FROM COMPLEXITY TO CREATIVITY

Emergent Patterns and Self-Organizing Dynamics in Mental, Computational and Physical Systems

By Ben Goertzel

FROM COMPLEXITY TO CREATIVITY

Emergent Patterns and Self-Organizing Dynamics

in Mental, Computational and Physical Systems

Ben Goertzel

Psychology Department

University of Western Australia

Nedlands WA 6009

Australia

Paper Version to be published by Plenum Press, 1996

CONTENTS

"A book for thinking --

and nothing besides"

Friedrich Nietszche,

preface to his never-written

book, *The Will to Power*

ACKNOWLEDGEMENTS

I owe particular thanks to those who assisted me, in one way or another, in the development of the ideas reported here. Far more than any of my previous books, this one has been a cooperative venture.

Of course, as the saying goes, any errors that remain in the book are my responsibility alone.

In no particular order, I must acknowledge:

Kent Palmer -- although I know him only by e-mail, he has probably given me much more interesting feedback on my work than anyone else. The work described here on magician systems, abstract algebras and hypercomplex fractals was largely inspired by my dialogue with Kent.

Tony Smith (of Georgia Tech) and Onar Aam, for numerous intriguing e-mail discussions, in a four-way dialogue with myself and Kent that has continued through the second half of 1995, and now, mid-way through 1996, is still remarkably active. The theory of consciousness given here

has benefitted particularly from their input: the basic idea of octonionic consciousness was given to me by Onar and Kent.

John Pritchard, a Columbia student and another e-mail acquaintance. His enthusiasm for and feedback on my book *Chaotic Logic* has been inspiring and educational; and his help in rethinking the Chaos Language Algorithm has been deeply appreciated.

Allan Combs, who has been extremely supportive of my forays into system theoretic psychology. His dual expertise in neuropsychology and transpersonal psychology has made him an excellent resource. In addition to the collaborative work on mood fluctuations reported here, his application of my "cognitive equation" to the dynamics of states of consciousness, as reported in his book *Radiance of Being*, has been most inspirational.

Matt Ikle' (now at the University of New Mexico, Alamosa), who collaborated on the infinite population size GA work reported here (contributing many thoughtful ideas as well as lots of Mathematica hacking), and is my current collaborator on the SEE model.

Malwane Ananda and Louis Yager, who helped out with the infinite population model in its early stages; and Gennady Bachman, who came up with the proof of the determinant theorem. Harold Bowman, who, as well as being a great friend, helped me to work through some of the mathematical details of the generalized Baker map and random magician networks. The magician system model itself was first worked out in detail by Harold and myself, in a series of early morning sessions in my Las Vegas townhouse; this was also when the connection between magician systems and abstract algebras was first conceived.

Three graduate students at UNLV. Hiroo Miyamoto and Yoshimasa Awata, who wrote the GA-driven fractal inverse problem program, as part of Hiroo's M.S. thesis. Andy Bagdunov, a computer science grad student at UNLV, who wrote a program which I used to evolve melodies by the genetic algorithm, at an early stage of my research on computer music (I never actually used his program, but the routines that he wrote for producing MIDI files have proved very useful).

George Christos, of the Curtin University math department, in Perth -- first, for sending me his papers on dream dynamics prior to my arrival in Australia; and second, for many intriguing discussions on dreaming, consciousness, neural networks, memory and related topics since my arrival in Perth.

The folks on *chaopsyc*, the Internet listserver of the Society for Chaos Theory in Psychology. This listserver has often been interesting and has, on several occasions, even been useful. Numerous sections of this book existed in their first incarnations as posts to *chaopsyc*.

The Computer Science Department of Waikato University, where much of this book was written, and a good bit of the work described in it was done. Although I only stayed there fourteen months (1994, and a bit of 1995), it was an extremely productive period for me, and I am grateful to have had the opportunity to work in such a remarkably friendly environment.

The Psychology Department of the University of Western Australia, where I am currently employed. The environment here is truly intellectually stimulating, much more so than any of my previous workplaces. In particular, Mark Randell and Mike Kalish have helped to keep alive my interest in psychological applications of complexity science, and have helped me to nurse along a few ideas, e.g. "mindspace curvature."

My sons Zarathustra and Zebulon. Zebbie has learned to talk during the two-year-long genesis of this book; when I first drafted this Acknowledgements section, one of his newest phrases was "Ben, no work!" He and his big brother Zar have held up extremely well as I have moved them from country to country over the past few years.

And last but not least, my wife Gwen, who deserves thanks both professionally (for her help in formulating the Chaos Language algorithm), and personally.

INTRODUCTION

Cybernetic pioneer Warren McCullough asked: "What is a man, that he may know a number; and what is a number, that a man may know it?" Thinking along much the same lines, my question here is: "What is a creative mind, that it might emerge from a complex system; and what is a complex system, that it might give rise to a creative mind?"

Complexity science is a fashionable topic these days. My perspective on complexity, however, is a somewhat unusual one:

I am interested in complex systems science principally as it reflects on abstract mathematical, computational models of **mind**. In my three previous books, *The Structure of Intelligence*, *The Evolving Mind*, and *Chaotic Logic*, I have outlined a comprehensive complex-systems-theoretic theory of mind that I now call the **psynet model**.

