes. It is quite clear that if a system always says the same thing in response to every test, then it is unlikely to give consistently interesting output, and is hence unlikely to be progressive. And I suggest that the converse is also true: that if a system is capable of incorporatingsensitive responses to data into its framework, then it is reasonably likely to say something interesting or useful about the context which generates the data.

Another way to phrase this idea is as follows: in general, **dialogicality and productivity are roughly proportional**. That is: in the real world, as a rule of thumb, any system which produces a lot of new pattern is highly dialogical, and any system which is highly dialogical produces a lot of new pattern.

The second of these assertions follows from the definition of dialogicality. The former, however, does not follow immediately from the nature of belief systems, but only from the general **dynamics of mind**; I will return to it in Section 9.7.

9.5.3 Circular Implication Structure

For a slightly different point of view on these issues, let us think about belief systems in terms of **implication**. Recall the passage above in which I analyzed the **origins** of Jane's paranoid belief system. I considered, among others, the following triad of implications:

My leg pain and my trouble at work are due to

conspiracies, so my problem with food probably is too

My trouble at work and my problem with food are

due to conspiracies, so my leg pain probably is too

My leg pain and my problem with food are due to conspiracies, so my trouble at work probably is too

In formulas, I let L denote the hypothesis that the leg pain is due to a conspiracy, W denote the hypothesis that the work problems are due to a conspiracy, and F denote the hypothesis that the food problems are due to a conspiracy, and I arrived at:

(L and W) --> F

(W and F) --> L

(L and F) --> W

(where each implication, in accordance with the theory of informational implication, had a certain degree determined by the properties of Jane's belief system).

The same basic implication structure can be associated with **any** belief system, not just a conspiratorial belief system. Suppose one has a group of phenomena, and then a group of hypotheses of the form "**this** phenomenon can be explained by my belief system." These hypotheses will support one another if a large number of implications of the form

this and **this** and ... **this**

can be explained by my belief system -->

that can be explained by my belief system

hold **with nontrivially high degree**. Earlier I reviewed conditions under which a collection of implications of this form can hold with nontrivially high degree. Our conclusion was that a high degree of **conservatism** is required: one must, when determining what follows what, not pay too much attention to hypotheses dissimilar to those which one has already conceived. If a high degree of conservatism is present, then it is perfectly possible for a group of beliefs to mutually support each other in this manner.

For a very crude and abstract example, consider the belief that the outside world is real, and the belief that one's body is real. One believes the outside world is real because one **feels** it -- this is G.E. Moore's classic argument, poor philosophy but good common sense ... to prove the external world is really there, kick something! And, on the other hand, why does one believe one's body is real and not an hallucination? **Not** solely because of one's internal kinesthetic feelings, but rather largely because of the sensations one gets when moving one's hand through the air, walking on the ground, and in general interacting with the outside world.

It doesn't take much acumen to see how these two phenomenological "proofs" fit together. If the outside world were an hallucination, then moving one's body through it would be no evidence for the reality of one's body. One has two propositions supporting one another.

According to the dynamics of the dual network, various belief systems will compete for **survival** -- they will compete not to have the processes containing their component beliefs reprogrammed. I suggest that circular support structures are an excellent survival strategy, in that they prevent the conception of hypotheses other than those already contained in the belief system.

But opposed to this, of course, is the fact that the conservatism needed to maintain a circular support structure is **fundamentally incompatible with dialogicality**. Circular support structures and dialogicality are **both** quality survival strategies, and I suggest that both strategies are in competition in most large belief systems. Dialogicality permits the belief system to adapt to new situations, and circular support structures permit the belief system to **ignore** new situations. In order to have long-term success, a belief system must carefully balance these two contradictory strategies -- enough dialogicality to consistently produce interesting new pattern, and enough circular support to avoid being wiped out when trouble arises.

The history of science, as developed by Kuhn, Feyerabend, Lakatos and others, shows that in times of crisis scientific belief systems tend to depend on circular support. In the heyday of Newtonian science, there was a little circular support: scientists believed the Newtonian explanation of W partly because the Newtonian explanations of X, Y and Z were so good, and believed the Newtonian explanation X partly because the Newtonian explanations of W, Y and Z were so good, et cetera. But toward the end of the Newtonian era, many of the actual explanations declined in quality, so that this circular support became a **larger and larger part** of the total evidence in support of each hypothesis of Newtonian explanation.

Circular implication structure is an inevitable consequence of belief systems being attractors for the cognitive equation. But the question is, how much is this **attraction** relied on as the sole source of sustenance for the belief system? If circular support, self-production, is the belief system's main means of support, then the belief system is serving little purpose relative to the remainder of the mind: it is monological. This point will be pursued in more detail in Chapter Twelve.

9.7. DISSOCIATION AND DIALOGUE

So, in conclusion, a belief system is: 1) a miniature dual network structure,

2) a structured transformation system, 3) an attractor for the cognitive equation.

What does this finally say about the proposed correlation between **dialogicality** and **productivity**?

It is, of course, conceivable that a monological system might create an abundance of new pattern. To say that this is highly unlikely is to say that, in actuality, new pattern almost always emerges from significant interaction, from **systematic testing**. But why should this be true?

The correct argument, as I have hinted above, proceeds on grounds of **computational efficiency**. This may at first seem philosophically unsatisfying, but on the other hand it is very much in the spirit of **pattern** philosophy -- after all, the very definition of pattern involves an abstract kind of "computational efficiency."

A monological system, psychologically, represents a **highly dissociated** network of belief processes. This network does of course interact with the remainder of the mind -- otherwise it would have no effects. But it restricts its interactions to those in which it can play an **actor** role; it resists being modified, or entering into symbiotic loops of inter-adjustment. This means that when a monological belief system solves a problem, it must rely **only, or primarily, upon its own resources**.

But the nature of thought is fundamentally **interactive** and **parallel**: intelligence is achieved by the complex interactions of different agents. A dialogical belief system containing N modestly-sized processes can solve problems which are of such an intrinsic computational complexity that **no** excessively dissociated network of N modestly-sized processes can **ever** solve them. For a dialogical system can solve problems by **cooperative computation**: by using its own processes to request contributions from outside processes. A monological system, on the other hand, cannot make a habit of interacting intensively with outside processes -- if it did, it would not be monological.

This, I suggest, is all there is to it. Despite the abstract terminology, the idea is very simple. Lousy, unproductive belief systems are lousy precisely because they keep to themselves; they do not make use of the vast potential for cooperative computation that is implicit in the dual network. This is the root of their conservatism and irrelevance. They are conservative and irrelevant because, confronted with the difficult problems of the real world, **any** belief system of their small size would necessarily be conservative and irrelevant, if it did not extensively avail itself of the remainder of the mind.

All this leaves only one question unanswered: why do monological systems arise, if they are unproductive and useless? The answer to this lies in the cognitive equation. Attractors can be notoriously stubborn. And this leads us onward....

Chapter Ten

BIOLOGICAL METAPHORS OF BELIEF

The train of thought reported in this chapter began in the fall of 1991. My father was writing *Turncoats and True Believers* (Ted Goertzel, 1993), a book about political ideologies, those who abandon them, and those who maintain them; he was collecting anecdotes from a variety of biographies and autobiographies, and he was struck by the recurrent patterns. In some intuitively clear but hard-to-specify sense, ideologues of all different stripes seemed to think alike.

My father has studied ideology for nearly a quarter century, and his approach is thoroughly rationalist: he believes that ideological belief systems coincide with irrational thought, whereas nonideological belief systems coincide with rational thought. This rationalism implies that adherents to **nonideological** belief systems should all think alike -- they are all following the same "correct" form of logical reasoning. But it says nothing about the nature of **irrationality** -- it does not explain why deviations from "correct" logical reasoning all seem to follow a few simple psychological forms.

He hoped to resolve the puzzle by coming up with a "litmus test" for belief systems -- a property, or a list of properties, distinguishing irrational, ideological reasoning from rational thought. For example, two properties under tentative consideration for such a list were:

1) adherents to ideological belief systems tend to rely on reasoning by analogy rather than logical deduction

2) adherents to ideological belief systems tend to answer criticism by reference to "hallowed" texts, such as the Bible or das Kapital.

But both of these properties were eventually rejected: the first because analogy is an essential part of logical deduction (as shown in Chapter Four); and thesecond because reference to hallowed texts is really a surface symptom, not a fundamental flaw in reasoning.

Every property that he came up with was eventually discarded, for similar reasons. Eventually he decided that, given these serious conceptual troubles, *Turncoats and True Believers* would have to do without a formal theory of justification -- a decision that probably resulted in a much more entertaining book! The present chapter, however, came about as a result of my continued pursuit of an explanation of the difference between "rational" and "ideological" thought.

I will not discuss political belief systems here -- that would take us too far afield from the cognitive questions that are the center of this book. However, the same questions that arise in the context of political belief systems, also emerge from more general psychological considerations. For I have argued that strict adherence to formal logic does not characterize sensible, rational thought -- first because formal logic can lead to rational absurdities; and second because useful applications of formal logic require the assistance of "wishy-washy" **analogical** methods. But if formal logic does not define rationality -- then **what does**?

In this chapter I approach rationality using ideas drawn from evolutionary biology and immunology. Specifically, I suggest that old-fashioned rationalism is in some respects similar to **Neo-Darwinism**, the evolutionary theory which holds the "fitness" of an organism to be a property of the organism in itself. Today, more and more biologists are waking up to the sensitive **environment-dependence** of fitness, to the fact that the properties which make an organism fit may not even be **present** in the organism, but may be **emergent** between the organism and its environment. And similarly, I propose, the only way to understand reason is to turn the analogy-dependence of logic into a **tool** rather than an obstacle, and view rationality as a as a property of the **relationship** between a belief system and its "psychic environment."

In order to work this idea out beyond the philosophical stage, one must turn to the dual network model. Productivity alone does not guarantee the survival of a belief system in the dual network. And unproductivity does not necessarily mitigate **against** the survival of a belief system. What then, I asked, **does**determine survival in the complex environment that is the dual-network psyche?

There are, I suggest, precisely **two** properties common to successful belief systems:

1) being an attractor for the cognitive equation

2) being productive, in the sense of creatively constructing new patterns in response to environmental demands

A belief system cannot survive unless it meets **both** of these criteria. But some belief systems will rely more on (1) for their survival, and some will rely more on (2). Those which rely mainly on (1) tend to be monological and irrational; those which rely mainly on (2) are dialogical, rational and useful. This is a purely **structural** and **systemic** vision of rationality: it makes no reference to the specific contents of the belief systems involved, nor to their connection with the external, "real" world, but only to their relationship with the rest of the mind.

In this chapter I will develop this approach to belief in more detail, using complex **biological** processes as a guide. First I will explore the **systematic creativity** inherent in belief systems, by analogy to the phenomenon of **evolutionary innovation** in ecosystems. Then, turning to the question of a belief system interacts with the rest of the mind, I will present the following crucial analogy: belief systems are to the mind as the immune system is to the body. In other words, belief systems **protect** the upper levels of the mind from dealing with trivial ideas. And, just like immune systems, they maintain themselves by a process of **circular reinforcement**.

In addition to their intrinsic value, these close analogies between belief systems and biological systems are a powerful argument for the existence of nontrivial **complex systems science**. Circular reinforcement, self-organizing protection and evolutionary innovation are deep ideas with relevance transcending disciplinary bounds. The ideas of this chapter should provide new ammunition against those who would snidely assert that "there is no general systems theory."

10.1. SYSTEMATIC CREATIVITY

As suggested in the previous chapter, a complex belief system such as a scientific theory may be modeledas a **self-generating structured transformation system**. The hard core beliefs are the initials I, and the peripheral beliefs are the elements of D(I,T). The transformations T are the processes by which peripheral beliefs are generated from hard core beliefs. And all the elements of D(I,T) are "components," acting on one another according to the logic of self-generating component-systems.

For example, in the belief systems of modern physics, many important beliefs may be expressed as equational models. There are certain situation-dependent rules by which basic equational models (Maxwell's Laws, Newton's Laws, the Schrodinger Equation) can be used to generate more complex and specific equational models. These rules are what a physicist needs to know but an engineer (who **uses** the models) or a mathematician (who develops the math **used by** the models) need not. The **structuredness** of this transformation system is what allows physicists to do their work: they can build a complex equational model out of simpler ones, and predict some things about the behavior of the complex one from their knowledge about the behavior of the simpler ones.

On the other hand, is the conspiratorial belief system presented above not also a structured transformation system? Technically speaking, it fulfills all the requirements. Its hard core consists of one simple conspiracy theory, and its D(I,T) consists of beliefs about psychological and social structures and processes. Its T contains a variety of different methodologies for generating situated conspiracy beliefs -- in fact, as a self-generating component-system, its power of spontaneous invention can be rather impressive. And the system is structured, in the sense required by continuous compositionality: similar phenomena correspond to similar conspiracy theories. Yes, this belief system is an STS, though a relatively uninteresting one.

In order to rule out cases such as this, one might add to the definition of STS a requirement stating that the set D(I,T) must meet some minimal standard of structural complexity. But there is no pressing need to do this; it is just as well to admit simplistic STS's, and call them simplistic. The important observation is that **certain** belief systems generate a **high** structural complexity from applying their transformation rules to one another and their initials -- just as written and spoken language systems generate a high structuralcomplexity from combining their **words** according to their grammatical words.

And the **meanings** of the combinations formed by these productive belief systems may be determined, to a high degree of approximation, by the **principle of continuous compositionality**. As expressions are becoming complex, so are their meanings, but in an approximately predictable way. These productive belief systems respond to their environments by continually creating **large quantities of new meaning**.

Above it was proposed that, in order to be **productive**, in order to survive, a belief system needs a **generative hard core**. A generative hard core is, I suggest, synonymous with a hard core that contains an effective set of "grammatical" transformation rules -- rules that take **in** the characteristics of a particular situation and put **out** expressions (involving hard core entities) which are tailored to those particular situations. In other words, the way the component-system which is a belief system works is that **beliefs, using grammatical rules, act on other beliefs to produce new beliefs**. Grammatical rules are the "middleman"; they are the part of the definition of f(g) whenever f and g are beliefs in the same belief system.

And what does it mean for an expression E to be "tailored to" a situation s? Merely that E and s **fit** together, in the sense that they help give rise to significant emergent patterns in the set of pairs {(E,s)}. That a belief system has a generative hard core means that, interpreted as a language, it is **complex** in the sense introduced in the previous paragraph -- that it habitually creates significant quantities of meaning.

The situatedness of language is largely responsible for its power. One sentence can mean a dozen different things in a dozen different contexts. Similarly, the situatedness of hard core "units" is responsible for the power of productive belief systems. One hard core expression can mean a dozen different things in a dozen different situations. And depending upon the particular situation, a given word, sentence or hard core expression, will give rise to **different** new expressions of possibly great complexity. To a degree, therefore, beliefs may be thought of as **triggers**. When flicked by external situations, these triggers release appropriate **emergent patterns**. The emergent patterns are not **in** the belief, nor are they **in** the situation; they are fundamentally a synergetic production.

10.1.1. Evolutionary Innovation

To get a better view of the inherent creativity of belief systems, let us briefly turn to one of the central problems of modern theoretical biology: **evolutionary innovation**. How is it that the simple processes of mutation, reproduction and selection have been able to create such incredibly complex and elegant forms as the human eye?

In *The Evolving Mind* two partial solutions to this problem are given. These are of interest here because, as I will show, the problem of evolutionary innovation has a close relation with the productivity of belief systems. This is yet another example of significant parallels among different complex systems.

The first partial solution given in *EM* is the observation that sexual reproduction is a surprisingly efficient optimization tool. Sexual reproduction, unlike asexual reproduction, is more than just random stabbing out in the dark. It is **systematic** stabbing out in the dark.

And the second partial solution is the phenomenon of **structural instability**. Structural instability means, for instance, that when one changes the genetic code of an organism slightly, this can cause disproportionately large changes in the appearance and behavior of the organism.

Parallel to the biological question of evolutionary innovation is the **psychological** question of evolutionary innovation. How is it that the simple processes of pattern recognition, motor control and associative memory give rise to such incredibly complex and elegant forms as the Fundamental Theorem of Calculus, or the English language?

One may construct a careful argument that the two resolutions of the biological problem of evolutionary innovation also apply to the psychological case. For example, it is shown that the multilevel (perceptual-motor) control hierarchy naturally gives rise to an abstract form of sexual reproduction. For, suppose process A has subsidiary processes W and X, and process B has subsidiaries X and Y. Suppose A judges W to work better than X, and reprograms W to work like X. Then, schematically speaking, one has

$A(t) = A', W, X$

$B(t) = B', X, Y$

$A(t+1) = A', W, W$

$B(t+1) = B', W, Y$

(where A' and B' represent those parts of A and B respectively that are not contained in W, X or Y). The new B, B(t+1), contains part of the old A and part of the old B -- it is related to the old A and B as a child is related to its parents. This sort of reasoning can be made formal by reference to the theory of genetic algorithms.

Sexual reproduction is an important corollary of the behavior of multilevel control networks. Here, however, our main concern will be with **structural instability**. Let us begin with an example from A. Lima de Faria's masterful polemic, *Evolution Without Selection* (1988). As quoted in *EM*, Lima de Faria notes that

> the 'conquest of the land' by the vertebrates is achieved by a tenfold increase in thyroid hormone levels in the blood of a tadpole. This small molecule is responsible for the irreversible changes that oblige the animal to change from an aquatic to a terrestrial mode of life. The transformation involves the reabsorption of the tail, the change to a pulmonary respiration and other drastic modifications of the body interior.... If the thyroid gland is removed from a developing frog embryo, metamorphosis does not occur and the animal continues to grow, preserving the aquatic structures and functions of the tadpole. If the thyroid hormone is injected into such a giant tadpole it gets transformed into a frog with terrestrial characteristics....
>
> There are species of amphibians which represent a fixation of the transition stage between the aquatic and the terrestrial form. In them, the adult stage, characterized by reproduction, occurs when they still have a flat tail, respire by gills and live in water. One example is... the mud-puppy.... Another is... the Mexican axolotl.
>
> The demonstration that these species represent transitional physiological stages was obtained by administering the thyroid hormone to axolotls. Following this chemical signal their metamorphosis proceeded and they acquired terrestrial characteristics (round tail and aerial respiration). (p. 241)

This is a sort of paradigm case for the creation of new form by structural instability. The structures inherent in water-breathing animals, if changed only a little, become adequate for the breathing of air. And then, once a water-breathing animal comes to breathe air, it is of course prone to obtain a huge variety of other new characteristics. A small change in a small part of a complex network of processes, can lead to a large ultimate modification of the product of the processes.

In general, consider any process that takes a certain "input" and transforms it into a certain "output." The process is **structurally unstable** if changing the process a little bit, or changing its input a little bit, can change the **structure** of (the set of patterns in) the output by a large amount. This property may also be captured formally: in the following section, the **first innovation ratio** is defined as the amount which changing the nature of the process changes the structure of the output, and the **second innovation ratio** is defined as the amount which changing the nature of the input changes the structure of the output.

When dealing with structures generated by structurally unstable processes, it is easy to generate completely new forms -- one need merely "twiddle" the machinery a bit. Predicting what these new forms will be is, of course, another matter.

10.1.1.1. The Innovation Ratios (*)

Let y and y' be any two processes, let z and z' be any two entities, and let e.g. y*z denote the outcome of executing the process y on the entity z. For instance, in *EM* y and y' denote genetic codes, z and z' are sets of environmental stimuli, and y*z and y'*z' represent the organisms resultant from the genetic codes y and y' in the environments z and z'. Then the essential questions regarding the creation of new form are:

1) what is the probability distribution of the "first innovation ratio"

$$\mathbf{d(S(y*z),S(y'*z))/d_{\#}(y,y')}?$$

That is: in general, when a process is changed by a certain amount, how much is the structure of the entities produced by the process changed? (d and $d_{\#}$ denote appropriate metrics.)

2) what is the probability distribution of the "second innovation ratio"

$$\mathbf{d(S(y*z),S(y*z'))/d_{\#}(z,z')}?$$

That is: when an entity is changed by a certain amount, how much is the structure of the entity which the process y **transforms that entity into** changed? For example, how much does the **environment** affect the structure of an organism?

If these ratios were never large, then it would be essentially impossible for natural selection to give rise to new form.

In *EM* it is conjectured that, where z and z' represent environments, y and y' genetic codes, and y*z and y'*z' organisms, natural selection **can** give rise to new form. This is not purely a mathematical conjecture. Suppose that for an arbitrary genetic code the innovation ratios had a small but non-negligible chance of being large. Then there may well be specific "clusters" of codes -- specific regions in process space -- for which the innovation ratio **is** acceptably likely to be large. If such clusters do exist, then, instead of a purely mathematical question, one has the biological question of whether real organisms reside in these clusters, and how they get there and stay there.

The **structural instability** of a process y may be defined as the average, over all y', of $d(S(y*z),S(y'*z))/d_{\#}(y,y') + d(S(y*z),S(y*z'))/d_{\#}(z,z')$ [i.e. of the sum of the first and second innovation ratios]. In a system which evolves at least partly by natural selection, the tendency to the creation of new form may be rephrased as **the tendency to foster structurally unstable processes**.

Several mathematical examples of structurally unstable processes are discussed in *EM*. It has been convincingly demonstrated that one-dimensional cellular automata can display a high degree of structural instability. And it is well-known that nonlinear iterated function systems can be structurally unstable; this is the principle underlying the oft-displayed Mandelbrot set.