This book is a continuation of the research programme presented in my previous books (and those books will be frequently referred to here, by the nicknames *SI*, *EM* and *CL*). One might summarize the trajectory of thought spanning these four books as follows. *SI* formulated a philosophy and mathematics of mind, based on theoretical computer science and the concept of "pattern." *EM* analyzed the theory of evolution by natural selection in similar terms, and used this computational theory of evolution to establish the evolutionary nature of thought. CL deviated much further from the main stream of contemporary science, and presented the "cognitive equation" -- an original mathematical formalism expressing the structure and dynamics of mind -- with applications to logic, semantics, consciousness, personality, belief systems, and the philosophy of reality. Now, the present book takes the striking, unorthodox models presented in *CL*, and seeks to draw them back into the fabric of current science, by

applying them to various real-world problems, and connecting them to other theories of complex system behavior.

The synthetic model of mind presented in *SI*, *EM* and *CL*, culminating in the cognitive equation, is given the name *psynet model*. The psynet model is presented in a new and hopefully more clear way, and the connections between the psynet model and other approaches to complex cognitive systems are drawn out in detail. The table of contents is broad, voyaging through dynamical systems, genetic algorithms, perceptual illusions, fractals, autopoietic systems, consciousness, time series analysis, dreams, personality theory, the World Wide Web, and creativity. But even this broad array of topics barely touches the possible applicability of the *psynet* perspective. The point is not to say the last word on any particular topic, but rather to sketch out a general point of view, which has detailed points to make about every aspect of the mind, and has many points to make about the more complex aspects of the non-mental world as well.

The diverse interconnections presented here teach us something about the psynet model, they teach us something about the mental and physical worlds, and, last but not least, they also teach us something about the other complex systems models that are discussed. It is very interesting to see how standard complex systems models must be **extended** in order to deal with the immense complexity of the mind. For instance, the theory of polynomial iterations must be extended to hypercomplex numbers, rather than merely complex numbers. Genetic algorithms must be extended to incorporate ecology and spatial distribution. Attractor theory must be extended to the study of emergent formal languages in trajectories. Neural network theories must be made to shift their emphasis to the structure of interconnection between neuronal groups or modules.

Some of the explorations presented here are fairly technical; others are almost entirely nontechnical. Some report concrete scientific research; others are more speculative. What ties the various chapters together, however, is the focus on the interface of complexity and mind. The relation of mind and complexity is a big question, and I certainly do not pretend to have fully resolved it. But I believe I have made some progress.

In order to guide the reader who may have fairly specific interests, I have divided the book into four parts. This division is only a rough one, but it serves to break up the journey from simple mathematical models to subtle human feelings into comprehensible segments.

Part I., *The Complex Mind-Brain*, outlines the psynet model and gives some related ideas that lie fairly close to the model itself -- the application of the model to brain modeling, and the relation of the model to theories of perception. Chapter 2, in particular, gives the conceptual framework for the remainder of the book.

Part II., *Formal Tools for Exploring Complexity*, is more technical and the nonmathematical reader might be wisest just to skim it over. It reviews ideas from dynamical systems theory, genetic algorithms, and abstract algebra, and shows how these ideas can be extended beyond what is usually done, to provide tools for thinking about and analyzing the mind. This material provides the conceptual and scientific inspiration for the ideas in the following Parts.

Part III., *Mathematical Structures in the Mind*, gives a series of loosely related applications of the ideas of Parts I and II to various psychological phenomena: consciousness, dreaming, language production, self-formation, and even the possibility of intelligence on the World Wide Web.

Finally, Part IV., *The Dynamics of Self and Creativity*, leaves the mathematical precision of Part II even further behind, and attempts to deal with the stickier problems of personality psychology. What do these complex systems models tell us about why people act the way they do? The culmination is the final chapter, which combines personality-psychology ideas with complex-systems ideas to arrive at a novel, synthetic theory of creativity.

Synopsis

I will now give a more detailed summary of the contents of the individual chapters.

Chapter One reviews some ideas regarding dynamical systems, genetic algorithms and autopoietic systems, which will be useful in following chapters.

Then, Chapter Two reviews the psynet model, the cornerstone of most of the following chapters. As compared to previous publications, the model is given a very dynamical twist -- it is interpreted to portray the mind as a collection of interacting, intercreating pattern-recognition and pattern-creation processes, residing on and dynamically creating a "mental graph." Mental entities such as thoughts, feelings, perceptions and actions are viewed as attractors of spatiotemporal pattern/process dynamics. The essence of mental structure is an high-level emergent meta-attractor called the "dual network," which consists of synergetic hierarchical and heterarchical structures.

Chapter Three gives a newly detailed psynet-theoretic model of the brain. Loose connections between the psynet model, the cell assembly theory of mind/brain, and Neural Darwinism have been made before. Here these general connections are used to formulate a specific psynet/cortex correspondence, in which the two aspects of the dual network are mapped onto the two orthogonal structures of the cortex (pyramidal neurons and cortical layers).

Chapter Four introduces a principle from perceptual psychology, form-enhancing distance distortion or "mindspace curvature." This principle emerges from the study of simple geometric illusions, but it has implications beyond vision. It is used to solve an outstanding problem within the psynet model as previously formulated, namely the initial impetus for the **formation** of the dual network.

Chapter Five -- beginning Part II -- turns to **pattern** -- a much-undervalued concept which is central to the psynet model. It is shown that the theory of algorithmic pattern allows one to give a complete formalization of complex systems science, by defining such key terms as system complexity and emergence. Then attention is turned to a new tool for recognizing patterns in dynamial data, the Chaos Language Algorithm or CLA. The CLA, it is argued, indicates how the psynet model can eventually be put to empirical test. An exploratory application of the CLA to data on mood fluctuations is reported.

In Chapter Six, the relation between the psynet model and genetic algorithms is elucidated. Some mathematical results about the dynamics of the GA are summarized, and it is argued that the behavior of the GA with crossover is more "mind-like" than the behavior of the GA with mutation only. Finally, it is shown that a spatially distributed genetic algorithm, with ecology included, can serve as a simple implementation of the psynet model. The genetic algorithm is viewed as an abstract "ideal form" of certain types of mental creativity.

Chapter Seven is an extended mathematical improvisation on the theme of "magician systems." Magician systems, collections of entities that collectively transform and annihilate each other, are central to the rigorous formulation of the psynet model. However, they have been studied very little, and it is not clear where they fit in along the spectrum of applied mathematics concepts. Here magician systems are formulated in terms of directed hypergraphs, and then in terms of hypercomplex algebras and yet more abstract, three-operation algebras. Magician system dynamics (and as a consequence, psynet dynamics) is shown to have to do with polynomial and rational iterations on these n-dimensional algebraic spaces.

Chapter Eight, starting off Part III, leaves mathematics behind and turn to a psychological problem: the question of consciousness. The nature of "raw consciousness" or raw experience is not entered into; the focus is rather on how this raw consciousness is elaborated into structured, subjective states of mind. The first half of the chapter is fairly non-adventurous, and merely expands upon the concept of the "perceptual-cognitive loop" as introduced in *CL*. The second half of the chapter is the most speculative part of the book; it presents a detailed algebraic theory of states of consciousness, based on the magician system algebras of Chapter Seven and the quaternionic and octonionic algebras in particular.

Chapters Nine and Ten apply the psynet model, and complex systems ideas in general, to two specific psychological problems: the nature of sentence production, and the purpose of dreams. First, sentence production is viewed as a process of fractal growth, similar to biological ramification. It is modeled in terms of L-systems, and evidence for this model is sought in the structure of child language.

Then, dreaming is considered in the context of the Crick-Mitchison hypothesis, which states that "the purpose of dreams is to forget." Hopfield net simulations of this hypothesis are considered, and then it is asked: how might similar phenomena be gotten to arise from the psynet model? The answer leads one to a theory of dreams that is surprising similar to conventional psychoanalytic theories. Dreaming does not simply help one forget, it helps one loosen the grip of overly dominant autopoietic thought-systems.

Chapter Eleven turns toward a topic that will preoccupy much of the remainder of the book: the **self.** It points out the importance of the psychosocial self for knowledge representation, and argues that until artificial intelligence embraces artificial selfhood, it is bound to failure. The emergence of the self from the dual network is discussed; and the notion of A-IS, or artificial intersubjectivity in artificial life worlds, is discussed more thoroughly than in *CL*.

Elaborating on this train of thought, Chapter Twelve raises the question of the psychology of the World Wide Web. As the Web becomes more intelligent, and becomes populated with intelligent

agents, might it someday become a mind? Might it develop self-awareness? The possibility of the Web developing a magician-system/dual-network structure, as described by the psynet model, is briefly discussed.

Finally, the last three chapters, constituting Part IV, turn to the difficult and imprecise, but very rewarding, topic of human personality. Chapter Thirteen reviews the notion of the **dissociated** self, and argues that a self is in fact a multi-part dynamical system. The essence of human personality, it is argued, lies in the dynamics of various subselves. Martin Buber's distinction between I-You and I-It interactions is reformulated in terms of emergent pattern recognition, and it is argued that healthy personalities tend to display I-You interactions between their various subselves.

In Chapter Fourteen, applications to the theory of romantic love and the theory of masochism are outlined. These applications are sketchy and suggestive -- they are not intended as complete psychoanalytic theories. The point, however, is to indicate how ideas from complexity science, as represented in the psynet model, can be seen to manifest themselves in everyday psychological phenomena. The same dynamics and emergent phenomena that are seen in simple physical systems, are seen at the other end of the spectrum, in human thought-feeling-emotion complexes.

Finally, in Chapter Fifteen, the nature of **creativity** is analyzed, in a way that incorporates the insights of all the previous chapters. The theory applies to either human or computer creativity (although, up to this point of history, no computer program has ever manifested even a moderate degree of creativity, as compared to the human mind). The existence of a dissociated "creative subself" in highly creative people" is posited, and the dynamics of this creative subself is modeled in terms of the psynet model and the genetic algorithm. The experience of "divine inspiration" often associated with creativity is understood as a result of a consciousness-producing "perceptual-cognitive loop" forming part of a greater emergent pattern. In general, previous complex systems models are seen as manifestations of particular aspects of the complex creative process. Creativity, the wellspring of complexity science and all science, is seen to require all of complexity science, and more, for its elucidation.

The Character of the Book

This book is written -- perhaps optimistically -- with two different audiences in mind. The first is the small community of researchers who are themselves involved in exploring the relations between complexity science and the mind. And the second is a much larger group, consisting of individuals with some knowledge of mathematics and computing, who have an interest in the mind, and in one or more of the topics touched on by complexity science, but who have not studied cognitive science or complexity science in any detail. The categories that come to mind here are psychologists, computer scientists, engineers, physicists, biologists, sociologists, mathematicians and philosophers -- but the list need not end there. For instance, the material on evolving computer art and music might be of particular interest to artists and musicians with a technical bent.

It bears repeating that what I provide here is not a systematic treatise, but rather a network of interconnected theoretical, computational and mathematical investigations. Like Nietzsche, I tend to believe that "the will to a system is a lack of integrity." The various chapters do build on each other, and they move in a common direction toward an understanding of creativity as a complex process. However, the path that they follow is not a straight one, and there are numerous ideosyncratic digressions along the way. What I present here are, in essence, structured improvisations on the themes of the psynet model, creativity and complexity science.