10.1.2. Structural Instability of Belief Systems

Now, let us see how structural instability ties in with the concepts of **monologicity** and **dialogicality**. Onemay consider the hard core of a belief system as a collection of processes y_1, y_2,.... Given a relevant phenomenon z, one of the y_i creates an **explanation** that may be denoted y_i*z. If similar phenomena can have dissimilar explanations, i.e. if y_i*z can vary a lot as z varies a little, then this means that the second innovation ratio is large; and it also fulfills **half** of the definition of dialogicality -- it says that the explanation varies with the phenomenon being explained.

The other half of the definition of dialogicality is the principle of **relevance** -- it says that $Em(y_i*z,z)$ should be nontrivial; that the explanation should have something to do with the phenomenon being explained. Part of the difficulty with maintaining a productive belief system is the tension between creativity-promoting structural instability and the principle of relevance.

And what does the **first** innovation ratio have to do with belief systems? To see this, one must delve a little deeper into the structure of belief systems. It is acceptable but coarse to refer to a belief system as a collection of processes, individually generating explanations. In reality a complex belief system always has a complex network structure.

Many explanation-generating procedures come with a collection of subsidiary procedures, all related to each other. These subsidiaries "come with" the procedure in the sense that, when the procedure is given a phenomenon to deal with, it either **selects** or **creates** (or some combination of the two) a subsidiary procedure to deal with it. And in many cases the subsidiary procedures come with their own subsidiary procedures -- this hierarchy may go several levels down, thus providing a multilevel control network.

So, in a slightly less coarse approximation to this dual network structure, let us say that each **hard core** process y_i generates a collection of subprocesses y_{i1}, y_{i2},.... For each i, let us consider the explanations of a fixed phenomenon z generated by one these subprocesses -- the collection $\{y_{ij}*z, j=1,2,3,...\}$. The first innovation ratio $[d(S(y_{ij}*z),S(y'*z))/d(\#y_{ij},y')]$ measures how much changing the subprocess y_{ij} changes the explanation which the subprocess generates. This is a measure of the ability of **y_i** to come up with fundamentally new explanations by exploiting structural instability. It is thus a measure of the **creativity** or **flexibility** of the hard core of the belief system.

Of course, if a belief system has many levels, the first innovation ratio has the same meaning on each level: it measures the flexibility of the processes on that level of the belief system. But considering creativity on many different levels has an interesting consequence. It leads one to ask of a given process, not only whether it is creative in generating subprocesses, but whether it generates subprocesses that are themselves creative. I suggest that successful belief systems have this property. Their component processes tend to be **creative in generating creative subprocesses**.

This, I suggest, is one of the fundamental roles of belief systems in the dual network. Belief systems are structured transformation systems that serve to **systematically create new pattern via multilevel structural instability**.

Earlier I explained how the linguistic nature of belief systems helps make it possible for them to generate complex explanations for novel situations. Linguistic structure allows one to determine the **form** of a combination of basic building blocks, based on the **meaning** which one wants that combination to have. Now I have also explained why linguistic structure is not enough: in order to be truly successful in the unpredictable world, a belief system must be systematically creative in its use of its linguistic structure.

10.2 BELIEF AS IMMUNITY

A belief system is a complex self-organizing system of processes. In this section I will introduce a crucial analogy between belief systems and a complex self-organizing **physical** system: the immune system. If this analogy has any meat to it whatsoever, it is a strong new piece of evidence in favor of the existence of a nontrivial complex systems science.

Recall that the multilevel control network is roughly "pyramidal," in the sense that each processor is connected to more processes **below** it in the hierarchy than **above** it in the hierarchy. So, in order to achieve reasonably rapid mental action, not **every** input that comes into the lower levels can be passed along to the higher levels. Only the most important things should be passed further up.

For example, when a complex action -- say, reading -- is being **learned**, it engages fairly high-level processes: consciousness, systematic deductive reasoning, analogical memory search, and so on. But eventually, once one has had a certain amount of practice, reading becomes "automatic" -- lower-level processes are programmed to do the job. Artful conjecture and sophisticated deduction are no longer required in order to decode the meaning of a sentence.

An **active belief** about an entity s may be defined as a process in the multilevel control hierarchy that:

1) includes a belief about s, and

2) when it gets s as input, deals with s without either

a) doing recursive virtually-serial computation regarding s, or b) passing s up to a higher level.

In other words, an active belief about s is a process containing a belief about s that **tells the mind what to do about s** in a reasonably expeditious way: it doesn't pass the buck to one of its "bosses" on a higher level, nor does it resort to slow, ineffective serial computation.

This definition presupposes that individual "processes" in the dual network don't take a terribly long time to run -- a noncontroversial assumption if, as in Edelman's framework, mental processes are associated with clusters of cooperating neurons. **Iterating** single processes or sequences of processes may be arbitrarily time-consuming, but that's a different matter.

All this motivates the following suggestive analogy: **belief systems are to the mind as immune systems are to the body**. This metaphor, I suggest, holds up fairly well not only on the level of purpose, but on the level of internal dynamics as well.

The central purpose of the immune system is to protect the body against foreign invaders (antigens), by first **identifying** them and then **destroying** them. The purpose of a belief system, on the other hand, is to protect the **upper levels** and **virtual serial capacity** of the mind against problems, questions, inputs -- to keep as many situations as possible out of reach of the upper levels and away from virtual serial processing, by dealing with them according to lower-level active beliefs.

10.2.1. Immunodynamics

Let us briefly review the principles of immunodynamics. The easy part of the immune system's task is the destruction of the antigen: this is done by big, dangerous cells well suited for their purpose. The trickier duties fall to smaller antibody cells: determining what should be destroyed, and grabbing onto the offending entities until the big guns can come in and destroy them. One way the immune system deals with this problem is to keep a large reserve of different antibody classes in store. Each antibody class matches (identifies) only a narrow class of antigens, but by maintaining a huge number of different classes the system can recognize a wide variety of antigens.

But this strategy is not always sufficient. When new antigens enter the bloodstream, the immune system not only tries out its repertoire of antibody types, it creates **new** types and tests them against the antigen as well. The more antigen an antibody kills, the more the antibody reproduces -- and reproduction leads to mutation, so that newly created antibody types are likely to cluster around those old antibody types that have been the most successful.

Burnet's (1976) theory of clonal selection likens the immune system to a population of asexually reproducing organisms evolving by natural selection. The fittest antibodies reproduce more, where "fitness" is defined in terms of match which antigen. But Jerne (1973) and others showed that this process of natural selection is actually part of a web of intricate self-organization. **Each antibody is another antibody's antigen** (or at least another "potential antibody"'s antigen), so that antibodies are not only attacking foreign bodies, they are attacking one another.

This process is kept in check by the "threshold logic" of immune response: even if antibody type Ab_1 matches antibody type Ab_2, it will not attack Ab_2 unless the population of Ab_2 passes a certain critical level. When the population **does** pass this level, though, Ab_1 conducts an all-out battle on Ab_2. So, suppose an antigen which Ab_2 recognizes comes onto the scene. Then Ab_2 will multiply, due to its success at killing antigen. Its numbers will cross the critical level, and Ab_1

will be activated. Ab_1 will multiply, due to its success at killing Ab_2 -- and then anything which matches Ab_1 will be activated.

The process may go in a circle -- for instance, if Ab_0 matches Ab_1, whereas Ab_2 matches Ab_0. Then one mightpotentially have a "positive feedback" situation, where the three classes mutually stimulate one another. In this situation a number of different things can happen: any one of the classes can be wiped out, or the three can settle down to a sub-threshold state.

This threshold logic suggests that, in the absence of external stimuli, the immune system might rest in total equilibrium, nothing attacking anything else. However, the computer simulations of Alan Perelson and his colleagues at Los Alamos (Perelson 1989, 1990; deBoer and Perelson, 1990) suggest that in fact this equilibrium is only partial -- that in normal conditions there is a large "frozen component" of temporarily inactive antibody classes, surrounded by a fluctuating sea of interattacking antibody classes.

Finally, it is worth briefly remarking on the relation between **network dynamics** and **immune memory**. The immune system has a very long memory -- that is why, ten years after getting a measles vaccine, one still won't get measles. This impressive memory is carried out partly by long-lived "memory B-cells" and partly by **internal images**. The latter process is what interests us here. Suppose one introduces Ag = 1,2,3,4,5 into the bloodstream, thus provoking proliferation of

Ab_1 = -1,-2,-3,-4,-5. Then, after Ag is wiped out, a lot of Ab_1 will still remain. The inherent learning power of the immune system may then result in the creation and proliferation of Ab_2 = 1,2,3,4,5. For instance, suppose that in the past there was a fairly large population of Ab_3 = 1,1,1,4,5. Then many of these

Ab_3 may mutate into Ab_2. Ab_2 is an **internal image** of the antigen. It lacks the destructive power of the antigen, but it has a similar enough shape to take the antigen's place in the ideotypic network.

Putting internal images together with immune networks leads easily to the conclusion that immune systems are structurally associative memories. For, suppose the antibody class Ab_1 is somehow stimulated to proliferate. Then if Ab_2 is approximately complementary to Ab_1, Ab_2 will also be stimulated. And then, if Ab_3 is approximately complementary to Ab_2, Ab_3 will be stimulated -- but Ab_3, being complementary to Ab_2, will then be **similar** to Ab_1. To see the value of this, suppose

Ag = 5,0,0,0,5

Ab_1 = -5,0,0,0,-5

Ab_2 = 5,0,0,-6,0

Ab_3 = 0,-4,0,6,0

Then the sequence of events described above is quite plausible -- even though Ab_3 itself will not be directly stimulated by Ag. The similarity between Ab_3 and Ab_1 refers to a **different subsequence** than the similarity between Ab_1 and Ag. But proliferation of Ag nonetheless leads to proliferation of Ab_3. This is the essence of analogical reasoning, of structurally associative memory. The immune system is following a chain of association not unlike the chains of free association that occur upon the analyst's couch. Here I have given a chain of length 3, but in theory these chains may be arbitrarily long. The computer simulations of Perelson and de Boer, and those of John Stewart and Francisco Varela (personal communication), suggest that the immune systems contains chains that are quite long indeed.

One worthwhile question is: what **good** does this structurally associative capacity do for the immune system? A possible answer is given by the speculations of John Stewart and his colleagues at the Institute Pasteur (Stewart, 1992), to the effect that the immune system may serve as a general communication line between different body systems. I have mentioned the discovery of **chains** which, structurally, are analogous to chains of free association. Stewart's conjecture is that these chains serve the as communication links: one end of the chain connects to, say, a neurotransmitter, and the other end to a certain messenger from the endocrine system.

10.2.2. Belief Dynamics

So, what does all this have to do with belief systems? The answer to this question comes in several parts.

First of all, several researchers have argued that **mental processes**, just like antibodies, reproduce differentially based on fitness. As discussed above, Gerald Edelman's version of this idea is particularly attractive: he hypothesizes that **types of neuronal clusters** survive differentially based on fitness.

Suppose one defines the **fitness** of a process P as the size of

$Em(P,N_1,...,N_k) - Em(N_1,...,N_k)$, where the N_i are the "neighbors" of P in the dual network. And recall that the structurally associative memory is **dynamic** -- it iscontinually moving processes around, trying to find the "optimal" place for each one. From these two points it follows that the probability of a process **not being moved** by the structurally associative memory is roughly proportional to its fitness. For when something is in its proper place in the structurally associative memory, its emergence with its neighbors is generally high.

This shows that, for mental processes, **survival** is in a sense proportional to fitness. In *The Evolving Mind* it is further hypothesized that fitness in the **multilevel control network** corresponds with survival: that a "supervisory" process has some power to reprogram its "subsidiary" processes, and that a subsidiary process may even have some small power to encourage change in its supervisor. Furthermore, it is suggested that successful mental processes can be **replicated**. The brain appears to have the ability to move complex procedures from one location to another (Blakeslee, 1991), so that even if one crudely associates ideas with **regions** of the brain this is a biologically plausible hypothesis.

So, in some form, mental processes do obey "survival of the fittest." This is one similarity between immune systems and belief systems.

Another parallel is the existence of an intricately structured **network**. Just as each antibody is some other antibody's antigen, each active belief is some other active belief's problem. Each active belief is continually putting questions to other mental processes -- looking a) to those on the level above it for guidance, b) to those on its own level as part of structurally associative memory search, and c) to those on lower levels for assistance with details. Any one of these questions has the potential of requiring high-level intervention. Each active belief is continually responding to "questions" posed by other active beliefs, thus creating a network of cybernetic activity.

Recall that, in our metaphor, the analogy to the "level" of antigen or antibody population is, roughly, "level" in the multilevel control network (or use of virtual serial computation). So the analogue of threshold logic is that each active belief responds to a question only once that question has reached **its** level, or a level not too far below.

As in the Ab_1, Ab_2, Ab_3 cycle discussed above, beliefs can stimulate one another **circularly**. One can have, say, two active beliefs B_1 and B_2, which mutually support one another. An example of this was given alittle earlier, in the context of Jane's paranoid belief system: "conspiracy caused leg pain" and "conspiracy caused stomach pain."

When two beliefs support one another, both are continually **active** -- each one is being used to support something. Thus, according to the "survival of the fittest" idea, each one will be replicated or at least reinforced, and perhaps passed up to a higher level. This phenomenon, which might be called **internal conspiracy**, is is a consequence of what in Chapter Eight was called **structural conspiracy**. Every attractor of the cognitive equation displays internal conspiracy. But the converse is not true; internal conspiracy does not imply structural conspiracy.

Prominence in the dual network increases with intensity as a pattern (determined by the structurally associative memory), **and** with importance for achieving current goals (determined by the multilevel control network). Internal conspiracy is when prominence is achieved through **illusion** -- through the conspiratorially-generated mirage of intensity and importance.

10.2.3. Chaos in Belief Systems and Immune Systems

Rob deBoer and Alan Perelson (1992) have shown mathematically that, even in an immune system consisting of two antibody types, chaos is possible. And experiments at the Institute Pasteur in Paris (Stewart, 1992) indicate the presence of chaotic fluctuations in the levels of certain antibody types in mice. These chaotic fluctuations are proof of an active immune network -- proof that the theoretical possibility of an interconnected immune system is physically realized.

Suppose that some fixed fraction of antibody types participates in the richly interconnected network. Then these chaotic fluctuations ensure that, at any given time, a "pseudorandom" sample of this fraction of antibody types is active. Chaotic dynamics **accentuates** the Darwinian process of mutation, reproduction and selection, in the sense that it causes certain antibody types to "pseudorandomly" reproduce far more than would be necessary to deal with external antigenic stimulation. Then these excessively proliferating antibody types may mutate, and possibly connect with other antibody types, forming new chains.

Of course, chaos in the narrow mathematical sense is not necessary for producing "pseudorandom" fluctuations -- complex periodic behavior would do as well, or aperiodic behavior which depends polynomially but not exponentially on initial conditions. But since we know mathematically that immune chaos is possible, and we have observed experimentally what looks like chaos, calling these fluctuations "chaos" is not exactly a leap of faith. Indeed, the very **possibility** of a role for immunological chaos is pregnant with psychological suggestions. What about chaos in the **human memory network**?

Chaos in the immune network may, for example, be caused by two antibody types that partially match each other. The two continually battle it out, neither one truly prevailing; the concentration of each one rising and falling in an apparently random way. Does this process not occur in the psyche as well? Competing ideas, struggling against each other, neither one ever gaining ascendancy?

To make the most of this idea, one must recall the basics of the dual network model. Specifically, consider the interactions between a set (say, a pair) of processes which reside on one of the lower levels of the perceptual-motor hierarchy. These processes themselves will not generally receive much attention from processes on higher levels -- this is implicit in the logic of multilevel control. But, by interacting with one another in a **chaotic** way, the prominences of these processes may on some occasions **pseudorandomly** become very large. Thus one has a mechanism by which pseudorandom samples of lower-level processes may put themselves forth for the attention of higher-level processes. And this mechanism is enforced, not by some overarching global program, but by natural self-organizing dynamics.

This idea obviously needs to be refined. But even in this rough form, it has important implications for the psychology of attention. If one views consciousness as a process residing on the intermediate levels of the perceptual-motor hierarchy, then in chaos one has a potential mechanism for pseudorandom changes in the focus of attention. This ties in closely with the speculation of Terry Marks (1992) that psychological chaos is the root of much **impulsive** behavior.

10.3. PSYCHIC ANTIMAGICIANS

I have been talking about beliefs "attacking" one another. By this I have meant something rather indirect: one belief attacks another by **giving the impression of being more efficient than it**, and thus depriving it of the opportunity to be selected by higher-level processes. One way to think about this process is in terms of the "antimagician" systems of Chapter Seven.

Also, I have said that belief systems may be viewed as component-systems, in which beliefs act on other beliefs to produce new beliefs. But I have not yet remarked that the process of beliefs destroying other beliefs may be conceived in the same way. When beliefs B and C are competing for the attention of the same higher-level process, then each time one "unit" of B is produced it may be said that one "unit" of anti-C is produced. In formal terms, this might be guaranteed by requiring that whenever f(g) = B, f(g,B) = C^. According to this rule, unless f and g vanish immediately after producing B, they will always produce one unit of anti-C for each unit of B.

The relationship between C and C^ strengthens the immunological metaphor, for as I have shown each antibody class has an exactly complement. In the immune system, an antibody class and its complement may coexist, so long as neither one is stimulated to proliferate above the threshold level. If one of the two complements exceeds the threshold level, however, then the other one automatically does also. And the result of **this** is unpredictable -- perhaps periodic variation, perhaps disaster for one of the classes, or perhaps total chaos.

Similarly, B and C may happily coexist in **different parts** of the hierarchical network of mind. The parts of the mind which know about B may not know about C, and vice versa. But then, if C comes to the attention of a **higher-level process**, news about C is spread around. The processes supervising B may consider giving C a chance instead. The result may be all-out war. The analogue here is not precise, since there is no clear "threshold" in psychodynamics. However, there are different levels of abstraction -- perhaps in some cases the jump from one of these levels to the next may serve as an isomorph of the immunological threshold.

Anyhow, the immunological metaphor aside, it is clear that the concept of an "antimagician" has some psychological merit. Inherently, the dynamics of belief systems are **productive** and not **destructive**. It is the multilevel dynamics of the **dual network** which providesfor destruction. Space and time constraints dictate that some beliefs will push others out. And this fact may be conveniently modeled by supposing that beliefs which compete for the attention of a supervisory process are involved with creating "anti-magicians" for one another.

Indeed, recalling the idea of "mixed-up computation" mentioned in Chapter Seven, this concept is seen to lead to an interesting view of the productive power of belief systems. Belief systems without antimagicians cannot compute universally unless their component beliefs are specifically configured to do so. But belief systems **with** antimagicians can compute universally even if the beliefs involved are very simple and have nothing to do with computation. It appears that, in this case, the discipline imposed by efficiency has a positive effect. It grants belief systems the automatic power of negation, and hence it opens up to them an easy path toward the production of **arbitrary** forms.

For instance, consider the following simple collection of beliefs:

A: I believe it is not a duck

B: I believe it is a duck

C: I believe it walks like a duck

D: I believe it quacks like a duck

E: I believe it is a goose

The mind may well contain the following "belief generation equations":

$F(F) = F$

$F(C,D) = B$

$B(B) = B$

$G(G) = G$

$G(E) = B\hat{}$

The self-perpetuating process F encodes the rule "If it walks like a duck, and quacks like a duck, it should probably be classified as a duck." The self-perpetuating process B encodes the information that "it" is a duck, and that if it was classified as a duck yesterday, then barring further information it should still be a duck today. And, finally, the self-perpetuating process G says that, if in fact it should be found out that "it" is a goose, one should **not** classify it as a duck, irrespective of the fact that it walks like a duck andquacks like a duck (maybe it was a goose raised among ducks!).

The entity F performs conjunction; the entity G performs negation. Despite the whimsical wording of our example, the general message should be clear. The same type of arrangement can model **any** system in which certain standard observations lead one to some "default" classification, but more specialized observations have the potential to overrule the default classification. The universal computation ability of antimagician systems may be rephrased in the following form: belief systems containing conjunctive default categorization, and having the potential to override default categorizations, are capable of computing anything whatsoever. **Belief systems themselves may in their natural course of operation perform much of the computation required for mental process.**

10.4. GOD, THE BIBLE AND CIRCULARITY

Now, in this final section, I will turn once again to the analysis of concrete belief systems. In Chapter Eight I considered one example of intense internal conspiracy -- Jane's paranoid belief system. But this may have been slightly misleading, since Jane's belief system was in fact an explicit **conspiracy theory**. In this section I will consider a case of internal and structural conspiracy which has nothing to do with conspiracies in the external world: the belief system of **Christianity**.

Christianity is a highly complex belief system, and I will not attempt to dissect it in detail. Instead I will focus on some very simple belief dynamics, centering around the following commonplace example of circular thought:

God exists because the Bible says so, and what the Bible says is true because it is the Revealed Word of God.

This "proof" of the existence of God is unlikely to convince the nonbeliever. But I was astonished, upon reading through a back issue of *Informal Logic*, to find an article attempting its defense.

The author of the article, Gary Colwell (1989), reorganizes the argument as follows:

(1) The Bible is the Revealed Word of God

(2) The Bible says that God exists

(3) God exists

His most interesting thesis is that, in certain cases, (1) is more **plausible** than (3). If one accepts this, it follows that demonstrating (3) from (1) is not at all absurd. Therefore, Colwell reasons, in practice the argument is not circular at all.

I do not agree with Colwell's argument; in fact I find it mildly ridiculous. But by pursuing his train of thought to its logical conclusion, one may arrive at some interesting insights into the creativity, utility and self-perpetuating nature of the Christian belief system.

10.4.1. The Bible and Belief

Let us review Colwell's case for the greater plausibility of (1), and pursue it a little further. I contend that, rather than removing the circularity of the argument, what Colwell has actually done is to **identify** part of the **mechanism** by which the circularity of the argument works in practice.

Colwell's argument for the greater plausibility of (1) is as follows:

It is not uncommon to hear of believers who relate their experience of having encountered God through the reading of the Bible. Prior to their divine encounter they often do not hold the proposition "God exists" as being true with anything approaching a probability of one half. Indeed, for some the prior probability of its being true would be equivalent to, or marginally greater than, zero. Then ... they begin to read the Bible. There in the reading, they say, they experience God speaking to them. It is not as though they read the words and then infer that God exists, though such an inference may be drawn subsequently. Rather, they claim that the significance of the words, the personal relevance of the words, and the divine source of the words are all experienced concomitantly. In reading the words they have the complex experience of being spoken to by God. The experienced presence of God is not divorced from their reading of the words....

Given that this experience of encountering God in the reading of the Bible is a grounding experience for the believer, from which he may only later intellectually abstract that one element that he refers to by saying that God exists, proposition (1) for such a believer may actually be more plausible than proposition (3).

Putting aside the question of how common this type of religious experience is, what is one to make of this argument?

I think that Colwell is absolutely right. It probably **is** possible for a person to find (1) more plausible than (3). For a person who has had the appropriate religious experience, the argument may be quite sensible and noncircular.

After all, when told that a young man has long hair, and asked to rate which of the following two sentences is more likely, what will most people say?

A: The young man is a bank teller

B: The young man is a bank teller and smokes marijuana

The majority of people will choose B. Numerous psychological experiments in different contexts show as much (for a review, see Holland et al (1975)). But of course, whenever B is true, A is also true, so there is no way B is more likely than A. The point is, intuitive judgements of probability or plausibility do not always obey the basic rules of Boolean logic. Even though (1) implies (3) (and in fact **significantly** implies (3) in the sense of Chapter Four), a person may believe that (1) is more likely than (3). Why not -- it is known that, even though B implies A, a person may believe B to be more likely than A.

What this means, I believe, is that the human mind is two-faced about its use of the word "and." If asked, people will generally make a common-language statement equivalent to "'and' means Boolean conjunction." But when it comes down to making real-life judgements, the human mind often interprets "and" in a non-Boolean way: it thinks as if "A and B" could be true even though A were false. Thus, God exists and the Bible is the Revealed Word of God" is treated as if it could be true even though "God exists" were false. In judging the plausibility or likelihood of "A and B," the mind sometimes uses a roughly **additive** procedure, combiningthe likelihood of A with the likelihood of B, when on careful conscious reflection a **multiplicative** procedure would make more sense.

But it seems to me that Colwell's argument contains the seeds of its own destruction. I grant him that in certain cases the inference from (1) to (3) may be reasonable -- i.e., given the a priori judgement of greater plausibility for (1). But nonetheless, the argument is **still** fundamentally circular. And I suspect that its circularity plays a role in the maintenance of religious belief systems.

I have known more than one religious individual who, when experiencing temporary and partial doubt of the existence of God, consulted the Bible for reassurance -- in search of the kind of experience described by Colwell, or some less vivid relative of this experience. But on the other hand, the same people, when they came across passages in the Bible that made little or no intuitive sense to them, reasoned that **this passage must be true because the Bible is the**

Revealed Word of God. Certain passages in the Bible are used to bolster belief in God's existence. But belief in the validity of the Bible -- when shaken by **other** passages from the Bible -- is bolstered by belief in God's existence. The two beliefs (1) and (3) support each other circularly. Considered in appropriate context, they may be seen to **produce one another**.

This psychological pattern may lead to several different results. In some cases the intuitive unacceptability of certain aspects of the Bible may serve to **weaken** belief in God. That is, one might well reason:

(1) The Bible is the Revealed Word of God

(2') The Bible is, in parts, unreasonable or incorrect

(3') Thus God is capable of being unreasonable or incorrect

And (3'), of course, violates the traditional Christian conception of God. This is one possible path to the loss of religious faith.

On the other hand, one might also reason

(1") God exists and is infallible

(2") The Bible is, in parts, unreasonable or incorrect

(3") The Bible is **not** the Revealed Word of God

This is also not an uncommon line of argument: many religious individuals accept that the Bible is an imperfect historical record, combining the Word of God with other features of human origin. For instance, not all Christians accept the Bible's estimate of the earth's age at 6000 years; and most Christians now accept the heliocentric theory of the solar system.

Finally, more interestingly, there is also the possibility that -- given appropriate real-world circumstances -- these two circularly supported beliefs might lead to **increased** belief in God. We have agreed that it is possible to believe (1) more strongly than (3). So, for sake of argument, suppose that after a particularly powerful experience with the Bible, one assigns likelihood .5 to (1), and likelihood .1 to (3). Then, what will one think **after** one's experience is done, when one has time to mull it over? Following Colwell's logic, at this point one will likely reason that, if (1) has likelihood .5, then the likelihood of (3) cannot be as low as .1. Perhaps one will up one's estimate of the likelihood of (3) to .5 (the lowest value which it can assume and still be consistent with Boolean logic). But **then**, now that one believes fairly strongly in the existence of God, one will be much more likely to attend church, to speak with other religious people -- in short, to do things that will encourage one to have **yet more intense** experiences with the Bible. So then, given this encouragement, one may have a stronger experience with the Bible that causes one to raise one's belief in (1) to .8. And after pondering this experience over, one may raise one's belief in (3) to .8 -- and so forth. The circularity of support may, in conjuction with certain properties of the real world in which the believer lives, cause an actual increase in belief in both (1) and (3).

So, whereas Colwell expresses "curiosity about the prominence that the putatively circular Biblical argument has received," I see no reason for curiosity in this regard. The Biblical argument in question really **is** circular, and it really **does** play a role in the maintenance of religious belief systems. The religious experience which he describes is indeed real, at least in a psychological sense -- but it does not detract from the circularity of the argument. Rather, it is connected with this circularity in a complex and interesting way.

10.4.2. Christianity as a Belief System

Let us rephrase this discussion in terms of pattern. "God exists" is a certain way of explaining events in the world. It explains some events -- say, a child being hit by a car -- very poorly. But it explains other events fairly well. To give an extreme example, several college students have reported to me that they do better on their **mathematics** tests if they pray beforehand. This phenomenon is explained rather nicely by the belief that God exists and intervenes to help them. My own preferred explanation -- the placebo effect -- is much less simple and direct.

Two related examples are the religious ecstasy some people experience in church, and the experience of "talking to God" -- either directly or, as discussed above, through the Bible. These subjective psychological phenomena are well explained by the hypothesis that God exists. Alternate explanations exist, but they are more complex; and the religious belief system is rather vigilant in sending out "antimagicians" against these alternatives.

Believing that "the Bible is the Revealed Truth of God" explains a few other things, in addition to those phenomena explained by "God exists." And, more importantly, it gives the believer a set of rules by which to organize her life: the Ten Commandments, and much much more. These rules promote **happiness**, in the sense defined above: they provide **order** where otherwise there might be only uncertainty and chaos. They actually **create** pattern and structure. They are a very effective "psychological immune system" -- protecting valuable high-level processes from dealing with all sorts of difficult questions about the nature of life, morality and reality.

So, one has an excellent example of internal conspiracy: belief in the Bible supports belief in God, and vice versa. And in very many cases this internal conspiracy is **also** a structural conspiracy: the two beliefs create one another. Belief in the Bible gives rise to belief in God, in an obvious way; and belief in the Christian God, coupled with a certain faith in the trappings of contemporary religion, gives rise to belief in the Bible. It is certainly possible to believe in the Christian God while doubting the veracity of the Bible; but in nearly all cases belief in the Christian God leads at least to belief in **large portions** of the Bible.

This is a **useful** belief system, in that it really does deal with a lot of issues at low levels, savinghigher levels the trouble. It is psychologically very handy. For example, it mitigates against the mind becoming troubled with metaphysical questions such as the "meaning of life."

And it does wonders to prevent preoccupation with the fear of death. It serves its immunological function well.

Next, as anyone who has perused religious literature must be well aware, the Christian belief system is systematically creative in explaining away phenomena that would appear to contradict Biblical dogma. It is precisely becuase of this that arguing evolution or ethics with an intelligent Christian fundamentalist can be unsettling. Every argument receives a response which, although clever and appropriate in its own context, is nonetheless strange and unexpected.

So, to a certain extent, the Christian belief system meets **both** the criteria for survival laid out at the beginning of the chapter. It is an attractor for the cognitive equation, a structural conspiracy, and it is creatively productive in the service of the dual network.

However, the Christian belief system clearly does have its shortcomings. It entails a certain amount of awkward dissociation. For instance, the Bible implies that the Earth is only a few thousand years old, thus contradicting the well-established theory of evolution by natural selection. In order to maintain the Christian belief system, the mind must erect a "wall" between its religious belief in the Bible and its everyday belief in scientific ideas. This is precisely the sort of dissociation that leads to **ineffective** thinking: dissociation that serves to protect a belief from interaction with that which would necessarily destroy it.

The prominence of this sort of dissociation, however, depends on the particular mind involved. Some people manage to balance a Christian belief system with a scientific world-view in an amazingly deft way. This is systematic creativity at work! For others, however, Christianity becomes stale and unproductive, separate from the flow of daily life and thought. The value of a belief system cannot be understood outside of the context of a specific believing mind. Just as a cactus is fit in the desert but unfit in the jungle, Christianity may be rational or irrational, depending on the psychic environment which surrounds it.

Chapter Eleven

MIND AND REALITY

Now, finally, with the cognitive equation and the theory of belief systems under our belt, we are ready to return to the "crucial connections" of Chapter Six -- to the intimate relationship between language, thought, reality, self and consciousness. In this chapter I will present several different views of the relationship between **psychology** and the **external world**.

In Section 1, using the ideas of the past two chapters, I will present the radical but necessary idea that **self and reality are belief systems**. Then, in Section 2, I will place this concept in the context of the theory of hypersets and situation semantics, giving for the first time a formal model of the universe in which **mind and reality reciprocally contain one another**. This "universal network" model extends the concept of the dual network, and explains how the cognitive equation might actually be considered as a **universal equation**.

Finally, in Sections 3-5, I will put forth a few speculative suggestions regarding how one might reconcile this idea with our contemporary understanding of the **physical** world. I will confront the well-known paradoxes of quantum mechanics, and argue that the resolution of these paradoxes may lie in the idea that **the world is made of pattern**. If this idea is correct, it will provide a basis for integrating the idea that reality is a belief system with modern physical science.

11.1. LANGUAGE, BELIEF AND REALITY

Nietzsche and Whorf, despite their fundamental theoretical differences, shared the following radical view: **external and internal reality are belief systems**. Further, they both maintained that **one of the main roles of consciousness and language is to maintain these belief systems**. Beings without consciousness and language, according to this perspective, do not perceive a split between external and inner reality.

Let us explore this proposition in detail. I have said that a language consists of a syntactic system appropriately coordinated with a semantic system. But this characterization says nothing about the possibility that the semantic system of a given written/spoken language **may also serve other purposes**. Perhaps this semantic system is also connected with various belief systems.

A belief system is itself a special kind of linguistic system. Each belief has a certain meaning, and the meanings, in order to be psychologically useful, must change roughly continuously with the syntactic construction of the beliefs.

On this rarefied level, the Nietszchean/Whorfian insight is simply that **different abstract "languages" may intersect one another semantically, while being quite different syntactically**. One of the languages is ordinary spoken language, and the others are **belief systems**, including the one which we call by the name "external reality."

In terms of efficiency, the sharing of a common semantic system by two different syntactic systems makes a lot of sense. Semantic systems are space-intensive -- they require the storage of a vast number of patterns/processes and the connections between them. Syntactic systems, on the other hand, are more time-intensive: they, like the slightly more general transformation systems discussed in Chapter Two, require the repetitive application of simple rules. Having two syntactic systems share the same semantic system conserves space, allowing the mind to pack a greater number of linguistic systems into the same space.

11.1.1. Reality as a Belief System

The belief system which we call external reality is a collection of processes for constructing three-dimensional space, linear time and coherent objects out of noise- and chaos-infused sense-data. Neurobiologists are just beginning to probe the most primitive levels of this belief system; the more sophisticated levels are completely out of reach. If the mind had to applyconscious and/or deductive reasoning to every batch of sense-data it received, it would be paralyzed. How long would it take to thoughtfully, logically determine the best interpretation of a given series of photons on the retina? For efficiency reasons, the mind instead applies certain **common sense beliefs** about the way the world is structured, and automatically or semi-automatically processes sense-data in terms of these beliefs.

The classic optical illusion experiments show that these common sense beliefs can be misleading. For instance, in the Ames experiment one looks through a peephole into a room with oddly angled walls, and one misjudges the relative positions of objects. But this is because one is applying irrelevant beliefs. Given enough exposure, the "external reality" belief system can use continuous compositionality (analogical structure) to adjust itself to minor changes of this sort. It can create new high-level beliefs to match the situation, by piecing together the same low-level beliefs that are pieced together to judge the relative positions of objects in an ordinary room.

11.1.2. Self as a Belief System

At first glance, "self" might seem to be a far simpler belief system than "reality." After all, what **beliefs** are involved in selfhood, beyond the simple faith that "I exist, and I act"? But a more careful investigation reveals that the sense of self is every bit as intricate as the sense of external reality. One's inner world is subtly guided by one's **body-concept**.

This point was emphasized repeatedly by Hubert Dreyfus in his *What Computers Can't Do* (1978). This book, which purports to be a disproof of the possibility of artificial intelligence, fails at its intended goal. But it is devastatingly effective as a diatribe against computer programs which attempt to simulate self in a disembodied way. Human intelligence, Dreyfus points out, is indivisible from the sense which we humans have of presence in a **body**. When we reason, we relate different ideas in a way that draws analogically on 1) the felt interrelations of parts of our body, and 2) the relation of our **body** with various external objects.

For example, the "detached" feeling of logical reasoning is not unrelated to the feeling of a separation between **self** and **world**. By learning to distinguish oneself from the external world, one learns moregenerally how to divide a continuum of patterns into **actor** and **acted-upon**. Thus I would predict that those who feel themselves more "at one" with the world will also be less likely to enjoy reasoning in a detached, "objective" way. This prediction is validated by the work on "boundaries" to be discussed a little later.

To see more vividly the reality of body-self interdependence, consider the phenomenon of the "phantom limb," discussed for example in Israel Rosenfield's recent book *Strange, Familiar and Forgotten* (1992). When a person loses her arm, she may instinctively **feel** the arm to be there for months or even years afterwards. This means that her sense of the existence of her arm is not tied to the physical sensations being sent from the arm, but rather persists "in itself."

From the point of view of classical psychological theories or modern cognitive science, this is rather difficult to explain; it requires complex theoretical contrivances. But from the **cognitive equation** perspective, it is virtually obvious. Since the self is a successful belief system, it must be an attractor for the cognitive equation. But if it is an attractor for the cognitive equation, then each one of its component beliefs must be **producible** by the others. The belief in the existence of one's right arm can be **produced** by the other beliefs in the self belief-system.

To put it less abstractly, we not only have processes for receiving data from our arms, we have processes for analyzing and transforming this data, and requesting more data. The theory of belief systems suggests that **this network of processes** is capable of producing the belief that the arm exists. And this is exactly what is observed in the phenomenon of the "phantom limb."

11.1.3. Intersections

Perhaps the most impressive example of intersection between the semantic system of spoken language and the semantic system of self/reality is the **imaginary subject**, discussed earlier in the context of Nietzsche's thought. Who can dispute the fact that, when we understand the world or self, we assume objects where there may not be any? The interpolation of imaginary subjects is a universal method for finding meaning. It ties **linguistic** constructions such as "I" and "flash" together with **biological** constructions like the phantom limb, and thefilled-in blind spot directly in front of every human being's nose.

But other examples are not lacking. For instance, in one of his most interesting papers, Whorf compares the Indo-European and Hopi concepts of time. The Hopi language, he claims, groups **future** and **imaginary** into one category, and **past** and **present** into another category. Correspondingly, he claims, their subjective "external worlds" are structured differently. Whereas we perceive a rift between the present and the past, they feel none. And whereas we tend to see the future as something definite, largely pre-determined, they tend to perceive it as nebulous and conjectural.

Whorf tends to imply that linguistic structure **causes** the structure of reality. But I don't see the point of introducing a Newtonian concept of causality. If one has two syntactic systems using the same semantic system, then **both** of them will influence the semantic system every time they access it. Each reference to a structurally associative memory has the potential to affect that memory's notion of association -- and thus its fundamental structure. Therefore, two linguistic systems that share the same memory network will influence one another quite directly -- each one will affect the structure of the common memory, which in turn will affect the direction of deduction in both systems.

It is possible that one of the two systems will have a greater effect on the common semantic system. But Whorf gives us no reason to believe that this is the fact of the matter in the case of spoken language and external reality. Evolutionarily and socially, these two systems must have originated together. Developmentally, in the mind of a child, the two arise together. And, finally, in day-to-day thought, the two operate symbiotically. Each time a person speaks, her semantic system is reinforced in ways that follow the demands of language; but each time a person perceives or reasons about reality, her semantic system is reinforced in ways that follow the structure of the belief system that is external reality.

11.1.4. Language, Conspiracy and Reality

If one accepts the idea that spoken language and external reality are interconnected linguistic systems, then one has the question of **why these systems survive**. Recall the idea that belief systems use three differentstrategies to maintain themselves: 1) effectiveness at protecting high-level processes from problems, 2) internal conspiracy. It seems quite plain that external reality excels in not only in the first category, but also in the second -- that **the belief system which we call external reality is a structural conspiracy which relies strongly on internal conspiracy for its survival**.

In other words, I suspect that is is common for belief in one aspect of external reality to reinforce or create belief in another aspect of external reality, and vice versa, even when those aspects of external reality have little or no support outside the belief system of external reality. The Sapir-Whorf hypothesis suggests that language is a key accomplice in this conspiracy.

This is a very deep and very radical hypothesis. And its complementary hypothesis is equally striking: that **the belief system which we call self also has the formal structure of internal conspiracy**. Whorf focused on outer reality more than on inner reality, but Nietzsche understood **both** to be constructs of language and consciousness. As already noted, he saw the "little word I" and the experience of "free will" as the most egregious possible instances of imaginary-subject postulation.

And, taking the whole process one level higher, these two internally conspiratorial belief systems combine to form a larger conspiracy. Belief in the self and free will encourages belief in an external reality. Belief in an external reality encourages belief in self and free will. The concepts "inner world" and "outer world" are each meaningless in isolation; they gain their meaning from one another. And the two systems involve many similar beliefs -- the postulation of imaginary subjects is one example, and the assumption of a linear time axis is another.

To make this a little clearer, consider the case of a person in doubt about the reality of the world around her. Two beliefs may pop into her mind: the belief that the wall in front of her is real, and the belief that the floor below her is real. Internal conspiracy suggests that these two beliefs will **reinforce one another**, increasing one another's strength just like two complementary antibody classes in the immune system.

Next, suppose that our heroine is **also** confused about **her own** reality -- about the effectiveness and substantiality of the mental process called her "self." Suppose, in order to test this hypothesis, she picks up a rock and throws it at the wall. Then two beliefs may occur to her: the belief that "she" is really in control of something, and the belief that the rock is really there. Internal conspiracy suggests that these two beliefs increase one another's strength: the more she believes she is in control, the more she is likely to believe the rock is real; and the more she believes the rock is real, the more she is likely to believe she is in control of something.

Next, **structural** conspiracy suggests that, as well as reinforcing one another, these basic beliefs are able to create one another. For instance, belief in the reality of the wall could be **created** by the belief in the reality of the floor, the ceiling, the lamp hanging on the wall, etc. And it could also be created in a different way, by reference to beliefs from the **self** system: for instance, belief in the controlling nature of the hand that punches the wall, the fingernail that scrapes the wall, or the voice that echoes off the wall.

11.1.5. Godfrey Vesey on Inner and Outer

It is interesting to contrast the Nietszchean/Whorfian view of self and reality with that of the contemporary philosopher Godfrey Vesey. In the Introduction to his insightful book *Inner and Outer* (1991), Vesey writes

> The essays in this collection are on a philosophical myth. I call it 'the myth of inner and outer.' It is behind what Gilbert Ryle calls 'the myth of the ghost in the machine.' But it is also behind what might be called 'the myth of a machine with a ghost in it', or, more generally, 'the myth of the world as external'. In brief, the myth divides what, to the philosophically unindoctrinated (and even to the indoctrinated in their non-philosophical moments) is undivided, into two distinct things -- one inner ('mental') and one outer ('physical').
>
> The myth manifests itself in philosophical theories of voluntary action, perception and communication. In regard to voluntary action, the myth finds expression in the theory that my raising my arm is really two distinct things, one of them inner (my performing a mental act ofwilling, a 'volition') and one of them outer (my arm rising).... In the case of communication, there is what Jonathan Bennett called 'the translation view of language': my saying something involves my translating inner things (ideas or thoughts) into outer things (audible sounds), and my understanding what someone has to say involves my translating outer things (audible sounds) into inner things (ideas or thoughts).