The technical level of the book is somewhat lopsided. The chapters on Dynamics and Pattern, Evolutionary Dynamics and Magician Systems are relatively technical, whereas the last chapters on personality dynamics are almost entirely nontechnical, and could probably be understood by the general reader. Of course, I would prefer everyone to read the whole book carefully. However, I must admit that a fairly decent understanding can probably be obtained by reading only the "easy" parts. What I would suggest for the reader with minimal mathematical background, and small tolerance for formulas, is to skip the three chapters mentioned above, flipping back to them only as cross-referencing and curiosity require.

Some readers may be disturbed by my heady mixture of "finished" research projects, raw philosophical speculations, and promising scientific beginnings. However, my feeling is that his mixture is entirely appropriate to the subject matter. After all, at the present time, both complex systems science and cognitive science may in themselves be considered "promising beginnings." Both are exciting intellectual adventures, in which philosophy and science are entangled in a most intriguing and confusing way. In this sense, the present book embodies Marshall McLuhan's principle "the medium is the message." Its network of deeply interlinked concepts, results and speculations reflects the structure of complex systems much more accurately than would a standard linear text focusing on a single narrowly defined "topic," or arguing for a particular, simply defined logical "point." By putting the various subfields of complexity science together, toward a common but complex goal, one begins to form an understanding of the way complex systems themselves are put together to make our universe.

PRELUDE

ON COMPLEX SYSTEMS SCIENCE

Complexity science is young enough that every book and article written has a measurable effect on the definition of the field. It thus seems worthwhile to pause for a moment and reflect on what complexity science **is** -- and what it is becoming.

Picking up the book *Complexity* by Roger Lewin in the Waikato University bookstore, and glancing at the back cover, I was surprised to read the following words:

Complexity theory is destined to be the dominant scientific trend of the 1990's.

But what is it?

It is the unifying theory which states that at the root of all complex systems lie a few simple rules. This revolutionary technique can explain any kind of complex system -- multinational corporations, or mass extinctions, or ecosystems such as rainforests, or human consciousness. All are built on the same few rules.

"Wow!" I thought, and eagerly flipped through the pages to discover these remarkable new rules which had, in all my years of studying systems theory and complexity science, somehow eluded me. To my great disappointment, however, all I found was a well-written description of things that I had known about for years: the Santa Fe approach to complex adaptive systems, the Gaia hypothesis, Daniel Dennett's theory of consciousness, and so on. If one puts aside all the advertising hype, what is left of complexity science? The truth is that, at present, complexity science is not so much a coherent science as a loose collection of subfields. Dynamical systems theory, fractal graphics, neural networks, genetic algorithms, cellular automata, mathematical immunology, and so forth, to name a few examples, are all thriving research areas, and all have a lot to say about the behavior of complex systems. But, as of this writing (late 1994), there really is no "unifying theory" of complex systems. The "theory which states that at the root of all complex systems lie a few simple rules" does not exist.

But there is a glimmer of truth underlying the hyperbole. There may be no "unifying theory," but there **is**, nevertheless, something very interesting going on with the science of complexity. There is a tantalizing pattern to the connections between the various research areas. Appropriately, the current state of complexity science is best described by ideas from complexity science itself. One might say that the various subfields are undergoing a slow, steady process of convergence -- if not convergence to a **fixed point**, as some classical philosophies of science would require, then at least convergence to some kind of **strange attractor**. Or, in a different vocabulary, one might hypothesize that we are now experiencing the beginning stages of a **phase transition**, in which the network of relationships that defines complexity science suddenly assumes a densely interconnected structure.

It is not clear, at this point, just how far the complex systems research programme is going to be able to push -- or in what direction. Clearly, many properties of real-world systems are indeed system-specific; complex-systems considerations can't tell you everything. But on the other hand, if complex systems ideas prove sufficiently informative in a sufficient variety of situations, they may come to be considered the general background against which more specific properties are to be viewed and investigated. This is clearly, it seems to my biased eye, the direction in which things are going.

At bottom, as has been pointed out forcefully by Sally Goerner (1994), complex systems science is all about interdependence. Computer technology has allowed us to explore the interdependences between various factors which, in previous decades, we were forced to treat as being "approximately" independent. Fractals, chaos, genetic algorithms and so forth all result from this emerging theoretical and empirical acknowledgement of interdependence. It is hardly surprising, then, that the various subdisciplines of complex systems science should also demonstrate an accelerating interdependence!

I suspect that there will never be "unified rules of complexity." Like many others, I believe that we are pushing toward an entirely different kind of science, toward a more fuzzily defined collection of theoretical schema and computational tools. The very **structure** of complex systems science is, like the details of the Mandelbrot set, something intriguing and unprecedented.

Complexity as a New Kind of Science

In evaluating this idea that complexity science is not only a new science, but a new **kind** of science, an historical analogy may be useful. Philosophers of science are well aware that history-based sciences like evolution theory and cosmology have to be treated differently from other sciences -- because they, by their very nature, deal with parts of the world that are difficult to experiment on. We cannot build new universes, or evolve new hominids, in order to test our theories. In experiment-based sciences, it is realistic to judge theories by whether they correctly predict the results of experiments, or whether they suggest interesting new experiments. In observation-based sciences like evolution and cosmology, however, the emphasis has to be placed on **conceptual coherence**. One is looking for theories that explain a wide variety of data with minimally complicated hypotheses. A premium is placed, in history-based science, on a theory's ability to predict the data that will be found in different situations. But the amount of data available is strictly restricted, and the quality of data is limited by sources beyond one's control. So, more than in experiment-based science, it is conceptual coherence that is the deciding factor.