I cannot accept Vesey's classification of the rift between inner and outer as "a philosophical myth," unknown to the "unindoctrinated." Surely the concepts of internal and external reality are more than erroneous theoretical constructs of some philosophers!

Look at Vesey's two examples: the idea that raising one's arm involves both an inner and an outer act, and the idea that language involve translating sound waves into ideas. Both of these examples represent the standard scientific perspective. We actually know which parts of the cerebellum must be activated in order to cause an arm to be lifted up. And we know which parts of the brain are stimulated by audible sounds, and which parts of the brain process those audible sounds that carry recognizable language. These examples are not philosophical myths, they are elementary neuroscience!

And, in addition to being good biology, they are also good common sense. We can have the thought of going to the freezer to get some ice cream, followed by the action of going to the freezer to get some ice cream -- these are two different things, and the first in some sense seems to cause the other. It is by analogy to this sort of situation that we analyze arm-raising in terms of a thought followed by an action. This is similar to (and related to) the postulation of an imaginary subject ... it is the postulation of an at least partially imaginary "cause and effect."

Similarly, when we speak, we often have the experience of **first** consciously formulating a sentence, **then** saying it. Although the process is not always so deliberate, even when it is not, we still tend to make the assumption that **all** speech consists of thought followed by action. This is a commonplace analogy, absolutely natural and inevitable in the functioning of the mental network.

In sum, what Vesey disparages as "a philosophical myth" is in fact absolutely essential both to everydaylife and to biological science. The concepts of inner and outer reality cannot just be dismissed out of hand. I agree with Vesey that they are not "correct" in any absolute sense. But I contend that they are **justified belief systems** in the sense of Chapter Ten, as well as being **internally conspiratorial** belief systems. They are impressively, incredibly dialogical -- the amount of new pattern which they create is far beyond our conscious comprehension.

It would be of great interest to study the structure of these belief systems in detail, with an eye toward understanding their dialogicality, their internal conspiratoriality, and their relationship with the deep structure of language. To a certain extent this quest is self-referential, since our tools for studying things are largely based on the concepts of internal and external reality. But, as I have repeatedly emphasized, self-reference is not necessarily a problem; it can be part of a solution. It is hard to imagine a research programme of greater importance or interest than this one.

11.1.6. Boundaries

All this talk of self and reality may seem overly abstract; disconnected from the actual business of thought. In Chapter Twelve, invoking the notion of **dissociation**, I will present a forceful argument that this is not the case. But dissociation is not the only connection between the self/reality system and ordinary, everyday behavior. In fact, the particular structure of a persons's self/reality system affects everything she thinks and does.

For example, we have seen that all thought, even the most "rational" and "logical," depends essentially on **belief systems**. But how, then, does a child's mind learn to develop belief systems? Nietszche was the first to arrive at the correct answer: **by analogy to, or direct use of, the self/reality belief system**.

For example, Nietzsche observed that the "little word I" is a paradigm case for **reification** in all its aspects. Language developed for speaking about the self involves postulation of an imaginary subject. This language is then used for thinking about all sorts of issues, and thus the tool of imaginary subjects spreads throughout all the belief systems of the mind.

Similarly, I propose, **every major aspect** of more specialized belief systems may be found to have itscounterpart in the one big belief system -- the self/reality system. One example of this involves the notion of **boundaries**, as developed by Ernest Hartmann in his intriguing book *Boundaries in the Mind*. Hartmann has developed a questionnaire designed to distinguish "thick-boundaried" people from "thin-boundaried" people. And through a comprehensive statistical analysis, augmented with numerous personal interviews, he has concluded that these two categories represent genuine personality types. **Thick-boundaried** people tend to place a large "distance" between themselves and the world -- they tend not to remember their dreams, they tend to be rigid in their beliefs and habits, not to be free in expressing their emotions. **Thin-**

boundaried people, on the other hand, seem to live partially in a dream-world, to be permissive and "liberal" in their beliefs, to express their feelings freely, and to be very sensitive to the emotions of others.

These results indicate that the **thickness** of the "boundary" which a person places between their self and their reality is a **quantitative** parameter which carries over into all aspects of life. Once someone's self/reality system erects a thin boundary, then that person's subsequent belief systems will tend to be of the "thin-boundary" type, making few rigid distinctions and permitting entities to blur into their opposites. On the other hand, once someone's self/reality system erects a **thick** boundary, then that person's subsequent belief systems will tend to place things into strict categories, to distinguish X and not-X most strenuously -- to be, in short, "thick-boundaried." This is a very strong piece of evidence that the self/reality belief system is used as a **model** for all subsequent instances of belief-system formation.

This example, as you may have guessed, was not selected arbitrarily. It is of paramount importance in the theory of the dual network. Recall that **consciousness**, in the dual network model, has to do with the iterative strengthening of **barriers** or **boundaries**. But the dual network model certainly does not imply that everyone's barrier-strengthening procedures are equally powerful. These procedures, like all others, **evolve** over the course of a lifetime. For one reason or another, in the course of developing an internal concept of reality, some infants evolve stronger boundary-strengthening processes than others. This psychological trait then carries through to their adult lives, influencing theirpersonalities and their methods of perceiving and categorizing the world.

11.2. COLLECTIVE REALITY

Hyperset theory shows that there is no logical problem with the philosophically attractive idea of reality as a belief system. Mind can belong to reality, while reality belongs to mind. **Mental** patterns in the brain can give rise to processes which themselves make up the brain. The contradiction is only apparent.

But what's the meat of the concept? If reality is a belief system, then what sort of belief system is it? One interesting answer to this question is provided by the situation semanticists, and their intriguing hyperset-based approach to the puzzle of **common knowledge**.

11.2.1. Reality as a Regress

I will begin obliquely, with an example that is not at all philosophically loaded. Consider two people staring into one another's eyes. Intuitively, one might say that each one of the two starers recognizes the following sequence:

I look at her look at me

I look at her look at me look at her

I look at her look at me look at her look at me

I look at her look at me look at her look at me look at...

Or, alternately, one might represent the situation by the circular formula

X = I look at her look at X,

(where I use the expression a = b to denote that a and b are equivalent set-theoretic entities, rather than merely that a is to be assigned the value b.)

What does this have to do with reality? Let us for the moment exclude phenomena such as mysticism, catatonia, extreme retardation, and schizophrenia -- let us consider a society in which everyone recognizes and thinks about essentially the same common externalreality. Then it is only reasonable to conclude that each member of society recognizes the following sequence:

Everyone recognizes the same phenomena

Everyone recognizes that everyone recognizes the same phenomena

Everyone recognizes that everyone recognizes that everyone

recognizes the same phenomena

Everyone recognizes that ...

Given this regress, it is tempting to sum the situation up by the hyperset formula

X = Everyone recognizes that X.

And if "everyone" is too strong, if one wishes to restrict consideration to some group such as the set of sane individuals, one may construct a similar regress leading up to the hyperset formula

X = Every sane person recognizes that X

There is one obvious complaint against this kind of analysis. The infinite regresses I have constructed are logically sensible but psychologically absurd, in the sense that the human mind has only limited recognition abilities. Biologically, at some point the sentences "Everyone recognizes that everyone recognizes that ... everyone recognizes the same reality" will become so long as to exceed the memory capacity of the human brain. So, if the regress is inevitably cut off after some finite point, then what good are the hyperset formulas, which are equivalent only to the **actually infinite** regresses?

However, this objection is far from fatal. To resolve the matter, one need only return to the **definition** of mind as patterns in brain. Suppose someone's brain contains the first twenty iterations of the regress

Everyone recognizes the same phenomena

Everyone recognizes that everyone recognizes the same phenomena...

This collection of twenty patterns is not at all unordered; there are significant patterns in **it**, relating to its obviously repetitive structure. And if hyperset patterns are permitted, then **one** of these patterns is clearly of the form "Take the first 20 iterations of the formula X = everyone recognizes that X, from the initial condition 'everyone recognizes the same phenomena'." This is a nice compact formula which allows one to quickly compute the collection in question.

The limiting hyperset form is part of a **pattern** in the first few iterations of the regress. So, even if the regress of **recognitions** is never explicitly completed, the hyperset formula that encapsulates the infinite regress may still be **part of the mind**. It all depends on whether, in the definition of mind, one interprets the word "pattern" to include "hyperset pattern" instead of just "computable pattern."

11.2.2. Common Knowledge

To make this line of thought a little more concrete, let us next turn to the Conway paradox. In their charming little book *The Liar*, Jon Barwise and John Etchemendy have expressed this conundrum in a particularly simple way:

Suppose you have two poker players, Claire and Max, and each is dealt some cards. Suppose, in particular, that each of them gets an ace. Thus, each of them knows that the following is a fact:

s = either Claire or Max has an ace

Now suppose Dana were to come along and ask them both whether they knew whether the other one had an ace. They would answer "no," of course. And if Dana asked again (and again...), they would still answer "no."

But now suppose Dana said to them, "Look, at least one of you has an ace. Now do you know whether the other has an ace?" They would again both answer "no." But now something happens. Upon hearing Max answer "no" Claire would reason as follows: "If Max does not know I have an ace, having heard that one of us does, then it can only be because he has an ace." Max would reason in the same way. So they both figure out that the other has an ace.

There is a big difference between the first situation Barwise describes, and the second. Intuitively, Dana's statement gave each of them some essential information. But yet, in a sense, Dana told them something that each of them already knew. This is the "paradox."

The intuitive solution of the paradox is that, prior to Dana's statement "at least one of you has an ace," the fact s was known to both of them, but it was not **common knowledge**. The puzzle which this "solution" raises is: what is common knowledge?

One approach is to declare that, by saying that s is common knowledge, one means

Max knows Claire knows s

Claire knows Max knows s

Max knows Claire knows Max knows s

Claire knows Max knows Claire knows Max knows s

...

Clearly, if one gives the name "G" to the group consisting of Claire and Max, then this is substantially the same as

Everyone in group G recognizes s

Everyone in group G recognizes that everyone in group G

recognizes s

Everyone in group G recognizes that everyone in group G

recognizes that everyone in group G recognizes s

...

This regress encapsulates, in a sense, the fact that s is **common knowledge** in the group G. But it is an unwieldy way of representing this fact. Much nicer to say, following Barwise,

X = Everyone in group G knows both X and s

This approach allows us to give a purely "sociological" definition of reality. One may say that a certain thing s is in the **reality** of a group of people if the hyperset

X = Everyone in the group recognizes both X and s

is a pattern in this group over some period of time.

As in the discussion at the end of the previous section, this does **not** imply that the minds involved must be capable of infinitely complex perception and memory. It just means that they carry out a long enough segment of the regress to make the limiting hyperset formula a **pattern** in this segment.

This is a subjective, rather than objective, definition of reality. What it means is that, when we look at a chair, instead of simply seeing a chair, what we see is first of all a regress of the form

Every sane person sees this as a chair

Every sane person knows that every sane person sees this as a chair, and also sees this as a chair

Every sane person knows that every sane person knows that every sane person sees this as a chair, and also sees this as a chair

and secondly a hyperset pattern in this regress:

X = every sane person knows that X, and also sees this as a chair

So far as I know, this is the first ever precise characterization of external reality as a subjective phenomenon. We have not yet arrived at a comprehensive model of mind and reality, but the idea of collective reality is a significant step along the way. It shows how a group of intelligent entities can generate a reality that is fundamentally, emergently **their own**.

11.2.3. The Universal Network

Now, finally, it is time to address the question of the fundamental relationship between mind and reality, from within this hyperset perspective. Let me introduce the word **universe**, to refer to the set containing both mind and physical reality. I suggest that the universe may be understood as a **collection of dual networks, linked at the bottom via certain "connector processes"**.

This is a very natural idea -- after all, the lowest levels of the dual network deal with immediate physical stimuli. So if a collection of dual networks are connected at the bottom, this means that there are processes interrelating the physical stimuli received by one network with the physical stimuli received by the other. These "connector processes" are the only physical reality there is.

And what form do these "connector processes" take? The arguments of the previous section imply that they must take the form

X = Everyone in the group G recognizes both X and s

In other words, these lowest-level connector processes which underly the collection of dual networks, themselves **refer** to the collection of dual networks. They **contain** the collection of dual networks. In this sense, one may say that **reality contains mind, while mind contains reality**.

What is the difference between simply "seeing a chair" and seeing a hyperset pattern of the form

X = every sane person knows that X, and also sees this as a chair ?

The main practical difference is, I suggest, one of **solidity**. Patterns of the form

X = every sane person knows X and s

should logically receive a great deal of protection from reorganization. This gets back to the mind's all-important grouping/scene-making/solidifying processes, which I have said to be intimately involved in consciousness.

11.2.3.1. Reorganization and Reality

But how exactly do these "scene-making" processes work? How do they determine what sort of coherent wholes to form out of the chaotic fragments of perception with which they are presented? They cannot go on internal clues alone -- they must rely largely on memory, on historical information regarding **what is really there**, or in other words **what is common knowledge**. Patterns which are of the "common knowledge" form are much more likely to emerge from their solidifying mechanisms.

Each mind learns to solidify those subnetworks which other minds have solidified. Thus there emerges a common core of "reality," by a kind of feedback relation: the more common knowledge there is, the greater incentive minds will have to reinforce common knowledge, and the more new common knowledge will be created.

So, reality is a self-referential, self-supporting system: each person believes in it because the other ones do. It is a belief system which transcends the boundaries of any one mind, and is supported only by the synergetic actions of many minds. One cannot refute the solipsistic proposition that there is only one mind, and all others are illusions. What is necessary for the maintenance of reality, however, is that these illusory minds must **act** as though they were living in a cooperatively created world. In other words, where reality is concerned, patterns of behavior are more fundamental than so-called "fundamental existence."

11.3. PHYSICAL REALITY AS A MENTAL CONSTRUCTION

Up till now, this book has been concerned with solving puzzles regarding the nature of mind. In this section and the two which follow it, however, I will take a break from proposing new solutions and present instead a **new problem**. This represents a bit of a digression from the main thread of the book, and the impatient reader may wish to skip ahead to Chapter Twelve, dipping back into this material later when time permits.

I do have some ideas regarding the solution of this new problem, but they are frankly speculative and not well developed. My main goal here is to draw attention to the problem itself, for it is a problem that, given its tremendous importance, has not received nearly the attention it deserves.

The problem is as follows: **how are physical structures built from mental structures?** Or, more pointedly: if reality is nothing more than a belief system, then why does this belief system obey beautiful, abstract principles like the Schrodinger equation and Einstein's gravitational field equation?

This question is an **inversion** of the point of view taken by systems theorists like Ilya Prigogine, Erich Jantsch and Hermann Haken (1984). For instance, in his classic treatise *The Self-Organizing Universe*, Eric Jantsch (1980) applies ideas from systems theory to analyze everything from microscopic particles to molecular soups to brains, societies, evolving ecosystems and galaxies. His philosophy is universalist: self-organization, he argues, is a phenomenon underlying **all levels** of structure and dynamics, perhaps **the** vital force of the cosmos. But his actual methodology is to takeideas developed for studying **physical** systems and "extrapolate them upward" toward the mental and social realms.

To a certain extent, it may well be possible to study mind and brain using **physical** ideas. What I am suggesting here, however, is that it may **also** be possible to do exactly the **opposite**: to "build down" from the complex to the simple, and somehow derive the laws of physics from the laws of psychology.

How, then, are physical structures built from mental structures? As already warned, I do not have a solution. It seems to me, however, that the most likely **source** for a solution is **quantum physics**, and more specifically the **quantum theory of measurement**. In the remainder of this chpater, therefore, after a few general philosophical comments, I will briefly review some of the discoveries of this odd branch of physics, and then explore their relationship with the pattern-theoretic psychology that was developed in the body of the book. This discussion will serve to make the basic question more concrete. And it will also lead us to some surprising discoveries -- such as the very close relationship between **quantum measurement**, **pattern philosophy**, and the **cognitive equation**.

This is admittedly a radical programme. But if one is **serious** about the idea that reality is a belief system, then one cannot avoid the question: where do these elegant mathematical properties of reality come from? Today the phrase "Foundations of Physics" refers to a technical subfield of theoretical physics. I venture the prediction that, in a hundred years time, it will refer to a branch of **mathematical psychology.**

So, let's get started. One way to conceptualize the huge gap between physics and psychology is to think about the two most basic aspects of physical reality: the **three dimensions of space** and the **one dimension of time**.

11.3.1. Euclidean Space

The ideas of Chapter Ten imply that three-dimensional Euclidean space is an element of a very very useful **belief system**. In the mental hierarchy of an individual conscious system, it lies well below consciousness, but well above the lowest "raw perception" levels. The postulate of three-dimensional space allows the organization of a vast amount of pattern in a remarkably convenient and productive way.

From this point of view, if the question "why **three** dimensions" has any answer at all, it should have an **system-theoretic** answer. There should be some reason why three dimensional space leads to a more **productive** belief system than two or four dimensional space. Maybe, as has been suggested, this has to do with the fact that three dimensional space is the only Euclidean space in which one can tie knots. Or perhaps it has to do with the fact that in three dimensions, but not in two, any finite graph can be drawn without the crossing of edges.

It may be, of course, that the question has no answer; that the three-dimensionality of our existence is a fluke, with no special meaning. The three-dimensional belief system is ingrained in our minds, brains and culture, but perhaps there are other organisms with mind/brains that naturally organize things in seven dimensions. My suspicion is that there is something special about **three** dimensions -- but this could be a case of bias, of "dimension-centrism"!

So Euclidean space is not fundamental. There is a sense in which **space** is fundamental, but space in this sense means nothing more than **separation**. It means that the mind can consistently perceive two different things **without** perceiving the patterns emergent between them, even though these emergent patterns are present in its memory, and not hard to find. The existence of space, in this sense, says simply that two things will often enter different "parts" of the lower levels of its dual network. It means that the lowest "perceptual" levels of the dual network can receive a variety of different input. This sort of space is essential to the universal dual network. But it comes with no inherent dimensional structure.

11.3.2. Linear Time

Next, what about time? Without pretending to have arrived at a definitive judgement on the matter, let us recall that, according to the cognitive equation, **time** may be equated with the **passage from substance to structure**. In other words, time is the process by which a collection of processes is replaced by those processes which are 1) produced by the actions of elements of A upon elements of A, **and** 2) patterns in the collection of entities formed by actions of elements of A on elements of A.

The cognitive law of motion therefore contains within it the assumption of **one-dimensional time**. A cognitive law of motion for two-dimensional time would involve replacing each collection of processes with **two** mutually noninteracting collections of processes, rather than just one. At the next time step, each of these two would then give rise to two new collections. This is not a completely fanciful idea; one could simulate a two-time-dimensional mind on a computer.

This would of course be **subjective** time, only indirectly connected with clock time. Clock time is a complex construct; it comes about as a consequence of the particular structure of space and it enters into the mind only as an outgrowth of other high-level concepts. We all know from personal experience how uncorrelated subjective time and clock time can be.

11.3.3. Conclusion

Both with space and with time, the gap between physics and psychology is apparent. The dual network model suggests an **abstract** notion of space, and the cognitive equation suggests an **abstract** notion of time. But one cannot equate **psychological** space and time with **physical** space and time. The movement from one to the other is vastly complex and apparently beyond the reach of contemporary science.

11.4. THE QUANTUM MIND

The **psychological** sense of building physical structures from mental structures is easy to see. To understand the **physical** sense of this point of view, we must begin with the commonplace observation that in quantum physics **measuring a phenomenon** is equivalent to **altering that phenomenon**. One cannot determine the position and momentum of an electron simultaneously, not with perfect accuracy -- because the position-determining measurement changes the particle's momentum, and the momentum-determining measurement changes the particle's position. This is the paradox of quantum measurement.

When no one is looking, quantum systems cannot be assumed to possess definite states; they exist in superpositions of physical states. An electron can spin either right or left, but when no one is looking, it isnot spinning either direction -- it is waiting. And the moment someone looks, it somehow decides which way to go.

This technical paradox gives rise to numerous conceptual troubles. For instance, there is the paradox of Schrodinger's Cat. Put a cat in a box together with a gun rigged to fire **only** if a certain electron turns out to be spinning left. Now until you look, the electron is spinning neither right nor left; it is in a state of suspension or superposition. But as soon as as you look, the electron assumes a definite state. So when is the cat shot? At the moment you look? What if your friend walks into the room a minute later -- from **her** view, the definite state should be assumed at the moment **she** looks.

One way of resolving this problem is to simply **define** consciousness as the reduction of quantum superposed states to definite states. This is the course proposed by John von Neumann, and taken up in *SI*. It is an attractive idea, although it does have certain puzzling implications. For instance, suppose, for the sake of argument, that a **mouse** is a conscious system. Then according to the quantum theory, the mouse's thoughts and perceptions play a role in shaping the universe. Einstein could not digest this; he said something like "I cannot believe that, when a mouse looks at the world, it is altered." He rejected Nietzsche's idea that

> A thing would be defined once all creatures had asked "what is that?" and had answered their questions. Supposing one single creature, with its own relationships and perspectives for all things, were missing, then the thing would not be defined.

The theory of the universal network sides with Nietzsche and quantum physics, and against Einstein's idea of an objectively, rationally ordered world.

The real problem with the quantum theory of consciousness, however, is the trouble of connecting it with the biology and psychology of consciousness. It is clear that, if the quantum theory/consciousness connection is to be taken seriously, something further must be done beyond merely **equating** consciousness with reduction. How does reduction from superposition to certainty correspond with the **solidification** that, in the dual network model, is the key function of consciousness? Unfortunately I will not resolve this question here. However, a bit more background regarding quantum theory should make the issue clearer.

11.4.1. Synchronicity

The paradox of quantum measurement ties in with the phenomenon of **nonlocal correlation**, which is surprisingly closely related to Carl Jung's notion of "synchronicity." In his book *Synchronicity: An Acausal Connecting Principle*, Jung suggested that coincidence is not always the result of chance; that there is an additional force in the universe which causes "appropriate", "meaningful" things to happen at certain junctures. This is not, strictly speaking, a psychological hypothesis. To many it seems more metaphysical than scientific. But, taking into account Bell's Inequality and the quantum theory of measurement, one may see it in a rather different light.

Bell's Theorem from quantum physics implies that systems which have interacted previously will be **correlated** in the future. The simplest example is two electrons, once coupled but now very distant -- if one is **observed** by some consciousness to spin one way then the other one automatically spins the other way. But this example is only the easiest to visualize; the same sort of thing happens with complex systems that interact then separate. When the entropy of the probability distribution of the possible states of **one** system is decreased through observation, the entropy corresponding to the other system is automatically decreased as well.

Stated a little differently, Bell's Theorem is about **emergent pattern**. It does **not** state that patterns in one part of the universe will cause similar patterns to emerge in other parts of the universe. But it **does** state that emergent patterns will spontaneously form, spanning distant systems which have been "physically unrelated" for a long time. That is what coincidence is: it is a **pattern emerging** between apparently unrelated events.

Therefore, according to accepted principles of quantum physics, **looking** at the world will in general cause certain emergent patterns -- certain coincidences -- to form. This scientifically validates Jung's basic intuition, in the abstract. I have trouble believing some of the examples which he gives in *Synchronicity*. I suspect that virtually all of the coincidences that occur in everyday life are genuine chance phenomena. But, interms of quantum physics, the **scientific possibility** is there for some coincidences to be more than that.

11.4.2. Wheeler's Vision

Over the last two decades, John Archibald Wheeler -- a leading gravitational physicist and the originator of the term "black hole" -- has become a sort of radical activist within the theoretical physics community. His goal is a physics which acknowledges the fact that, while physical reality creates observers (such as humans), observers also create physical reality. And he has argued that contemporary scientific ideas are largely inappropriate for this goal.

[N]o alternative is evident but a loop, such as: Physics gives rise to observer-participancy; observer-participancy gives rise to information; and information gives rise to physics.

Is existence thus based on "insubstantial nothingness"? Rutherford and Bohr made a table no less solid when they told us it was 99.99... percent emptiness. Thomas Mann may exaggerate when he suggests that "we are actually bringing about what seems to be happening to us," but Leibniz reassures us that "although the whole of this life were said to be nothing but a dream and the physical world nothing but a phantasm, I should call this dream or phantasm real enough if, using reason well, we were never deceived by it...."

Directly opposed to the concept of universe as machine built on law is the vision of a world self-synthesized. In this view, the notes struck out on a piano by the observer-participants of all times and places, bits though they are, in and by themselves constitute the great wide world of space and time and things....

First, elementary quantum phenomena brought to a close by an irreversible act of amplification. Second, the resulting information expressed in the form of bits. Third, this information used by observer-participants -- via communication -- to establish meaning. Fourth, from the past through the billennium to come, so many observer-participants, so many bits, so muchexchange of information, as to build what we call existence.

In the language of hypersets and functions, what Wheeler is proposing is that

a) mind = f(physical reality)

b) physical reality = g(mind),

for some functions f and g. In totally non-mathematical terms, this just means:

a) mind is defined in some way by physical reality

b) physical reality is defined in some way by mind

This proposal, made by a leading physicist, is obviously very much in the spirit of this chapter. I am not the only one to consider the possibility of reconciling of the **psychological** view of external reality as a belief system, and the **physical** view of external reality as a medium of specific dimensionality obeying specific dynamic equations.

11.4.3. Physics and Pattern

In recent years, a new approach to quantum measurement has emerged -- the **statistical** approach, pioneered by the physicist Asher Peres (1990). In compressed form, the essence of the approach is that **measurement is related to the *statistical* coupling of the measuring system and the object being measured**. This idea, I suggest, may be precisely what is needed in order to connect the physical world with the psychology of belief systems.

As Peres puts it, "a measuring apparatus **must** have macroscopically distinguishable states," where **macroscopic** is defined to mean "incapable of being isolated from the enviroment." Peres's thermodynamic arguments show that what is physically meant by "macroscopic" is nothing other than "statistically coupled with the environment." But a measurement device is defined as something with macroscopic states. Therefore, measurement is conceptually bound up with statistical correlation.

The same idea was hinted at years earlier by no less a physicist than Richard Feynman:

Proposal: only those properties of a single atom can be measured, which can be correlated (with finite probability) with an unlimited number of atoms.

Let us think about this carefully. A correlation is, essentially, a way of predicting the behavior of a whole group of entities from the behavior of a small subset of the group. In other words, a **correlation** in a collection of particles is a **pattern** in that collection. It is an "approximate pattern," according to the technical definition; but it is a pattern nonetheless.

What are we to make of Feynman's reference to an **infinite** number of atoms? Obviously there is not an infinite number of atoms in the universe, so if taken literally this implies that measurements never exist. But if one thinks in terms of pattern, the role of the infinite number of atoms here is easy to understand. A correlation among an infinite collection of atoms is bound to be a pattern in the collection of atoms, no matter who is determining what is a pattern and what is not. But a correlation among only finitely many atoms is, to a much greater extent, a matter of opinion: some observers may recognize it as a pattern, while others may not.

Thus, the statistical approach to quantum measurement implies that every property of a single atom which can be measured is actually a **pattern emergent between the atom and other atoms**. And how can one **tell** if a group of atoms are statistically correlated? Well, only by measuring them. But if **measuring** means **detecting a statistical correlation** -- then it follows that the atoms themselves are never directly measured, only collections of "properties" that are in fact statistical correlations among large groups of atoms.

One thing that this suggests is the radical possibility that **the physical universe is an attractor for the "cognitive equation."** It is known that each particle may be produced by certain configurations of other particles -- this is shown by the well-known catalogue of scattering diagrams. Capra, in his *Tao of Physics*, has illustrated this point for a nontechnical audience in a masterful way. The statistical approach to measurement implies that, furthermore, each particle is in fact definable as a collection of **patterns** among other particles (the specific patterns in question are statistical correlations).

This may seem to be a somewhat extravagant conclusion. If one wants to be less ambitious, however, one may at least conclude the following: if mind is pattern, and if all that we can physically measure are emergent patterns, then it follows that physical reality is in no way separate from mental reality. Insofar as we can measure it, physical reality is just a certain subset of the collection of patterns that makes up the mind. The only question is how the mind came up with the **temporal** patterns governing the behavior of those patterns that we call particles. For these "temporal patterns" are nothing other than the laws of physics.

11.4.4. Consciousness Revisited

Finally, what does all this have to say about quantum theory and **consciousness**? The verdict is unclear.

If the physical world consists of patterns, then the difference between the quantum world and the classical world has to do with the transition probabilities between patterns. In other words, it has to do with whether, given the problem of computing the joint probability of two independent events A and B, one

1) multiplies the probability of A by the probability of B (the classical view), or

2) uses the path summation formula (the quantum view)

The latter method involves the **interpenetration** of the two distinct events, A and B. The quantum theory of consciousness states that conscious intervention renders this kind of interpenetration impossible. In the context of the theory of consciousness given earlier, this implies that the **barriers** erected by consciousness around the patterns it processes somehow prevent **quantum-physical** interpenetration, as well as memory reorganization. Is this a sensible idea, or merely a surface correspondence between two fundamentally different things?

11.5. FEYNMAN INTEGRALS AND PATTERN PSYCHOLOGY

The previous section was one long sequence of suggestive speculations. Now I will cap the chapter off with an appropriate **grand finale** -- the biggest and most suggestive speculation of all. I will put forth theradical possibility that the laws of mind may be used to partially **deduce** the laws of physics, and perhaps even to resolve some of the pressing problems of modern physics.

This may seem to be a crazy idea. But one must recall that the hottest physical theory of the decade, string field theory, implies that the universe is a 26-dimensional space rolled up into a very thin 4-dimensional cylinder. In this light, it is hard to pronounce **any** approach to fundamental physics overly bizarre.

11.5.1. Perception and Paths

The early Gestalt psychologists showed that, given a number of possible ways of perceiving a figure, the mind will tend to choose the **simplest**. Similarly, the philosophical axiom called "Occam's razor" states that, all else equal, the **simplest** of a collection of competing explanations should be preferred. Phrased in terms of pattern theory, these two insights boil down to the same thing: that the mind tends to make the choice of **least algorithmic complexity** (where algorithmic complexity is measured relative to the perceiving mind). In *The Structure of Intelligence*, this view of induction and perception is discussed in great detail.

What if, then, one applies this rule of perception to **particle paths**? In quantum physics, a particle does not take one definite path from point A to point B; it takes "all paths at once." An action is assigned to each path; then these actions are summed up in a special way, yielding the probability that the particle goes from A to B. But there are numerous technical problems with the standard methods of assigning probabilities to the different paths. If one considers that the various paths **do not exist** except as perceived by some mind, then one immediately arrives at the conclusion that the probability of a path should be chosen proportionally to its algorithmic information, relative to the mind which is observing the path.

This would provide a "psychological" derivation of the dynamics of the physical world: the Schrodinger equation, Newton's Laws, special relativity and perhaps even general relativity. It would not immediately resolve the question of where **the spacetime containing the paths** comes from. However, Wheeler (1979) has proposed that spacetime itself may be obtained by amethod formally similar to path summation; this is the concept of "quantum foam." Perhaps, given a spacetime A at time t, all possible spacetimes for time t+1 exist at once, each one with a certain "generalized action." Then, summing up these actions according to the Feynman formula, one obtains the probability of going from spacetime A to spacetime B.

Whether this idea yields acceptable physical conclusions is not yet clear. At very least, however, it illustrates the viability of combining physical and psychological ideas. The two views of external reality are complementary and perhaps synergetic; they do not contradict one another.

11.5.2. The Feynman Path-Summation Formula (*)

Let q_it_i denote the proposition that a quantum system is in state q_i at time t_i. In his classic 1948 paper, Richard Feynman showed that the quantum-mechanical probability of a transition from q_1t_1 to q_2t_2 is given by $|(q_1t_1|q_2t_2)|\mathit{2}$, where I denotes the integration functional and

$(q_1t_1|q_2t_2) = I\ [e\mathit{iS(q)/h}]$ **(*)**

The integral is taken over all classical paths from q_1t_1 to q_2t_2; S(q) is the Lagrangian of the path q, and

h = **(**)**

is the normalized Planck's constant.

This version of quantum dynamics is not only elegant but remarkably generalizable. All contemporary theories of particle physics -- from quantum electrodynamics to electroweak theory, chromodynamics, grand unified field theory and even string theory -- can be cast in the form of equation **(*)**, with different interpretations for q and different forms for S (Feynman, 1950; Bailin and Love, 1986; Rivers, 1987; Ramond, 1981; Green, Schwartz and Witten, 1987). The integration variable q becomes not a classical path but a classical field, or a field defined over a Grassmann algebra, etc. -- but the basic concept remains the same. In a general context, equation **(*)** says that a quantum system assumes **all possible spacetime configurations** consistent with its observed behavior -- it is a "sum over all possible spacetime configurations." But, for simplicity's sake, I will continue to refer to **(*)** as a "sum over all possible paths."

Given the tremendous importance of oscillatory integrals of the form **(*)**, it is a curious fact that the entity "dq" has received no proper definition. As a standard particle physics text puts it, this differential is "just a fancy way of hiding our lack of knowledge about the measure" (Ramond, 1981).

Because **(*)** is purely oscillatory, one cannot define it directly using Wiener measure. Attempts to get around this problem have been few and far between. Feynman himself simply used approximations to the integral, without formally taking the limit. And that is still a common approach. But among more theoretically inclined physicists, the most popular strategy for understanding **(*)** is analytic continuation: one removes the i to obtain a real integral, defines the real integral in terms of Wiener measure, then obtains the integral in **(*)** as the continuation of this real integral onto the imaginary axis. This allows one to study Feynman integrals using standard methods from statistical mechanics (Simon, 1979). But it is intuitively most unsatisfactory. It does not represent **(*)** as a sum over all possible paths.

In 1967, Ito came up with a clever functional-analytic definition for "dq," but his method only works for a limited class of action functionals S; it does not generalize to relativistic quantum theory. A little later, Morette-deWitt (1974) suggested an interesting variation on Ito's approach. And, most impressively, in 1976 Albeverio and Hoegh-Krohn used the Parseval relation to give a fairly general Fourier-transform-theoretic definition of **(*)**. But none of these tricks is really satisfactory from a physical, intuitive point of view. They still do not represent **(*)** directly as a sum over all possible paths.

11.5.3. The Psychological Connection (*)

So, what is the solution? How can the gap between **equation** and **intuition** be bridged? One option which has not been explored is to introduce the physical **Church-Turing Hypothesis** -- the idea that the physical world must be **computable**. This principle, pursued by Joseph Ford (1985), Edward Fredkin (Fredkin and Toffoli, 1982; see also Wright, 1989) and others in different areas of physics, states quite simply that **uncomputable entitiesdo not physically exist**. If one accepts the computability principle, then it follows that, when computing path integrals, one should not integrate over uncomputable paths. But the number of computable paths is only countable, and thus the computability principle may well render **(*)** much less formidable.

There is, of course, a catch. The problem of defining **(*)** has typically been cast in the form: find a **measure** on the space of all possible paths from q_1t_1 to q_2t_2, under which oscillatory integrals of the form **(*)** can exist under general conditions. But if one is to make sense of the concept of integrating over computable paths only, one must weaken the concept of measure to that of **finitely additive measure**. A finitely additive measure (f.a.m.) is a nonnegative-valued set function m which obeys the rule

m(A **union** B) = m(A) + m(B)

whenever A and B are measurable and disjoint. As the name suggests, to go from a measure to an f.a.m., countable additivity is replaced by finite additivity. One can easily define the Lebesgue integral with respect to an arbitrary f.a.m. Many of the nice results of measure theory do not carry over; but if one could obtain convergence, this would be a small price to pay.

What sort of f.a.m. might be appropriate here? **This** is where the psychological connection comes into play. If one accepts that physical reality is psychically constructed, then it follows that those paths that are **simpler** to the constructing mind should have a higher probability of being followed. In other words, the probability of a path should be proportional to its **algorithmic information content** relative to the mind doing the measuring. This idea imposes the pattern-theoretic analysis of mind on the **physical world**, in an elegant, if technical, way.

The Feynman path summation formula itself may be seen as an incredibly intense pattern in the lower levels of the mental network. The Feynman formula implies that P[A and B] need not equal P[A]*P[B]; but nothing in the dual network model implies that the classical rules of probability must hold. In our everyday world, ordinary probability theory approximates the quantum probability formulae tolerably well. But the dual network model would apply just as accurately were this not the case.

A specific particle path is a somewhat less intense pattern in the lower levels of the dual network. But thesimpler a path is, the more intense it can be as a pattern. Gestalt laws of perception specify that, out of many possible ways of seeing something, the **simplest** will tend to be chosen. This is also implied by the pattern-theoretic analysis of induction: given a number of possibilities, the mind will automatically assign a higher probability to the algorithmically simpler choices. What is being suggested in the section is that this rule of perception should be included as a part of the **laws of physics**. For, after all, the physical world does not exist until it is perceived.

11.5.4. Perturbation Theory (*)

To see the possible usefulness of this kind of f.a.m., let us recall how **(*)** is actually used to study concrete examples of particle behavior. At present there are two fundamental strategies, perturbation expansions, and lattice approximations; but the former is by far the more popular. In the perturbation approach, one first lets t_1 and t_2 tend to infinity in **(*)**, thus arriving at an entry of the "scattering matrix" S. Then, one expands the integrand in a Taylor series in terms of some coupling parameter, and integrates the series term by term, obtaining a "perturbation expansion" of **(*)**. Finally, Feynman diagrams are read off from the first two terms of this perturbation series, giving an excellent intuitive and quantitative model of particle interactions.

The trouble is, when one proceeds in this way, one tends to obtain infinite integrals. Thus one must use the technical procedure of renormalization, which allows one to "subtract off" these infinities, leaving only finite integrals. In the case of quantum electrodynamics, renormalization gives results that agree with experiment to a remarkable degree. The results for chromodynamics, electroweak theory and grand unified theory are not so clear, partly because for the Lagrangians involved in these theories, tractable perturbation expansions are very difficult to come by.

But it seems quite plausible that, if one uses an appropriate f.a.m. defined in terms of algorithmic information, one might be able to get **(*)** to converge for the action functionals involved in physics. This would imply that the infinite integrals which necessitate renormalization are not inherent in **(*)**, but are rather an artifact of the method of perturbation expansion.

The reason to suspect that algorithmic-information f.a.m.'s might allow one to bypass these divergences is quite simple: these f.a.m.'s have a certain natural decay property. They are not smoothly peaked like Gaussian measures, but they are peaked on a very coarse scale. In short, algorithmic-information f.a.m.'s impose an **effective cutoff** on **(*)** in a natural way, an effective cutoff which is qualitatively quite different from the artificial cutoffs imposed in renormalization theory. Lacking a detailed analysis, one can at least say that these f.a.m.'s suggest that, **once one commits oneself to a computable universe, an effective cutoff point is inevitable**.

11.5.5. Conclusion

So, what's the bottom line? The jury is emphatically out on the speculative physical theory of this section, on the use of algorithmic information f.a.m.'s to simplify Feynman integrals. But my purpose in outlining this theory here is to illustrate in detail **the possibility of integrating psychology with physics**. The view of the physical world as a belief system does not contradict the existence of detailed theories of physics. Far from it: the two views are complementary, and beyond this they have an immense potential to enhance one another.

Chapter Twelve

DISSOCIATIVE DYNAMICS

I have analyzed the mind as a collection of interconnected, intercreating processes; and I have proposed that the overall connectivity structure of this collection is that of a **dual network**. The dual network structure, however, is extremely flexible; it encompasses many possible patterns of connectivity. One parameter which varies widely among these possible connectivity patterns is the **degree of modularity**.

Fodor (1987) has argued that human perceptual processes are strongly modularized, in the sense that most vision processes need connect only with other vision processes, most hearing processes need connect only with other hearing processes, and so forth. The origin of this modularity is as yet unknown -- some of it probably results from straightforward genetic programming, but the greater part of it may well **self-organize** as a part of the infant brain's growth process. Neural Darwinism suggests that, if there did arise significant connections between low-level vision processes and low-level hearing processes, these connections would quickly disappear due to lack of utility.

In a similar way, it is quite possible for **higher levels** of the dual network to become modularized. In this chapter I will use the word **dissociation** to refer to modularization which occurs as a result of childhood or adult mental dynamics, as opposed to modularization which is present in the brain at birth. For instance, I will talk about **personality dissociation** --

dissociation involving subnetworks involving significant portions of the self/reality belief system.

This is an abstraction and generalization of the revolutionary concept of "dissociation" introduced by psychologist Pierre Janet around the turn of the century. For nearly three quarters of a century Janet's work was ignored, assumed to have been superseded by psychoanalytic ideas. In 1970, however, Ellenberger's *Discovery of the Unconscious* rescued Janet's work from obscurity and brought it to the attention of the psychological community. Since that time, dissociative phenomena have been studied with increasing vigor, mainly in the context of multiple personality and post-traumatic stress disorder; and they have been related with modern brain science in a satisfying way (Modell, 1980; Mitchell, Osborne and O'Boyle, 1985).

Here I will integrate the classical concept of dissociation with the dual network model and the cognitive equation, thus arriving at a striking new framework for understanding mentality, one which synthesizes and (hopefully) clarifies all the ideas of the previous chapters. The central claim of this framework is that partial personality dissociation is central to the formation of structurally conspiratorial belief systems; which are in turn essential to productive, creative **logical thought**. Or, in a formula: no powerful intelligence without strong internal conflict. What might at first seem an obstruction to logic, is in fact necessary to the **evolution** of useful logic-guiding systems within the mental network.

This synthetic framework serves to bring the abstract psychology of the previous chapters closer to everyday human life. For this reason, it should be of interest not only to psychologists, but to anyone concerned with better understanding their **own** mentality. It gives, for perhaps the first time, a sensible idea of how the creative diversity and **per**versity of human personality might emerge from the evolutionary dynamics of neural pathways.

And in the final section, I will argue that this framework is pregnant with implications, not only for human psychology, but for **engineering**. I will propose a new kind of computer science called **A-IS**, or **artificial intersubjectivity**. A-IS centers around the idea of programs which socially interact with one another and hence develop interrelated, dissociated personality structures. Only in this way, I contend, could computers ever simulate or supersede the wonderfully chaotic reasoning of the human brain.

12.1. MULTI-MENTALITY

'The World is One!' -- the formula may have become a sort of number-worship. 'Three' and 'seven' have, it is true, been reckoned as sacred numbers; but abstractly taken, why is 'one' any more excellent than 'forty-three,' or than 'two million and ten'?