The theory of evolution by natural selection is a wonderfully simple idea. It explains the observed diversity of species, and the nature of the fossil record; it connects closely with genetics, and the practices of animal and plant breeding. Similarly, the Big Bang theory explains a huge number of astronomical phenomena, including the distribution of matter in the universe, the cosmic background radiation. Neither of these theories is directly testable in the classic sense -- but both are fabulous science.

Just as history-based sciences have to be treated differently, so, I claim, does complexity science. By its very nature, it deals with high-level emergent patterns, which behave differently from the simple physical quantities studied by other natural sciences. Complex systems have a variety and variability that makes reliable experimentation difficult. Different types of complex systems are similar in many abstract, subtle ways, which are difficult to rigorously characterize. They are not similar enough to be accurately modelled with the same equations, but they are similar enough that, when viewed on a coarse level, they display behaviors that are "identical in character."

Many researchers believe, with Lewin, that complexity science will eventually become like traditional, experiment-based science -- that researchers will find precise equations for the relations between complex systems. They believe that the current focus on qualitative models and vaguely-defined "archetypal" behavior is just a transitional phase. Of course, this is possible. But, looking at what complexity science actually **is**, we see something quite different from sciences like, say, solid-state physics, protein chemistry, mineralogy or haematology. We see a science that is based on fuzzily-defined but intuitively recognizable **abstract forms** -- a science

that manifests, in the most striking possible way, Plato's notion of abstract Ideas which are only approximated by real entities, in various different ways.

A science of this nature, I claim, should be judged differently than other kinds of science. It should be judged, firstly, by the archetypal patterns it provides, and how these patterns help us to understand particular complex systems; and secondly, by the analytical and computational tools it provides, and how these tools help us to tease out the ways the archetypal patterns manifest themselves in particular systems. At any rate, that is the spirit in which the ideas reported here were conceived, and are presented.

Complexity and Psychology

The main disciplines touched on here are mathematics, computer science and psychology. The relation between complexity science and mathematics and computing needs little note. The relation between complexity science and the discipline of **psychology**, however, may be a less obvious matter.

The first point to be noted is the remarkable extent to which theoretical psychology is currently **fragmented**. The concepts used to model perception are by and large completely different from those used to model personality, which in turn are different from those used to model cognition, etc. While there are bound to be significant differences between different aspects of mental process, it seems likely that there is more underlying commonality than current theory reveals.

Two examples will serve to illustrate this point. First: memory. While cognitive psychologists tend to speak in terms of memory systems, the cognitive neuroscientists tell us that memory, as an independent entity, does not exist. The brain does not store inactive "memory tokens"; it contains active processes, which construct the neural activity patterns we interpret as memories, and often do other things as well.

In a similar way, until very recently, perceptual psychologists tended to speak of perceptual processing systems as independent modules. However, input from cognitive neuroscience has consistently contradicted this, and has taught us that much of vision processing (to name only the sense mode that has received the most attention) is based on interconnections with cognitive, motor and emotional centers of the brain.

In one case after another it is revealed that, what psychologists model as separate systems operating under their own principles, are actually carried out by the brain in an interconnected and unified way. This fact has led several researchers to construct general "complex systems theories" of mental process, emphasizing the interconnectedness and the self-organizing nature of the mind/brain. My own psynet model falls in this category, as does Kampis's component-systems model, and the Abrahams' dynamical-systems psychology.

From the point of view of psychology, these system-theoretic models are excessively general (and my own model is no exception here). They speak of the general structure of mental systems, rather than of what is particular to human intelligence. However, they are superior to mainstream psychological theories in one very important respect: they give a concrete vision of mind as a

whole. Thus it is of intense interest, psychologically speaking, to extend these models in a more concrete and definite direction, by looking for general, complex-systems-theoretic principles of **human** mental function.

This is the importance of complexity science for psychology. But what of the importance of psychology for complexity science? The study of psychological systems, as pursued here, has several lessons to teach us regarding complexity science in general. Most of all, it leads to a vision of complexity which centers on **emergent pattern** and **autopoiesis**, rather than the numerical details of low-dimensional iterations.

Of course, everyone recognizes that complexity is all about "emergence," in some sense -- but in practice, most work in complex systems science still has to do with studying numerical variables. I believe that, particularly in discussing **very** complex systems such as psychological systems, it is necessary to push even further in the direction of emergence, and to treat algorithmic pattern/processes, rather than numerical variables, as the main topics of discourse. The self-organization and interproduction of **patterns** must be kept at the forefront. **Magician system models**, as introduced in *Chaotic Logic*, are stressed here, as a way of capturing this kind of abstract, autopoietic pattern dynamics.

This new angle on standard complex systems models presented here serves to balance the "hard science" bias which one finds in most complexity science. Complexity science evolved primarily within physics, computer science and applied mathematics, with some input from biology, chemistry, economics, etc. The complexity/psychology connection has only very recently attracted significant attention. Here I look at complex systems ideas and models from a theoretical-psychology perspective -- to see how thse models must be extended in order to deal with that most complex of systems, the mind. This exercise is, I believe, valuable for cognitive science and complexity science in particular, as well as for the psynet model in particular.

Background Reading in Complexity Science

Despite the wide sweep of the ideas, this is fundamentally a research monograph and not a textbook. In particular, it is not a good introduction to the study of complex systems. Thus it may be appropriate to give a few suggestions for background reading for the reader who is not "up to speed" regarding the emerging science of complexity.