-- William James

This quote is humorous, but at the same time it makes a very serious point. "Uni-" means one, so that the very word **universe** conceals a philosophical presupposition. Why should **unity** be a fundamental character of the world? Who says the world doesn't have diversity, rather than unity, at its core? Why not a **multiverse**, rather than a universe?

What form might a multiversal world take? William James was interested in **subjective realities** -- in the "semi-real" realities perceived by individual minds. He wanted to understand the world as an indeterminately large group of interacting, intersecting subjective realities. And he wanted to tie this in with the pragmatist idea that only **observable properties** are real. But he was disappointingly vague on the details.

By replacing the word "World" with the word "Mind" in James's quote, one obtains an equally valid **bon mot**:

'The Mind is One!' -- the formula may have become a sort of number-worship ... why is 'one' any more excellent than 'forty-three' or 'two million and ten'?

William James broke new ground with his theory of the "stream of consciousness"; he was also one of the first to seriously question the unity of mental experience. In the "stream of consciousness" metaphor, he did not rule out the possibility of **rocks** or **islands** in the stream, breaking up the flow into several distinct pieces.

Perhaps the deepest-ever insight into the fundamental **diversity** of the psyche was achieved by the novelist Fyodor Dostoevsky. In *The Idiot*, for example, the angelic but tragically unstable Prince Myshkin represents an aspect of Dostoyevsky's own consciousness. Myshkin thinks only good of other people; his only desire is to help. Now this certainly does not describe the man Fyodor Dostoevsky. But Dostoevsky felt and acted thisway at certain times; Myshkin was one of his **sub**personalities.

And in *The Brothers Karamazov*, the four brothers Ivan, Dmitri, Alyosha and Smerdyakov may be understood to represent separate "voices" in Dostoyevsky's mind, independent "streamlets" of Dostoevsky's consciousness. Alyosha is a less pathological Myshkin, the Myshkin sub-personality tempered by the realism of the rest of Dostoevsky's mind. Dmitri is a sensualist, a confused womanizer and gambler; Dostoevsky, under the influence of his Dmitri aspect, gambled his savings away many times. Ivan is a writer and philosopher, tirelessly agonizing over the problem of God in the modern world. Finally, the half-brother Smerdyakov represents the "worst of Dostoevsky," the evil, petty, vindictive, cunning sub-personality that all of us possess to some degree. As his diaries suggest, Dostoevsky viewed his own life as a constant struggle between these various sub-personalities, these competing modes of consciousness.

In recent years, psychologists have rediscovered this Dostoevskyan notion of multi-consciousness. Multiple personality patients like Sybil and Billy Milligan are virtually household names. And several psychological theorists have proposed that the kind of "dissociation" apparent in multiple personality is different in **extent** rather than **kind** from the mental dissociation observable in the ordinary person.

This is what Somerset Maugham meant when he wrote

There are times when I recognize that I am made up of several persons and that the person that at that moment has the upper hand will inevitably give place to another.

Maugham did not have multiple personality disorder -- each of the "several persons" making up his psyche was aware that its name was Somerset Maugham, and was aware of **most** if not all of the experiences had by the other "persons." But Maugham, like Dostoevsky, was a good enough self-observer to recognize that his mind was to some degree **dissociated**; that it consisted of several largely disconnected "functional personality units."

Ronald Fischer (see McKellar, 1979) reports experiments in which people are asked to memorize material under the influence of alcohol. When sober they exhibit poor recall -- but when given alcohol again theirmemory **improves**. This illustrates the phenomenon of "state-dependent memory." John does not have one unified memory -- drunken-John has his own memory, as does sober-John. The experience reported by Somerset Maugham is one step beyond this -- several personality units, each possessing its own "unit-dependent memory" as well as access to a shared memory store. And multiple personality disorder is but one step further: an amazingly large part of the shared memory store is divvied up among the various independent personality units.

The fundamental multiplicity of mind and world was expressed beautifully by the Russian philosopher Mikhail Bakhtin in his masterwork *Problems of Dostoevsky's Poetics*:

It should be pointed out that the single and unified consciousness is by no means an inevitable consequence of the concept of a unified truth. It is quite possible to imagine and postulate a unified truth that requires a plurality of consciousnesses, one that cannot in principle be fitted into the bounds of a single consciousness, one that is, so to speak, by its nature **full of event potential** and is born at a point of contact among various consciousnesses. ...

Not a single objective world ... a **plurality of consciousnesses, with equal rights and each with its own world**, combine but are not merged in the unity of the event.

12.2. DISSOCIATION AND THE DUAL NETWORK

As we have seen, the concepts of multi-reality and multi-consciousness are far from novel; they date back at least a century, to Janet, Dostoevsky and James. Up to this point, however, these ideas have not received a systematic theoretical analysis. I suggest that the **dual network model** provides the key to understanding dissociative psychological phenomena.

Recall that the dual network model analyzes mind in terms of two semi-autonomously functioning networks: an **associative memory network**, which self-organizes itself according to the principle that related entities should be stored "near" each other; and a **perceptual-motor hierarchy**, which operates according to the multi-levellogic of a flexible command structure. And it makes the central hypothesis that **these two networks are superposed**.

This superposition implies a roughly "fractal" structure for the associative memory network. And, more to the point, it implies that, if a section of memory is somehow split off or "dissociated" from the rest of memory, then a section of the mind's **control network** is also split off, as an automatic consequence. This explains, in one immensely simple step, how the attempt to suppress unpleasant memories can lead to the creation of an autonomously acting and remembering psychological unit. In other words, as will be shown in detail below, it explains the basic phenomenon of **traumatic memory** and post-traumatic stress syndrome.

Multiple personality is a little more complex: it has to do with the **self**, an intricate self-referential construction and a complex belief system. However, we may make a few simple observations. Post-traumatic stress syndrome is often a consequence of a **single painful event** -- e.g. watching a close friend die a bloody death. Multiple personality, on the other hand, is generally a consequence of **repeated painful events**, usually **beginning in early childhood**. Very often these events are incestual rape, or severe child abuse.

In post-traumatic stress syndrome, the painful event usually occurs **after the person's self is formed**. The person already has a unified self-image, so if his mind wants to shut off offending memories, it has to shut them **away** from the well-formed self. In multiple personality, though, the painful events occur while the person's self is still forming. Therefore, the "split off" memories are subjected to the self-formation process, just as much as the rest of the dual network. While not a complete explanation, this gives some idea of why multiple personality disorder should exist, and why different types of traumas should give rise to different psychological problems.

12.2.1. Dissociation and the World

On a more philosophical level, the dual network perspective makes clear that there is not so much difference between

1) the various personalities of a person suffering from multiple personality disorder (MPD)

2) the various personalities which exist in the world

3) the various sub-personalities of a normal person

Just as MPD results from the splitting-up of a single person's "dual network," so do individual personalities result from the splitting-up of the **universal dual network**. This idea unifies Dostoevsky's psychological idea of multi-consciousness with James' philosophical idea of a multiversal world. It is a dramatic conclusion -- but at the same time it is a new beginning. For it opens up a whole new way of looking at the mind and world: as **multiple** phenomena.

Far from being isolated pathologies, dissociative mental disorders are natural and **necessary** features of mental life. In other words, all mental action is a kind of interplay between different "personalities" -- different semi-autonomous agents, which help to mold one another's reality, which possess individual "senses of identity," which partially share the same memory, and which compete with one another for attention.

Aside from traumatic experiences, what might **cause** a section of the dual network to split off and become semi-autonomous? The answer to this is surprisingly simple. Two things only are required. In order to split off and **survive** on its own, a subnetwork must first be **complete in itself**, in the sense of being a strong attractor of the cognitive equation. And second, it must have relatively **few connections** with the other parts of the mental networks -- otherwise its autonomy would not last.

From this description it should be obvious that a dissociated subnetwork has something in common with a structurally conspiratorial belief system. The difference is not absolute, it is one of degree. A "subnetwork" is expected to have a more marked dual network structure than a belief system, which may contain few levels and display the dual network structure only to a small degree.

One way to distinguish the two is with **depth-to-breadth ratio**. A belief system tends to involve a sizeable collection of beliefs on roughly the same level of abstraction -- the same level of the **hierarchical** mental network. It is "shallow" but "broad." On the other hand, a dissociated subnetwork like a subpersonality tends to span a fairly large number of different levels of the hierarchical network; its depth exceeds its breadth. In other words, a dissociatedsubnetwork contains all the levels needed to **do**, whereas a belief system only **guides** other systems in doing.

12.2.2. The Subtlety of Personality

A dissociated personality subnetwork or **subpersonality** is centrally concerned with two things:

1) constructing the reality perceived by the mind, and

2) constructing the self-image "perceived" by the mind

Separate personality subnetworks are interconnected in the sense that they access, to a great extent, the same memory store. And they also have in common certain parts of the self/reality system, particularly the lower and more basic levels.

What makes human beings so interesting is that, by altering the common aspects of the self/reality system, and by altering the associative memory structure, each subpersonality affects the **environment** in which the other subpersonalities live. Thus, relations between subpersonalities of a mind are somewhat more intense than relations between people in the physical world. Perhaps the best physical-world analogy for the subpersonalities of a single mind is a **community of psychokinetics**, each one living a normal life, but also continually altering the physical world in response to the alterations made by the others. In such a community, one could never be sure what was "objectively there," and what was merely placed there by somebody else for some particular purpose. This is precisely the situation with which subpersonalities are presented.

12.3. TRAUMATIC MEMORIES

Evolutionary psychology reveals that partial personality dissociation is not only normal but **necessary for efficient mental functioning**. In the history of psychology, however, the main role of the concept of dissociation has been in the characterization of various **pathological** mental conditions. To help bridge the gap between these two perspectives, in this section I will discuss perhaps the simplest form of pathological dissociation: **traumatic memory**, and the related "post-traumatic stress syndrome."

Rape and violent wartime combat might seem to be rather memorable occurences. But sometimes traumatic experiences such as these are not stored in a person's memory in the ordinary way. Instead, they seem to enter the mind and disappear; they are shut off from conscious memory and reflection, until in certain situations, they pop up intensely and unexpectedly, rendering the "rememberer" mentally dysfunctional.

In the words of van der Kolk and van der Hart (1991),

> Lack of proper integration of intensely emotional arousing experiences into the memory system results in dissociation and the formation of traumatic memories. Janet called these new cores of consciousness "subconscious fixed ideas." [T]raumatic memories of the arousing events may return as physical sensations, horrific images or nightmares, behavioral re-enactments or a combination of these. Since fixed ideas have their origin in a failure to make sense of a past experience, they fulfill no further useful function and lack continued adaptive value.

Janet's term "fixed ideas" is reminiscent of the dynamical term "fixed point." It is suggestive of the idea that traumatic memory systems, like structurally conspiratorial belief systems, are **attractors** to the cognitive equation.

All in all, the phenomenon of traumatic memory fits in well with the dual network view. Why do the traumatic memories "split off" and become autonomous? Because, it seems, certain experiences are simply **difficult to connect** with the remainder of the mental network. The mind tirelessly seeks to improve its organization, to cut-and-paste parts of the traumatic-memory subnetwork with elements from the rest of the mind. But these attempts fail; they lead only to nightmares, re-enactments of the traumatic experience, and so forth.

And why does the mind fail in its attempts to re-organize and integrate the traumatic experiences? **Not**, as one might think, primarily because there may be few connections to be drawn, but rather because those connections that **could** be drawn would be painful ones. When reorganization hits on a **real** connection, this connection itself causes severe unfulfillment of expectations, which is the definition of strong emotion. Moreover, the specific nature of this unfulfillment is a feeling of **decreasing order** -- a feeling of disruption of previously coherent thought systems. This is precisely, according to Paulhan (1880) and *SI*, the definition of **unhappiness**.

But when "correct" reorganizations are continually rejected because of induced unhappiness, the very **reorganization** processes become confused. They futilely seek to adjust and improve their algorithms and strategies. The behavioral result is that traumatized individuals react to stressful situations with irrelevant movements, emotions and thoughts that represent fragments of their traumatic memories. As Janet (1904) put it, it is "as if their personality development has stopped at a certain point and cannot expand any more by the addition or assimilation of new elements." In the most extreme case, there is the phenomenon of **re-enactment**. A person may repeatedly go through the exact words and physical motions of a traumatic experience, yet still be unable to answer simple questions regarding these words and motions. This implies that the traumatic memories are not integrated with the higher-level verbal and cognitive sections of the dual network.

Normally we retain high-level patterns in our experiences, and very little of the experiences themselves. But the situation with traumatic memories is just the opposite. They have not been subjected to the usual rearrangement-based **pattern recognition** processes, because these processes proved too painful. Instead, they have been retained as a full-fledged subnetwork of perceptions and actions, untouched by rearrangement. In Janet's words,

The person must not only know how to do it, but must also know how to associate the happening with the other events of his life, how to put it in its place in that life-history which each of us is perpetually building up and which for each of us is an essential element of his personality.

What sort of therapy helps people suffering from traumatic memories? What is needed is to get the relevant rearrangement processes back to their prior state of productivity. One useful strategy is to introduce ideas which are **related** to the traumatic memories, but easier to integrate into the remainder ofthe memory. For instance, many women stigmatized by rape have been helped by imagining that **they** have all the power in the world, and are applying it to the perpetrator. This allows the specific memories of the rape to be cut-and-pasted with other elements of memory, in a less painful way.

So, in sum, what differentiates traumatic dissociation from healthy dissociation? Traumatic memories are a case of **forced dissociation**. They represent **combat** between the hierarchical and heterarchical structures of the dual network. Integrative rearrangements of the traumatic memories are "successful" by the standards of the associative memory network; they lead to common pattern. But they are rejected by the control network due to the unhappiness they generate, the abundance of unfulfilled commands. Successful, healthy dissociation, on the other hand, is harmonious with the entire dynamic of the dual network: it involves a division into successfully functioning parallel subnetworks, which deal with different, relatively unrelated problems. For this very reason, healthy dissociated subnetworks are able to deal with the common segment of memory without fear of wreaking havoc.

A traumatic memory subnetwork must isolate itself from the rest of the memory, or else risk causing distracting, troublesome pain. Thus traumatic subnetworks can never truly be functional. No subnetwork of such small size can be truly complete in itself -- the task of intelligence is too difficult for that.

12.4. DISSOCIATION AND THE STRUCTURE OF BELIEF

I have said that the self/reality belief system is a tool for guiding the construction of other belief systems. Boundary-setting, discussed in Section 12.2.6, is one example of this "guiding" dynamic at work. Another example, I suggest, is the topic of this chapter: dissociation. A small child learns dissociation in the context of her self/reality belief system. This dissociated structure, then, serves to create other autonomous mental structures -- in particular, structurally conspiratorial belief systems, which are crucial for the production of creative **logical** reasoning.

To explain **why** this should be true, let us begin chronologically. Once a child learns that she must **act** different ways in different situations, then she will inevitably develop relatively autonomous personality subnetworks corresponding to the different situations. These subnetworks will not achieve the degree of separation observed in multiple personality patients, but they may well have different likes and dislikes, and different ways of responding to the same stimulation.

This process may also be looked at linguistically, in terms of the theory of language given above. As a child learns that the same **words** have significantly different **meanings** in different situations, she will develop a semantic system with distinct subsystems, and these subsystems will take the form of semi-autonomous subnetworks of her dual network. And the same thing that happens with spoken language, will also happen with the **language of behaviors** (as discussed earlier) -- thus resulting in semiautonomous **personality** subnetworks.

Now: these different sub-personalities, though they may have arisen in specific social situations, may well emerge on cue in situations different from those which elicited them. The way a person deals with any given issue may be determined by **different sub-personalities** at different times. Thus there is a kind of **evolutionary competition** among subpersonalities.

The result of this competition, I suggest, is that a sub-personality will flourish to the extent that it can create belief systems which

a) support its interests, and

b) stand little chance of being destroyed by other sub-personalities

Quite clearly, the best way to achieve (b) is to create **structurally conspiratorial** belief systems. If a belief system depends on outside factors for its survival, these factors may well shift when the controlling subpersonality shifts. But if a belief system can survive on its own, then it has a much better chance of "waiting out" an hostile environment.

To see the importance of this, recall the conclusion reached in Chapter Ten, that productive belief systems tend to be those that receive significant support both externally **and** conspiratorially. External needs are too strongly fluctuating to be relied upon as a sole source of support.

But how do these structurally conspiratorial belief systems develop in the first place? Yes, they areattractors of the cognitive equation, so they may be arrived at by "accidental iteration." But, to use evolutionary terminology, how much better to have a force explicitly **selecting for** structural conspiracy! This is exactly what dissociated personality networks do. Each of the competing subnetworks **specifically reinforces** related subnetworks that operate by structural conspiracy, and are hence not easily disrupted by competitors.

My contention is that this specific selective pressure is mentally **necessary**. It is not necessary for the maintainance of structural conspiracies, which by definition maintain themselves. Rather, it is useful for the maintainance of belief systems that, while **close** to being structurally conspiratorial, are not yet truly self-supporting. The iteration of the cognitive equation is **mind-wide**; it is not restricted to the individual subnetworks that happen to be converging to attractors on their own. It will tend to mix up subnetworks even if they are somewhat close to being autonomous. The extra push toward autonomy may often be needed; and personality dissociation may thus be a crucial part of the development of **effective thinking and acting**.

12.4.1. Dissociation and Logic

The social uses of dissociation are obvious. In today's society, it rarely pays to have the same personality at work and at home. But what I am claiming here is something much stronger and more radical. I am claiming that partial personality dissociation is not only socially but **cognitively** necessary. By biasing the selection of belief systems toward the structurally conspiratorial, it also biases the selection of belief systems toward the productive. Or in other words: no dissociated personality, little chance of systematically creative belief production.

And this brings us back toward **logic** and **reasoning**. Logic, if you recall, requires a semantic, analogical system to guide it. And the **quality** of a chain of logical reasoning depends at least as much on the productivity of this system as on the cleverness of the deductive rules. The conclusion? Without dissociation, ideological, paranoid and otherwise pathologically conspiratorial belief systems would be rare. But so would be productive belief systems; and hence, so would be creative logical thought.

This, finally, is the true meaning of the phrase **chaotic logic**. Dissociated personality networks, and the structurally conspiratorial belief systems which they encourage, are attractors of the cognitive equation, supporting apparently chaotic dynamics. But without these strange attractors, the rich reserve of analogies required for deductive logic would never be created. Logic thrives on chaos. And, conversely, logic itself is a crucial tool of these belief systems and sub-personalities; it aids them in maintaining their attractor status ... chaos thrives on logic.

12.4.2. The Meta-Dynamics of Paranoid Belief

A complete treatment of the practical psychological implications of abstract "dissociative dynamics" would be out of place here. However, it seems worthwhile to give at least a few hints. Toward this end I will now briefly return to Jane's paranoid belief system, discussed extensively in earlier chapters.

Now I will be able to say a little more about the **possible origin** of this paranoid system. But as before, I must emphasize that this analysis is not intended as a definite **diagnosis** of Jane's specific problems, but only as an **illustration** of certain general principles.

Jane demonstrates many, many different dissociated subpersonalities. Chief among these, however, are: 1) an **obsessive** subpersonality, in which the world is perceived as hostile and in need of constant mocking scrutiny; and 2) a **happy-go-lucky** subpersonality, in which she makes an excellent impression on others, and is good-natured almost to the point of being giddy. These are not full personalities; they share most of the same memories. But on the other hand, they are not merely **moods** either; they are alternate systems of perceiving and classifying data.

The alternation between these two subpersonalities might perhaps be characterized as "manic depression." But obviously there is more to it than that. It would seem that, at very least, there is an unusually complex form of manic depression at work here.

In the obsessive subpersonality, Jane is overly attentive to facial expressions, the colors of clothing, the letters on license plates, and so forth; she is constantly categorizing things in unusual ways. She demonstrates perceptual patterns that might be called "compulsive," and her behavior tends toward the unusual and offensive. She will often act out specifically toshock people; cursing, flashing, making faces, and so forth.

In the happy-go-lucky subpersonality, on the other hand, Jane is open-minded and accepting toward other people's ideas. She tends not to notice details of her surroundings, and her behavior is generally quite unexceptional, except for perhaps a slight overexuberance. She is a pleasant companion and a good conversationalist.

The worst of Jane's depressed moods seem to occur when she is in her obsessive subpersonality, and she is unable to find an **external source** to blame for her problems (most of which are caused, of course, by the paranoid behavior of the obsessive subpersonality). The happy-go-lucky subpersonality is not so concerned about these problems, and thus is not worried about where to place the blame. But every time the obsessive subpersonality comes back again, it needs to once again begin its quest for an external source to blame.

Therefore, obviously, it is in the interest of the obsessive subpersonality to create a blame-placing belief system which will **persist** even when the happy-go-lucky subpersonality is in charge. How can this be done? One way, of course, is to create a **structurally conspiratorial** blame-placing belief system; a system that will maintain itself indefinitely, that will keep itself going even when the reigning subpersonality has no use for it. Perhaps the obsessive subpersonality will experiment with many different strategies for apportioning blame; but those which are less conspiratorial will be less likely to survive the fluctuations of control. Personality dissociation provides a **selective force** in favor of structural conspiracies -- such as Jane's paranoid belief system.

12.4.1.1. A More Detailed Model

In slightly more detail, one may say that the obsessive subpersonality contains the following beliefs:

D_0 = I am unloved

D_1 = I am good and lovable

D_2 = They are bad

This system is not in itself an attractor for the cognitive equation; it is partially self-supporting, but it also relies on the remainder of the mind.

The dynamics here are simple enough. D_0 chips away at D_1; but D_1, acting on D_0, helps to produce D_2. And D_2, acting on D_1 and D_0 collectively, helps to produce D_1, thus counteracting the effect of D_0 (if one is not loved by bad people, that increases rather than decreases one's goodness).

But the problem is that D_0 is a **self-reproducing** belief: it is a pattern in the behavior which it produces. It would seem that perhaps D_0, and the behavior systems to which it is connected, are **in themselves** an attractor for the cognitive equation. For the behavior system is produced by D_0 and its own internal dynamics; and D_0 is produced by the behavior system.

The effect of D_0 on D_1 is so strong that D_1 is powerless to counteract it, even via D_2. So what could be more natural than to counteract D_0 by making D_2 self-perpetuating -- by making it a structural conspiracy. **This** is what is accomplished by the conspiratorial belief system described earlier. This entire belief system, with all its complex dynamics, is merely a way of making D_2 as strong as possible.

This, on a deeper level, is the meaning of Jane's refusal to take blame. Taking blame for anything subtracts from D_1, which is already in serious trouble. But the conspiratorial belief system within D_2 works along with D_1 to counteract the powerful effect of the self-reproducing belief D_0 -- which is, most likely, the root of the whole problem.

This is still a very partial, sketchy analysis of Jane's situation. But it does serve to illustrate the perverse complexity of the mind. One sees belief-system attractors grow within subpersonality attractors, and spawn new belief-system attractors in the common memory, generating a hierarchy of chaotic pattern dynamics -- and all to counteract the runaway self-perpetuating growth of a single belief of the utmost simplicity: "I am unloved."

12.4.3. Dissocation and Creativity

In Jane's case, dissociative dynamics led to an undesirable, overly rigid belief system. But precisely the opposite result is also possible. To give a little bit more of the flavor of the implications of dissociative dynamics, I will now discuss very briefly two famous thinkers, and comment on the role dissociativedynamics played in the development of their thought. These two thinkers, Jung and Nietzsche, are extreme cases; they were more dissociated than most. But they provide an excellent illustration of how belief systems, once they have been made conspiratorial by dissociative dynamics, may also benefit from dissociation in more complex ways.

12.4.3.1. Carl Jung

In his autobiography, Carl Jung analyzed his life work as a result of cooperation and competition between two subpersonalities, whom he called "Number One" and "Number Two." Number One was scientific and practical; Number Two was spiritual and cared little for the material world. Each of the personalities erected its own belief systems: Number One a rational, objectivist view of the world, and Number Two a mystical perspective.

And each of these belief systems turned out to be strong enough to withstand those times when the **non-supporting** subpersonality was in control. The result was a mental network capable of incredibly powerful, uniquely creative logical reasoning. The **competition** between the two subpersonalities necessitated the development of much more robust belief systems than would otherwise have been necessary. And the robustness, the structural conspiracy of these belief systems, was crucial in providing analogies to guide Jung's masterful trains of thought.

For a simple example, consider Jung's concept of an "archetype" -- an abstract concept-structure or **meta-idea** which appears in myths, thoughts and dreams. A simple example is the "resurrection" theme of "the hero and rescuer who, although he has been devoured by a monster, appears again in a miraculous way having overcome whatever monster it was that swallowed him." This archetype may be found in a rather high percentage of movies, novels and television shows!

These archetypal images are not specific **pictures** -- the hero need not be big and strong, and the monster need not be a huge ugly green beast. The archetype is a **structure** -- in this case, it is a structure which consists of **roles** and **types of events**. Each role (hero, rescuer, monster) and each type of event (rescue, devouring, miraculous reappearance) is simply a certain collection of patterns, and each one may be fulfilled in a number of different ways. As Jung put it,

> Again and again I encounter the mistaken notion that an archetype is determined in regard to its content, in other words that it is a kind of unconscious idea (if such an expression be admissible). It is necessary to point out once more that archetypes are not determined as regards their content, but only as regards their form and then only to a very limited degree. A primordial image is determined as to its content only when it has become conscious and is therefore filled out with the material of conscious experience.... The archetype in itself is empty and purely formal, nothing but a... possibility of representation which is given *a priori*. The representations themselves are not inherited, only the forms, and in that respect they correspond in every way to the instincts, which are also determined in form only. The existence of the instincts can no more be proved than the existence of the archetypes, so long as they do not manifest themselves concretely. (Jung, 1934)

The collection of all archetypes, Jung called the collective unconscious. It is -- or so he hypothesized -- an inherited, a priori part of every human mind. Archetypes subtly guide all our feelings and acts.

Jung did not discover the notion of "archetype" by scientific, logical analysis; he discovered it by pure intuition, by **seeing** the mind as an abstract structure and thus understanding its dynamics. This was clearly a Number Two process. But a pure, spiritual intuition into the mind would not survive the scrutiny of Number One. In order to keep its insight, Number Two had to form the concept of "archetype" into a powerful, self-maintaining ideational system. Once this was done, then Number One not only refrained from destroying the concept; it latched onto it, refined and improved it, yielding the scientific notion of archetype that we have today.

12.4.3.2. Friedrich Nietzsche

For a more complex example, consider the philosopher Friedrich Nietszche, on whose work I have drawn so liberally in these pages. Nietzsche demonstrated at least two prominent subpersonalities. One was a mild-mannered, friendly and quiet philologist, who hated seeing pain and avoided causing anyone offense. Theother was the brilliant, arrogant madman whom one sees in works such as *Thus Spake Zarathustra* and *Ecce Homo*. Following the example of Jung, let us call these Number One and Number Two.

Number One forced Nietszche to hide the radical nature of his philosophy from casual acquaintances. On one occasion, when he saw a horse about to be whipped by its master, it caused him to stop and vigorously hug the horse. Number Two, on the other hand, impelled Nietszche to forsake classical philology and spend his life in a passionate quest to destroy all the ideas he had been raised to believe in: religion, morality, reality, truth. As has often been observed, Nietszche's philosophy encapsulates the contradiction between these two emotional views of the world.

Number Two supported an incredibly productive belief system of mistrust and skepticism, a disputative belief system capable of seeing the holes in any argument, and of combining and manipulating abstract ideas with great dexterity. This belief system was not merely nihilistic; it consisted of a repository of clever tools for demonstrating the falsehood and vanity of any point of view. One sees this system at its best in aphoristic works such as *Human, All Too Human*, *Dawn* and *The Gay Science*.

Number One, on the other hand, silently upheld the values against which Nietzsche's work railed. It supported a more traditional philosophical belief system: it perceived an underlying order in the universe, it respected the difference between right and wrong, and it had a powerful sense of spirituality. This was the belief system which governed Nietzsche's personal life. Number Two wrote tirades against asceticism; but Number One was responsible for Nietzsche's own ascetic lifestyle.

Nietszche's most dramatic ideas, the eternal recurrence and the will to power, may be seen as the result of **synthesizing aspects of these two conflicting belief systems**. The eternal recurrence is a cynic's version of afterlife. The will to power is a will superseding all notions of "free will" -- with its militaristic "order of rank," it is a morality "beyond good and evil." Zarathustra's beautiful sermons display an atheistic spirituality beyond all traditional concepts of Godliness. Much of the strength of Nietzsche's thought results from its dual source: two productive,structurally conspiratorial belief systems, usually competing but occasionally collaborating.

12.4.3.2. Conclusion

This very cursory study of two great thinkers indicates an important aspect of dissociative dynamics: namely, the possibility of **synergy** between competing belief systems. Two dissociated subpersonalities need not become unified in order for their respective belief systems to combine with one another. Jung and Nietszche are two examples of creativity emerging from the synergy between the structurally conspiratorial belief systems of **different subpersonalities**. In neither case was a complete synthesis attained; but in both cases, the interaction and partial reconciliation of conflicting systems proved tremendously productive. The role of this sort of synergy in everyday life and thought would seem to be a very fertile area for future investigation.

12.5. DISSOCIATION IN THE UNIVERSAL NETWORK

What is the difference between the dissociated subpersonalities of a given mind, and the separate minds in the world? After all, as noted above, the different minds in the world are just semi-autonomous subnetworks of the **universal** network. Are we all perhaps just subpersonalities of one particularly advanced multiple personality patient?

In fact there are two main differences between a collection of subpersonalities and a collection of minds. The first is that subpersonalities are mainly conscious **in sequence**, not in parallel. There is certainly **some** parallelism going on: one sub-personality may passionately declare "I love you" while another simultaneously and silently ridicules the remark. One subpersonality may raise a gun to shoot someone, while a competing subpersonality causes the legs to buckle, preventing the murder from occurring. But these are extreme cases; there is much **more** parallelism among the different personalities in the real world.

And the second outstanding difference regards **memory access**. Dissociated subpersonalities, although largely disconnected from one another, still have access to a common memory store. In normal mental functioning, every personality has access to **almost** every memory in thebrain; the main thing is that different memories are more **easily** accessible to certain subpersonalities than to others. State-dependent memory is important but not all-pervading.

Different personalities in the world, on the other hand, do not appear to have access to a common memory store. They are connected at the bottom, via physical reality, but this would seem to be the extent of their interconnection. This feature is shared by the various personalities of multiple personality patients -- for instance, one personality may speak Italian while the other does not. But even in these exceptional cases, there is still **some** degree of common memory, much more than between two different people.

Rupert Sheldrake's (1981) theory of the **morphogenetic field** attempts to destroy this distinction, claiming that each person's memory is **aphysically connected** to everyone else's. So that, for instance, once a thousand people learn the formula for solving cubic equations, a small "trace" of that knowledge becomes available to everyone, thus making the process of learning that particular formula universally easier. But this dramatic prediction remains unproven.

So, in sum, I have argued that the difference between the people in the world and the subpersonalities in one mind is a matter of **degree** rather than absolute distinction. There are serious differences in the amount of parallel consciousness and the existence or amount of common memory. However, there is a significant simlarity in that, just as different subpersonalities collectively create their "environment," different people collectively create their **reality**.

12.5.1. Why Not One Mind Only?

These observations lead to some rather interesting philosophical ideas regarding the nature of our collectively constructed external reality. Subpersonalities are behooved to encourage structural conspiracies, so that the processes they create will not be destroyed by other subpersonalities. And by the very same reasoning, **minds** will do well to create reality structures which are structurally conspiratorial, so that the reality structures they create will not be disrupted by other minds. This suggests that the immensely conspiratorial nature of the reality belief system ispartly due to its construction at the hands of **competing individual consciousnesses**.

In other words: a reality created by one consciousness alone would probably not be very interesting; it would have little generativity, because of the lack of structurally conspiratorial subcomponents. The competition of different minds encourages structural conspiracy and hence creativity. This is a novel, thought-provoking answer to the old philosophical puzzle of the **multiplicity** of consciousness. Why not, as the Buddhists would have it, one mind only? Because that path leads to a **boring** world. If **intricate structure** is a criterion of value, then multiple consciousnesses are valuable indeed.

This does not exactly give a **reason** for the multiplicity of consciousness. But it does give something to go on: the fact that the universal network is a multiple-consciousness attractor for the cognitive equation. If one accepts this equation, the only other thing to be taken on faith is that, starting from wherever it did, the universe eventually converged on the universal network structure. And chaos implies that there need be no real "reason" for this. Convergence to one attractor rather than another can be the result of pure chance.

12.5.2. The Future of Reality

These ideas may appear to be "out there" -- philosophical meanderings unrelated to any issues of practical substance. However, this perception is far from accurate. The ideas of this section are not merely theories about the relation between mind and reality, they are **computational** theories about the relation between mind and reality. And this means that they fall into the category of theoretical science, rather than philosophy. For, although current technology does not permit the relevant tests to be carried out in a reasonable time frame, these theories are **in principle** empirically testable.

To see this, consider the possibility of **virtual reality** technology, which would allow us to put our consciousnesses into simulated bodies living in simulated physical realities. Given this technology, it would be easy to experiment with different methods of collective reality construction. For instance, one could easily verify whether or not one consciousness is enough tocreate an intricate world -- whether or not, as I have claimed, defensive structural conspiracy is required.

Short of full virtual reality, it is also possible to conceive of **simulated realities**: collections of artificially intelligent programs that collectively construct their own simulated world. Though less dramatic, this would also permit direct empirical test of theories about the mind/reality relation.

12.7. ARTIFICIAL INTERSUBJECTIVITY

The dual network model and the cognitive equation are **computational** models. In this final section, I will briefly explore the possibility of using them to do practical computation: to design a **computer program** which displays the sensitive interplay of chaos and logic that today is only associated with human minds. I will describe a new type of algorithm, which I call an **artificial intersubjectivity**, or an **A-IS**.

12.5.1. AI and Alife

Let us approach this "new type" of program obliquely, by way of the two most exciting branches of modern computer programming, **artificial intelligence** and **artificial life**....

In artificial intelligence, first of all, one seeks to write programs that will display the full range of behaviors that humans term "intelligent." There are already programs that display **many** of the behaviors that we call intelligent -- doing arithmetic, algebra and calculus, flying jet planes, playing championship chess, recognizing voices, etc. But these programs are invariably narrow in focus: each one does its **schtick**, and is unable to generalize its intelligence to other contexts. A true artificial intelligence would be able to **learn**, and **learn how to learn**, just like a person. It would not necessarily need to know how to do long division like a pocket calculator -- but it would need to be able to **learn** to do long division, to recognize faces, to play new games,....

In the 1960's and early 1970's, it was widely believed that one could achieve artificial intelligence by programming a sufficiently clever "thought algorithm." Now, however, this is no longer believed to be the case. Today it seems to be a terribly long way from voicerecognition and championship chess to true intelligence. The modern AI community is torn between the "old-fashioned" programming approach and the even older, recently rediscovered "connectionist" approach, which seeks to write programs loosely modeling brain function. Connectionism has succeeded in many instances where old-fashioned programming repeatedly failed. But on the other hand, connectionism seems to be even **less** competent at dealing with logical reasoning and other aspects of linguistic thought.

When one writes a program to imitate the brain, on a coarse or a fine level, one is writing a program that is in a sense **chaotic** and **unpredictable**. One knows what the program does, but not **how** it does it. Thus, the "connectionist" approach to AI has given up on the programme of

first logically understanding an action, **then** writing a program to simulate it, based on this understanding. This unpredictable aspect of connectionist programs leads us to our second type of programming: artificial life, or **Alife**. Alife seeks to take the self-organizing unpredictability of connectionism and apply it to the simulation of **biochemical** or **ecological** rather than neural systems.

For instance, several different groups of researchers have run computer simulations of **self-organizing systems of enzymes**, with an eye toward understanding the dynamics underlying the evolution of life. Bagley and Farmer (1992) have treated the origins of metabolism in this way; whereas Boerlijst and Hogeweg (1992) have modeled the well-known hypercycle theory of the origins of reproduction.

And, on a higher level of organization, various researchers have simulated "artificial ecosystems" from ant farms (Collins and Jefferson, 1992) to systems of coevolving parasites (Hillis, 1992). Richard Dawkins (1986) has investigated "biomorphs," artificial life-like shapes generated by a process of progressive evolution.

12.5.2. A-IS

In AI, one seeks programs that will respond "intelligently" to **our** world. In Alife, one seeks programs that will evolve interestingly within the context of their **simulated** worlds. It is, of course, not difficult to synthesize these two research programmes to obtain the idea of "artificially intelligent artificiallife" -- synthetically evolved life forms which display intelligence with respect to their simulated worlds.

But A-IS, artificial intersubjectivity, constitutes a large step beyond this hybrid concept "AI Alife." What I am suggesting is to simulate a **system of intelligences collectively creating their own "virtual" reality**. The universal network model gives us a blueprint for the joint construction of realities; it remains only to put this blueprint into action by making appropriate computational simplifications.

12.5.2.1. The Nature of Human Intelligence

What would be the **point** of this formidable programming exercise? There are at least **two** good reasons for pursuing A-IS. First of all, consider: what if we humans are only intelligent with respect to **the reality which we have collectively created for ourselves**? Setting aside the unanswerable question of the "ultimate" existence of an objective reality, what if we are only intelligent with regard to the **subjective** reality which we collectively, culturally construct and live within?

This proposition may be taken in two ways. First of all, if one considers intelligence as an optimization problem, as was done in Chapter Three, then the conclusion becomes almost inevitable on an **evolutionary** level. After all, the **general** problems of global optimization and pattern recognition are unsolvable. The human brain consists of some general-purpose optimization routines, plus a whole host of special case tricks tailored to the environment for which it evolved. The **structure** of this network of tricky processes may be universal -- but two entities with the same global structure don't necessarily have the same **abilities**.

But whether or not one accepts this "evolutionary" point, the same conclusion follows even more surely on a **psychological** level. For as we have already shown in Section Two, the **adult human mind** is specifically tailored to its culturally constructed collective reality. We have to **learn to think** -- an infant doesn't know how; and the evidence shows that a child left to mature in isolation will never adequately learn how. We learn to think by practicing on examples that have to do with the self/reality belief system; and this belief system develops properly only in a **social** context, i.e. only in the context of **explicitly creating subjective reality jointly with other minds**.

So, in sum, it seems quite certain that **the process of thinking is inseparable from the process of participating in the collective construction of a reality**. And **this** fact indicates the necessity for a new type of programming, one that might be roughly characterized as "AI + Alife + feedback between the two" -- a community of artificial intelligences, acting in an artificial collective subjective world, and simultaneously acting **on** that world.

One might argue that collective construction of reality is not enough -- that the "adult/child" relationship is necessary for the development of intelligence; that one cannot become a "mental adult" except under the tutelage of another "mental adult." But of course, this is a chicken/egg problem ... who was the first "mental adult"? On the other hand, the idea that collective reality construction is necessary for intelligence presents no chicken/egg problem, since there can quite well have been a first **tribe**, a first **group** of organisms biologically capable of some degree of intelligence.

Perhaps, indeed, a high degree of intelligence requires a few dozen or a few thousand generations of co-creating minds working gradually toward "mental adulthood." But even if this is true (which I rather doubt), it is not a **fundamental** obstacle to the concept of A-IS, of computer-simulated intersubjective reality construction. After all, in Alife one routinely simulates thousands of generations of evolution. In Theodore Sturgeon's classic story "Microcosmic God," a scientist breeds organisms called "Neoterics" which evolve so fast that they zoom beyond mankind in a matter of months. Robert L. Forward's novel *Dragon's Egg* pursues a similar theme, except that the rapidly evolving organisms are not human creations but the natural fauna of a neutron star. With sufficiently fast computers, this science-fictional "souped-up evolution" process could be simulated, allowing artificial intersubjectivity to evolve over numerous generations.

The universal network model gives us a handy, elegant way of achieving this type of **artificially intersubjective** program. Namely, simulate a collection of **dual networks connected at the bottom**. The bottom levels are the collective subjective reality; the upper levels are the individual thought processes of the "intelligences." Under appropriate conditions, the presence of a common subjective reality will cause thevarious networks to "converge" to a common belief system regarding their "external world." This belief system will inevitably include a role for **themselves** -- an "imaginary subject."

12.5.2.2. Dissociation and A-IS

And this leads us to the **second** good reason for pursuing A-IS: only by developing a natural self/reality dynamics can a mind develop dissociated personalities which encourage structural conspiracies. Productive structural conspiracies are necessary for systematic, clever logical thought. Therefore, by creating a **community** of collective-reality-constructing AI agents, we will implicitly be creating AI agents which are **adept and creative at directing their logical reasoning**. This creativity will not necessarily be a clone of human creativity, because the specific belief systems and dissociated subpersonalities may be different. But there is no reason that computer creativity achieved in this way could not equal or exceed human creativity in utility and power.

This is a fundamentally new approach to computer reasoning. Neither connectionism nor old-fashioned rule-based AI comes anywhere near to acknowledging the complex process dynamics of intelligence. Alife and connectionist AI may support various types of self-organization and chaotic dynamics; but only A-IS can fully manifest the systematic self-generation that is **chaotic logic**.

12.5.2.3. The Question of Implementation

From the point of view of current implementation, there are two serious problems with the A-IS idea: memory and speed! It would be possible to run a stripped-down version of A-IS on contemporary massively parallel supercomputers, such as the larger "Connection Machines" manufactured by Thinking Machines, Inc. But although one could surely obtain interesting results in this way, one would not be doing justice to the concept of A-IS. Agents of relatively little intelligence will be able to develop collective reality dynamics of relatively little subtlety.

Each human brain contains maybe 10*11* simple numeric processors. Even with more efficient techniques at our disposal, it seems unlikely that we can get by with onlya few hundred thousand analogous processing units, which is what today's most powerful parallel computers offer. The most promising path toward developing true A-IS, I suspect, is nanotechnology (Drexler, 1986), or **molecular computing**. Using molecular computing techniques, it may be possible, in the not-too-distant future, to **grow** computers, to create computers which add onto themselves like crystals do. If this possibility should be actualized, then it will not be too long before A-IS becomes a practical science.

AFTERWORD

We are surrounded by complex systems; they touch every aspect of our lives. Our bodies, minds, and environments are all incredibly, perhaps incomprehensibly, complex. And yet, until very recently, there has never been anything close to a **science** of complex systems.