First of all, since I have critiqued the blurb on the cover of Lewin's *Complexity*, I must say that -- while it does suffer, to a certain extent, from the usual sins of omission and lionization -- this is an unusually lucid and intelligent book. Lewin does an excellent job of getting across the general flavor of complexity science as it existed in the early 1990's. This era has already passed, and by now complexity science has become far more diffuse and diverse -- but it is a time period well worth reading about.

A different kind of review of complexity science, more focussed on ideas rather than personalities, is Coveney and Highfield's book *Frontiers of Complexity* (1995). I have used this book as the main textbook for a liberal arts class, and the students found it enjoyable and readable despite its sometimes intricate scientific detail.

Next, although it is not fashionable to admit it, modern complexity science owes a great deal to the half-century-old discipline of **general systems theory**. Erich Jantsch's classic *The Self-Organizing Universe* is an excellent overview of the accomplishments of general systems theory. Sally Goerner's *The Evolving Ecological Universe* (1994) ties chaos and complexity together with modern work in thermodynamic systems theory and also with more general social trends. Finally, the works of Gregory Bateson, most notably *Mind and Nature* (1980), are also not to be missed.

Regarding dynamical systems theory in particular, in addition to popular treatments like Gleick's *Chaos* (1988), there is an abundance of quality texts on all different levels. A popular choice is Robert Devaney's *Chaotic Dynamical Systems* (1988), which is aimed at upper level undergraduate and first year graduate students in mathematics. *A Visual Introduction to Dynamical Systems Theory for Psychology*, by Fred Abraham, Ralph Abraham and Chris Shaw (1991), gives a nice overview of basic concepts as well as describing many solid scientific applications.

Regarding genetic algorithms and evolutionary computation, the picture is not so clear -- no one has yet written a comprehensive textbook surveying the full variety of EC theory and applications. The reader who wishes more general background information on EC is therefore referred to not one but four good books reviewing aspects of the "state of the art" in evolutionary computation. Goldberg's *Genetic Algorithms for Search, Machine Learning and Optimization* (1988) gives the classic treatment of the bit string GA. Michaelewicz, with *Genetic Algorithms + Data Structures = Evolution Programs* (1993), has written the bible for floating point based GA's and "evolution programs." Koza's massive tomes, *Genetic Programming* and *Genetic Programming II* (1990, 1993) make the best case so far for the genetic programming approach. Regarding classifier systems, the best reference is still the classic work *Induction* (Holland et al, 1986).

The Artificial Life conference proceedings volumes, especially *Artificial Life II* (Langton et al, 1992), give an excellent survey of applications of evolutionary computation, dynamical systems theory and other ideas to the problem of constructing artificial life forms. Also, Stuart Kauffman's *The Origins of Order* (1993), though by no means easy reading, summarizes a long list of research projects in complexity science accomplished by Kauffmann and his colleagues over the years.

Finally, the small but important subfield of "system-theoretic cognitive science" is referred to frequently in the following pages. This research is grounded in classical systems theory as much as in modern complexity science; it is exemplified by books such as Francisco Varela's *Principles of Biological Autonomy* (1978), Vilmos Csanyi's *Evolutionary Systems: A General Theory* (1989), George Kampis's *Self-Modifying Systems in Biology and Cognitive Science* (1991), and my own *Chaotic Logic, Structure of Intelligence* and *Evolving Mind*.

PART I

THE COMPLEX MIND/BRAIN

CHAPTER ONE

DYNAMICS, EVOLUTION, AUTOPOIESIS

1.1 INTRODUCTION

In this brief chapter, I will introduce a few preliminary ideas required for understanding the rest of the book: dynamics, attractors, chaos, genetic algorithms, magician systems, and algorithmic patterns in dynamics.

These concepts are diverse, but every one of them is a variation on the theme of the first one: **dynamics**. For "dynamics," most broadly conceived, is just change, or the study of how things change. Thus, for example, evolution by natural selection is a kind of dynamic. Comparing natural selection with physical dynamics, as portrayed e.g. by the Schrodinger equation, gives one a rough idea of just how broad the concept of dynamics really is.

Physics has focused on dynamics at least since the time of Newton. Psychology and the other human sciences, on the other hand, have tended to focus on statics, ignoring questions of change and process -- a fact which is probably due to the incredible **dynamical complexity** of human systems. Only now are we beginning to find mathematical and computational tools which are up to dealing with the truly complex dynamical systems that we confront in our everyday mental and social lives.

There is a vast and intricate body of mathematics called "dynamical systems theory." However, the bulk of the concrete results in this theory have to do with dynamical systems that are governed by very special kinds of equations. In studying very complex dynamical systems, as we will be doing here, one is generally involved with conceptual models and computer simulations rather than analytical mathematics. Whether this state of affairs will change in the future, as new kinds of mathematics evolve, is extremely unclear. But at present, some of the most interesting dynamical systems models are so complex as to almost completely elude mathematical analysis.

An example is the genetic algorithm, a discrete and stochastic mathematical model of evolution by natural selection. Another example is the autopoietic system, a model of the **self-producing** nature of complex systems. Both these dynamical systems models will be discussed in detail in later chapters, and briefly introduced in this chapter. It is these models, I believe, which are most relevant to modeling extremely complex systems like the mind/brain. In order to model psychological complexity, we must leave traditional dynamical systems theory behind, and fully confront the messiness, intricacy and emergent pattern of the real world.

In the last section of this chapter, the relation between dynamics and **structure** will be discussed, and dealt with in a formal way. The abstract theory of pattern, developed in my previous publications, will be introduced and used to give formal definitions of such concepts as "dynamical pattern" and "system complexity." The concept of pattern unifies all the various dynamics discussed in the previous sections. For these dynamics, like all others, are in the end just ways of getting to appropriate **emergent patterns**.

Finally, before launching into details, it may be worthwile to briefly reflect on these topics from a more philosophical perspective. Change and structure -- becoming and being -- are elementary philosophical concepts. What we are studying here are the manifestations of these concepts within the conceptual framework of conceptual science. Any productive belief system, any framework for understanding the world, must come to terms with being and becoming in its own way. Science has, until recently, had a very hard time dealing with the "becoming" aspect of very complex systems, e.g. living and thinking systems. But this too is changing!

1.2 ATTRACTORS

Mathematical dynamical systems theory tends to deal with very special cases. For instance, although most real-world dynamical systems are most naturally viewed in terms of **stochastic** dynamics, most of the hard mathematical results of dynamical systems theory have to do with **deterministic** dynamics. And even within the realm of deterministic dynamics, most of the important theorems involve quantities that are only possible to compute for systems with a handful of variables, instead of the hundreds, thousands or trillions of variables that characterize many real systems. A course in dynamical systems theory tends to concentrate on what I call "toy iterations" -- very simple deterministic dynamical systems, often in one or two or three variables, which do not accurately model any real situation of interest, but which are easily amenable to mathematical analysis.

The great grand-daddy of the toy iterations is the logistic map, $x_{n+1} = rx_n(1-x_n)$; a large portion of Devaney's excellent book *Chaotic Dynamical Systems* is devoted to this seemingly simple iteration. In the realm of differential equations, the Lorenz equation is a classic "toy model" (its discretization, by which its trajectories are studied on the computer, is thus a toy iteration). Implicit in the research programme of dynamical systems theory is the assumption that the methods used to study these toy iterations will someday be generalizable to more interesting dynamical systems. But this remains a promissory note; and, for the moment, if one wishes to model real-world systems in terms of dynamical systems theory, one must eschew mathematical theorizing and content oneself with qualitative heuristics and numerical simulations.

But even if the deeper results of dynamical systems theory are never generalized to deal with truly complex systems, there is no doubt that the conceptual vocabulary of dynamical systems theory is useful in all areas of study. In the following pages we will repeatedly use the language of dynamical systems and attractors to talk about psychological systems, but it is worth remembering that these concepts did not (and probably could not) have come out of psychology. They were arrived at through years of painstaking experimentation with simple, physics-inspired, few-variable dynamical systems.

So, let us define some terms. A **dynamical system**, first of all, is just a mapping from some abstract space into itself. The abstract space is the set of "states" of the system (the set of states of the real-world system modelled by the mapping; or the abstract dynamical system implicit in the mapping). The mapping may be repeated over and over again, in discrete time; this is an iterative dynamical system. Or it may be repeated in continuous time, in the manner of a differential equation; this is a "continuous" dynamical system. In general, continuous dynamical systems are more amenable to mathematical analysis, but discrete dynamical systems are more

amenable to computer simulation. For this reason, one often has cause to transform one kind of dynamical system into the other.

A trajectory of a dynamical system is the series of system states that follows from a certain initial (time zero) state. For a deterministic dynamical system, a trajectory will be a simple series, for a stochastic dynamical system, it will be a constantly forward-branching collection of system states. When doing computer simulations of dynamical systems, one computes particular sets of trajectories and takes them as representative.

The key notion for studying dynamical systems is the **attractor**. An attractor is, quite simply, a **characteristic behavior** of a system. The striking insight of dynamical systems theory is that, for many mathematical and real-world dynamical systems, the initial state of the system is almost irrelevant. No matter where the system starts from, it will eventually drift into one of a small set of characteristic behaviors, a small number of attractors. The concept of "attractor" is, beyond all doubt, the most important contribution of dynamical systems theory to the general vocabulary of science.

Some systems have fixed point attractors, meaning that they drift into certain "equilibrium" conditions and stay there. Some systems have periodic attractors, meaning that, after an initial transient period, they lock into a cyclic pattern of oscillation between a certain number of fixed states. And finally, some systems have attractors that are neither fixed points nor limit cycles, and are hence called **strange attractors**. An example of a two-dimensional strange attractor, derived from equation (3) below, may be found in Figure 1. The most complex systems possess all three kinds of attractors, so that different initial conditions lead not only to different behaviors, but to different **types** of behavior.

The formal definition of "strange attractor" is a matter of some contention. Rather than giving a mathematical definition, I prefer to give a "dictionary definition" that captures the common usage of the word. A strange attractor of a dynamic, as I use the term, is a collection of states which is: 1) invariant under the dynamic, in the sense that if one's initial state is in the attractor, so will be all subsequent states; 2) "attracting" in the sense that states which are near to the attractor but not in it will tend to get nearer to the attractor as time progresses; 3) not a fixed point or limit cycle.

The term "strange attractor" is itself a little strange, and perhaps deserves brief comment. It does not reflect any **mathematical** strangeness, for after all, fixed points and limit cycles are the exception rather than the rule. Whatever "strangeness" these attractors possess is thus purely psychological. But in fact, the psychological strangeness which "strange attractors" originally presented to their discoverers is a thing of the past. Now that "strange attractors" are well known they seem no stranger than anything else in applied mathematics! Nevertheless, the name sounds appealing, and it has stuck.