Mainstream "simple-systems" science can give us dazzling details about the structure and function of our cells, molecules and atoms; and it can explain for us the flickerings and motions of objects so distant that it would take millions of years to reach them. It can help us cure diseases, and instruct us how to build computers, bridges, cars, airplanes, houses, nuclear weapons, precision surgical tools, et cetera et cetera.

But virtually all of these achievements were arrived at by the same "meta-method": study a complex phenomenon by

1) breaking it down into its component parts

2) studying the component parts

3) using information about the component parts to obtain information about the whole.

This method, often called "reductionism", does not seem to work very well for studying complex, self-organizing phenomena. It would seem that something beyond reductionism is needed, some new methodology better suited to complex systems.

This observation was the raison d'etre of the mid-century cybernetics/ general systems theory movement. And it is the focal point of an increasing amount of contemporary research: in physics, in biology, in computer science, in psychology, in chemistry,.... We have no completely general theory of complex system dynamics, but we have a wealth of interesting details and moderately general insights. The theory of chaotic dynamical systems hasgiven us a fairly good understanding of phenomena like weather, heartbeats, and smell. By putting together neural network theory, dynamical systems theory and information theory, we can begin to understand significant aspects of the mind and brain. By synthesizing insights from mathematics, biology and physics, we can begin to understand biological evolution.

My goal in writing this book was see whether, by combining current ideas regarding complex system dynamics with the **pattern-theoretic psychology** developed in my earlier books, it might not be possible to work out a **dynamics of mind**. This is, everyone will agree, a task at which reductionist science has utterly failed.

We began, if you recall, with four "intuitive equations":

Linguistic system = syntactic system + semantic system

Belief system = linguistic system + self-generating system

Mind = dual network + belief systems

Reality = minds + shared belief system

Now we are in a position to understand how much, and how little, these system-defining equations reveal. The cognitive equation gives the **flow** of mind, and these equations describe attractors which direct this flow. To take the "flow" metaphor one step further, the system-defining equations are something like complexly-contoured continents, guiding the flow of the vast chaotic ocean that is pattern space. But yet they are not **quite** like continents, because they are themselves formed from the flow of the ocean itself.

As emphasized throughout, all this is only a beginning. We have considered a decent number of concrete examples -- but not enough. The abstract ideas given here must be fleshed out by further contact with the nitty-gritty details of real languages, real trains of thought, real cultures, real belief systems, real personalities, real subjective realities.

However, I do feel that some genuine insight has been gained. Previously uncharted regions have been tentatively explored. The first few steps have been taken toward understanding that most mysterious and most essential process by which **logic** interfaces with **self-organizing habit** ... by which order synergizes with chaos to form the complex patterns of becoming that we call -- **mind**.

REFERENCES

Abraham, Fred, Ralph Abraham and Chris Shaw (1991). A Visual Introduc tion to Dynamical Systems Theory for Psychology, Aerial Press, Santa Cruz CA

Aczel, P. (1988). Non-Well-Founded Sets, CSLI Lecture Notes, Palo Alto

Albeverio, S. and R. Hoegh-Krohn (1976). Mathematical Theory of Feynman Path Integrals, Springer-Verlag, New York

Albus, J.S. (1991). "Outline for a Theory of Intelligence," IEEE-TSMC

Arbib, Michael, and Mary Hesse (1986). The Construction of Reality, Cambridge University Press, Cambridge

Ashby, Ross (1954). Design for a Brain, Chapman and Hall, London

Atlan, Henri (1988). "Measures of Biologically Meaningful Complexity," in Measures of Complexity, ed. Peliti et al, Springer-Verlag NY

Au, T.K.F. (1983). "Chinese and English Counterfactuals -- the Sapir-Whorf Hypothesis Revisited." Cognition 15, pp. 155-87

Avron, Arnon (1990), "Relevant Logic and Paraconsistency -- A New Approach," Journal of Symbolic Logic 55-2, 707-732

Bailin, D. and A. Love (1986). Introduction to Gauge Field Theory, Adam Hilger, Boston

Bagley, R., D. Farmer and W. Fontana, "Spontaneous Emergence of a Metabolism," in (Langton et al, 1992).

Barnsley. Michael. (1989). Fractals Everywhere, Academic, Boston

Barwise, Jon (1989). The Situation in Logic, CLSI Lecture Notes No. 17, CLSI Publications, Stanford

Barwise, Jon and John Etchemendy (1988). The Liar

Barwise, Jon and Larry Moss (1991). "Hypersets," Mathematical Intelligencer 13-4, 31-44

Barwise, Jon and J. Perry (1981). Situations and Attitudes, MIT Press, Cambridge MA

Bates, E. and B. MacWhinney (1987). "Competition, Variation and Language Learning," in Mechanisms of Language Acquisition, Ed. B. MacWhinney, Erlbaum, Hilldale NJ

Bateson, Gregory (1980). Mind and Nature: A Necessary Unity. Bantam, NY

Bennett, C.H. (1982) "The Thermodynamics of Computation -- A Review," International Journal of Theoretical Physics, Vol. 21, p. 905-940

Blakeslee, Sandra (1991). "Brain Yields New Clues on its Organization for Language," New York Times, Sept. A, p. C1

Bloom, Alfred (1981). The Linguistic Shaping of Thought: A Study in the Impact of Language on Thinking in China and the West, Erlbaum, Hilldale NJ

Boerlijst, M. and P. Hogeweg (1992). "Self-Structuring and Evolution: Spiral Waves as a Substrate for Pre-biotic Evolution," in (Langton et al, 1992).

Brock, W.A. (1991). Nonlinear Dynamics, Chaos and Instability, MIT Press 1991

Brown, Jason (1988). The Life of the Mind, Erlbaum, Hilldale NJ

Bundy, Alan (1991). "A Science of Reasoning," in Computational Logic, Edited by Lasser and Plotkin, MIT Press, Cambridge MA

Burnet, C. MacFarlane (1976). A Homeostatic and Self-Monitoring Immune System. In Immunology, Ed. by C.M. Burnet. SanFrancisco: Freeman.

Callero, Peter (1991). "Toward a Sociology of Cognition," in The Self-Society Dynamic, Edited by J. Howard and P. Callero, Cambridge University Press, New York

Chaitin, Gregory (1974). "Information-Theoretic Computational Complexity," IEEE Transactions on Information Theory, IT-20, pp. A-15

Chaitin, Gregory (1978). "Toward a Mathematical Definition of Life," in The Maximum Entropy Principle, ed. Levine and Tribus, MIT Press, Cambridge MA

Chaitin, Gregory (1987). Algorithmic Information Theory, Cambridge Press, NY

Changeux, Jean-Pierre (1985). Neuronal Man, Pantheon, NY

Chomsky (1956). "Three Models for the Description of Language," IRE Trans. Info. Theory 2, p. 113

Collins, R. and D. Jefferson (1992). "AntFarm: Towards Simulated Evolution," in (Langton et al, 1992).

Colwell, Gary (1989). "God, the Bible and Circularity," Informal Logic XI.2, p. 61-73

Cooper, David (1973). Philosophy and the Nature of Language, Longman, London

Csanyi, Vilmos (1989). Evolutionary Systems: A General Theory, Duke University Press, Durham NC

da Costa (1984). "On the Theory of Inconsistent Formal Systems," Notre Dame Journal of Formal Logic, pp. 497-5A

deBoer and Perelson (1990). "Size and Connectivity as Emergent Properties of a Developing Network." Preprint.

Dawkins, Richard (1986), The Extended Phenotype, Freeman 1982

Dennett, Daniel (1991). Consciousness Explained, Little, Brown and Co., Boston

Deutsch, D. (1985). "Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer," Proc. R. Soc. Lond. A 400, 97-117

Devaney, Robert (1986). An Introduction to Chaotic Dynamical Systems, Benjamin/Cummings, Menlo Park CA

Devlin, Keith (1991). Logic and Information

Dewart, Leslie (1989). Evolution and Consciousness, University of Toronto Press, Toronto

Dowty, Wall and Peters (1981): Introduction to Montague Semantics, Reidel, Boston

Drexler, Eric (1986). The Engines of Creation: The Coming Era of Nanotechnology, Doubleday NY

Dreyfus, Hubert (1992). What Computers Still Can't Do, MIT Press, Cambridge MA

Edgar, P. (1990). Measure, Topology and Fractal Geometry, Springer-Verlag, NY

Edelman, Gerald (1987). Neural Darwinism. Basic Books, NY

Edelman, Gerald (1989). The Remembered Present: A Biological Theory of Consciousness. Basic Books, NY.

Ellenberger, H.F. (1970). The Discovery of the Unconscious, Basic, New York

Etzione (1968). The Active Society, The Free Press, NY

Fadeev, L.D. (1969). "The Feynman Integral for Singular Lagrangians," Theor. Math. Phys. 1, 1-13

Feyerabend, Paul (1970). Against Method, University of Minneapolis Press, Minneapolis

Feynman, R.P. (1948). "Spacetime Approach to Non-relativistic Quantum Mechanics," Rev. Mod. Phys. 20, 367-387

Feynman, R.P. (1950). "Mathematical Formulation of the Quantum Theory of Electromagnetic Interaction," Phys. Rev. 80, 440-457

Fodor (1987). Psychosemantics, MIT Press, Cambridge MA

Ford, J. (1985). "How Random Is a Coin Toss?" Physics Today

Fredkin, E. and T. Toffoli (1982). "Conservative Logic," International Journal of Theoretical Physics 21

Freeman, W. (1991). "The Physics of Perception," Scientific American, pp. 34- 41

Frege, Gottlob (1893/1952): On Sense and Reference, in Translations from the Philosophical Writings of Gottlob Frege, ed. by P.T. Geach and M. Black, Basil Blackwell Oxford, pp. 56-78

Frege, Gottlob (1884/1959). The Foundations of Arithmetic, Basil Blackwell, London

Goertzel, Ben (1989). "A Multilevel Approach to Global Optimization," Ph.D. thesis, Temple University, Phila. PA. To appear in Journal of Optimization Theory and Applications.

Goertzel, Ben (1990). "Information, Complexity and Meaning". Talk delivered at Irvine AMS Conference, Nov. 1990

Goertzel, Ben (1992a). "Measuring Static Complexity," Journal of Mathema tics and Mathematical Sciences

Goertzel, Ben (1992b). "Self-Organizing Evolution," Journal of Social and Biological Structures

Goertzel, Ben (1992c). "Structural Complexity of Sequences, Images and Automata," to appear in Proceedings of International Conference on Finite Fields, Coding Theory, and Advances in Communication and Computing.

Goertzel, Ben (1993). The Structure of Intelligence, Springer-Verlag

Goertzel, Ben (1993a). The Evolving Mind, Gordon and Breach

Goertzel, Ben (1993b). "A Cognitive Law of Motion," in Proceedings of the First, Second and Third Annual Conferences of the Society for Chaos Theory in Psychology, Ed. Robin Robertson, Gordon and Breach, New York

Goertzel, Ben (1993c). "Belief Systems as Attractors," in Proceedings of the First, Second and Third Annual Conferences of the Society for Chaos Theory in Psychology, Ed. Robin Robertson, Gordon and Breach, New York

Goertzel, Ben (1993d). "Evolutionary Dynamics in Minds and Immune Systems," in Chaos in Psychology, Ed. Abraham and Gilgen, Greenwood, New York

Goertzel, Ted (1993). Turncoats and True Believers, Prometheus Press

Goffman, E. (1959). The Presentation of Self in Everyday Life, Doubleday, New York

Goffman, E. (1961). Asylums, Doubleday, New York

Goldberg, D. (1980). Genetic Algorithms for Search, Optimization and Machine Learning, Addison-Wesley, New York

Goodwin, Richard (1990). Chaotic Economic Dynamics, Oxford University Press, New York

Green M., J. Schwartz and E. Witten (1987). Superstring Theory, v. 1-2, Cambridge, 1987

Hacking, Ian (1991). "Speculation, Calculation and the Creation of Phenomena," in Beyond Reason, Edited by G. Munevar, Kluwer Academic, New York

Haken, Hermann (1983). Synergetics: An Introduction, Springer-Verlag, NY

Hartmann, Ernest (1991). Boundaries in the Mind, Basic NY

Hillis, W.D. (1992). "Co-Evolving Parasites Improve Simulated Evolution as an Optimization Procedure," in (Langton et al, 1992).

Hofstadter, Douglas (1985). Metamagical Themas, Basic, New York

Hoijer, H. (1953). "The Relation of Language to Culture." In Anthropology Today, Edited by A.L. Kroeber, University of Chicago Press, Chicago, pp. 554-73

Holland, J.H. (1975) Adaptation in Natural and Artificial Systems, Ann Arbor: University of Michigan Press

Howard, J.A. and Peter Callero, Editors (1991). The Self-Society Dynamic, Cambridge University Press, New York

Hoyle, F. and J. Narlikar (1974). Action at a Distance in Physics and Cosmology, Freeman, New York, p. 166-173

Hubel, David (1988). Eye, Brain and Vision, Freeman, NY

Ito, K. (1967). "Generalized Uniform Complex Measures in the Hilbertian Metric Space with Their Application to the Feynman Path Integral," Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, U. California Press, Berkeley

Jackendoff, R. (1987). Consciousness and the Computational Mind, MIT Press, Cambridge MA

Janet, P. (1904). "Un Amnesie et la Dissociation des Souvenirs par l'Emo tion," Journal de Psychologie 1, 417-453

Jantsch, Erich (1980). The Self-Organizing Universe, Pergamon, NY

Jerne, Niels Kaj (1973). The Immune System. Scientific American

Jung, Carl G. (1936). The Archetypes and the Collective Unconscious

Jung, Carl G. (1952). Synchronicity: An Acausal Connecting Principle

Kampis, George (1991). Self-Modifying Systems in Biology and Cognitive Science, Pergamon, New York

Kanerva, Pentti (1988). Sparse Distributed Memory, MIT Press, Cambridge MA

Kandel, A. (1986). Fuzzy Mathematical Techniques with Applications, Addison-Wesley, New York

Kauffmann, Louis (1976). Sign and Space, privately distributed manuscript

Kauffmann, Stuart (1993). The Origins of Order

Kohler, Wolfgang (1975). Gestalt Psychology, New American Library, NY

Kohonen, Teuvo (1984). Self-Organization and Associative Memory,

Springer-Verlag NY

Kolmogorov, A.N (1965) "Three Approaches to the Quantitative Definition of Information", Prob. Information Transmission 1, pp. 1-7

Koppel, Moshe (1987) "Complexity, Depth and Sophistication." Complex Systems 1, pp. A87-A91

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

Laing, R.D. (1972). Knots, Pantheon, NY

Lakatos, Imre (1978). Philosophical Papers, Cambridge University Press, Cambridge England

Lakoff, George (1987). Women, Fire and Dangerous Things. University of Chicago Press, Chicago

Langton, Chris, C. Taylor, J.D. Farmer, and S. Rasmussen (Editors) (1992). Artificial Life, Addison-Wesley, New York

Lima de Faria, A. (1988). Evolution Without Selection, Elsevier, New York

Lofgren, L. (1968). "An Axiomatic Explanation of Complete Self-Reproduc tion," Bull. Math. Biophysics 30, 415-425

London and Bauer (1983). In Quantum Theory and Measurement, ed. Wheeler and Zurek, Princeton University Press, Princeton, pp. 217-259

Lovelock, James (1988). The Ages of Gaia. New York: Bantam.

Lucy, John Arthur (1987). "Grammatical Categories and Cognitive Processes: An Historical, Theoretical and Empirical Reevaluation of the Linguistic Relativity Hypothesis." Dissertation, University of Chicago.

MacNamara (1986). A Border Dispute: The Place of Logic in

Psychology, MIT Press, Cambridge MA

Mandler (1985). Cognitive Psychology: An Essay in Cognitive Science, Erlbaum Press, Hillsdale NJ

Margolus, H. (1987). Patterns, Thinking and Cognition, University of Chicago Press, Chicago

Marks, Terry (1992). "Chaos and Impulsivity," preprint

McKellar, Peter (1979). Mindsplit, Dent, London

Mill, J.S. (1843). System of Logic, Longman's, London

Minsky, Marvin (1986). The Society of Mind, MIT Press, Cambridge MA

Mitchell, D., E. Osborne and M. O'Boyle (1985). "Habituation Under Stress," Behav. Neural Biol. 43, 212-217

Modell, A. (1990). Other Times, Other Realities, Harvard University Press, Cambridge MA

Montague, Richard (1974): Universal Grammar, in Formal Philosophy: Selected Papers of Richard Montague, ed. by R.H. Thomason, Yale University Press, New Haven, pp. 222-246

Morette-deWitt (1974). "Feynman Path Integrals, I. Linear and affine techniques, II. The Feynman-Green Function," Commun. Math. Phys. 37, 63-81

Nietszche, Friedrich (1968). Beyond Good and Evil, in Basic Writings of Nietszche, tr. Walter Kauffman, Random House, NY, p. 216

Nietszche, Friedrich (1968a). The Will to Power, tr. Walter Kauffman, Random House, NY

Partee, Barbara H. (1975): Montague Grammar and Transformational Grammar, Linguistic Inquiry 6, pp. 203-300

Paulhan, J. (1887). Laws of Feeling

Paz, Octavio (1984): Selected Poems, New Directions, NY, p. 54

Peirce, Charles Sanders Santiago (1935). Collected Works of Charles S. Peirce, Vol. 5, Harvard University Press, Cambridge, p. 139

Peirce, Charles Sanders Santiago (1966). Selected Writings, Dover, New York

Penrose, Roger (1988). The Emperor's New Mind

Perelson, Alan (1989). "Immune Network Theory." Immunological Reviews 1A, 5-36.

Perelson, Alan (1990). "Localized Memory in Idiotypic Networks." J. Theoretical Biology 146, 483-499.

Pinker, Steven (1987). "The Bootstrapping Problem in Language Acquisition," in Mechanisms of Language Acquisition, Ed. B. MacWhinney, Erlbaum, Hilldale NJ

Poundstone, W. (1988). Labyrinths of Reason, Doubleday NY

Prigogine, Ilya and I. Stengers (1984). Order Out of Chaos, Bantam

Ramond, P. Field Theory: A Modern Primer, Benjamin/Cummings, Reading

Read, Stephen (1988). Relevant Logic, Basil Blackwell Inc., New York

Rivers, R. (1987). Path Integral Methods in Quantum Field Theory, Cambridge University Press, Cambridge 1987

Rock, I. (1988). The Logic of Perception, MIT Press, Cambridge MA

Rosen, R. (1991). Life Itself, Columbia University Press, NY

Rosenfield, Israel. (1989). The Invention of Memory, Knopf, NY

Rosenfield, Israel. (1992). Strange, Familiar and Forgotten, Knopf 1992

Rota, G.-C. (1985). "The Barrier of Meaning," Letters in Mathematical Physics A, pp. 97-115

Rumelhart, D.E., McLelland, J.L. (1986) and the PDP Research Group. Parallel Distributed Processing. Cambridge MA: MIT Press

Ruse and Dubose, ed.'s (1985). Models of the Visual Cortex, Wiley, NY

Searle, J. (1983). Intentionality: An Essay in the Philosophy of Mind. Cambridge University Press, Cambridge

Schultz (1990). Dialogue at the Margins, University of Wisconsin Press, Madison

Sheldrake, Rupert (1987). A New Science of Life, Blond and Mercer, London

Shimony, A. (1986). In Quantum Concepts in Space and Time, edited by R. Penrose and C. J. Isham, Clarendon, Oxford

Simon, B. (1979). Functional Integration and Quantum Physics, Academic Press, New York

Smith-Lovin, Lynn (1991). "An Affect-Control View of Cognition and Emotion," in The Self-Society Dynamic, Edited by J. Howard and P. Callero, Cambridge University Press, New York

Solomonoff, L. (1964) "A Formal Theory of Induction, Parts I. and II.," Information and Control 7, pp. 1-22 and 224-254

Spencer-Brown, G. (1972), Only Two Can Play This Game, Julian Press NY

Spencer-Brown, G. (1972). Laws of Form, Julian Press, NY

Stewart, John (1992). Informal talk given at MSRI Workshop on Mathema tical Physiology, Berkeley CA, August 1992

Tarski, Alfred (1935): The Concept of Truth in Formalized Languages, in Logic, Semantics, Metamathematics, ed. by A. Tarski, Clarendon, Oxford, pp. 152-278

van der Kolk, B. and van der Hart, O. (1991). "The Intrusive Past: the Flexibility of Memory and the Engraving of Trauma," American Imago 48, No. 4, pp. 425-454

van Gelder, Tim (1990). "Compositionality: A Connectionist Variation on a Classical Theme," Cognitive Science 14, pp. 355-384

Varela, Francisco (1978). Principles of Biological Autonomy, North-Holland NY

Vesey, Godfrey (1991). Inner and Outer, St. Martin's Press, New York

Wheeler, John (1979). "Frontiers of Time," in Problems in the Foundations of Physics, Edited by T. di Francia, North-Holland 1979

Whiting, H. (1984). Human Motor Actions: Bernstein Reassessed, North-Holland, Amsterdam

Whorf, Benjamin (1956). Language, Thought and Reality, MIT Press, Cambridge MA

Wolfram, Stephen (1986). Cellular Automata: Theory and Applications, World Scientific, Singapore

Wright (1989). Three Scientists and Their Gods: Looking for Meaning in an Age of Information, Harper and Row, New York