Strictly speaking, this classification of attractors applies only to **deterministic** dynamical systems; to generalize them to stochastic systems, however, one must merely "sprinkle liberally with probabilities." For instance, a fixed point of a stochastic iteration $x_{n+1} = f(x_n)$ might be defined as a point which is fixed with probability one; or a p-fixed point might be defined as one

for which $P(f(x) = x) > p$. In the following, however, we will deal with stochastic systems a different way. In the last section of this chapter, following standard practice, we will think about the IFS random iteration algorithm by transplanting it to an abstract spaces on which it is deterministic. And in Chapter Six I will study the genetic algorithm by approximating it with a certain **deterministic** dynamical system. While perhaps philosophically unsatisfactory, in practice this approach to stochasticity allows one to obtain results that would become impossibly complicated if translated into the language of true stochastic dynamical systems theory.

Chaos

A great deal of attention has been paid to the fact that some dynamical systems are **chaotic**, meaning that, despite being at bottom deterministic, they are capable of passing many statistical tests for randomness. They **look** random. Under some definitions of "strange attractor," dynamics on a strange attractor are **necessarily** chaotic; under my very general definition, however, this need not be the case. The many specialized definitions of "chaos" are even more various than the different definitions of "strange attractor."

Let us look at one definition in detail. In the context of discrete dynamical systems, the only kind that will be considered here, Devaney defines a dynamical system to be chaotic on a certain set if it displays three properties on that set: sensitive dependence on initial conditions, topological transitivity, and density of repelling periodic points.

I find it hard to accept "density of repelling periodic points" as a necessary aspect of chaos; but Devaney's other two criteria are clearly important. Topological transitivity is a rough topological analogue of the more familiar measure-theoretic concept of ergodicity; essentially what it means is that the dynamics thoroughly mix everything up, that they map each tiny region of the attractor A into the whole attractor. In technical terms an iteration f is topologically transitive on a set A if, given any two neighborhoods U and V in A, the iterates $f_n(U)$ will eventually come to have a nonempty intersection with V.

Sensitivity to initial conditions, on the other hand, means that if one takes two nearby points within the attractor and uses them as initial points for the dynamic, the trajectories obtained will rapidly become very different. Eventually the trajectories may become **qualitatively** quite similar, in that they may have the same basic shape -- in a sense, this is almost guaranteed by topological transitivity. But they will be no more or less qualitatively similar than two trajectories which did not begin from nearby initial points.

Although Devaney did not realize this when he wrote his book, it has since been shown that topological transitivity and density of periodic points, taken together, **imply** sensitivity to initial conditions. Furthermore, for maps on intervals of the real line, topological transitivity and continuity, taken together, imply density of periodic points. So these criteria are intricately interconnected.

The standard tool for quantifying the degree of chaos of a mapping is the **Liapunov exponent**. Liapunov exponents measure the **severity** of sensitive dependence on initial conditions; they tell

you **how fast** nearby trajectories move apart. Consider the case of a discrete-time system whose states are real numbers; then the quantity of interest is the ratio

$$R_n = |f^n(x) - f^n(y)|/|x-y| \quad (1)$$

where f^n denotes the n-fold iterate of f. If one lets y approach x then the ratio R_n approaches the derivative of f^n at x. The question is: what happens to this derivative as the elapsed time n becomes large? The Liapunov exponent at x is defined as the limiting value for large n of the expression

$$\log[\, f^{n'}(x)\,]/n \quad (2)$$

If the difference in destinations increases slowly with respect to n, then the trajectories are all zooming together, the ratio is less than one, and so the Liapunov exponent is negative. If the trajectories neither diverge nor contract, then the ratio is near one, and the Liapunov exponent is the logarithm of one -- zero. Finally, and this is the interesting case, if the difference in destinations is consistently **large** with respect to n, then something funny is going on. Close-by starting points are giving rise to wildly different trajectories. The Liapunov exponent of the system tells you **just how different** these trajectories are. The bigger the exponent, the more different.

A system in one dimension has one Liapunov exponent. A system in two dimensions, on the other hand, has two exponents: one for the x direction, computed using partial derivatives with respect to x; and one for the y direction, computed using partial derivatives with respect to y. Similarly, a system in three dimensions has **three** Liapunov exponents, and so forth.

A positive Liapunov exponent does not guarantee chaos, but it is an excellent practical indicator; and we shall use it for this purpose a little later, in the context of the **plane quadratic iteration** as given by:

$$x_{n+1} = a_1 + x_n*(a_2 + a_3*x_n + a_4*y_n) + y_n*(a_5 + a_6*y_n)$$

$$y_{n+1} = a_7 + x_n*(a_8 + a_9*x_n + a_{10}*y_n) + y_n*(a_{11} + a_{12}*y_n)$$

(3)

This is an example of a very simple dynamical system which mathematics is at present almost totally unable to understand. Consider, for instance, a very simple question such as: **what percentage** of "parameter vectors" $(a_1,...,a_{12})$ will give rise to strange attractors? Or in other words, how common is chaos for this iteration? The answer, at present, is: go to the computer and find out!

1.3 THE GENETIC ALGORITHM

Now let us take a great leap up in complexity, from simple polynomial iterations to the dynamic of **evolution**.

As originally formulated by Darwin and Wallace, the theory of evolution by natural selection applied only to **species**. As soon as the theory was published, however, theorists perceived that natural selection was in fact a very general dynamic. Perhaps the first to view evolution in this way was Herbert Spencer. Since the time of Darwin, Wallace and Spencer, natural selection has been seen to play a crucial role in all sorts of different processes.

For instance, Burnet's theory of clonal selection, the foundation of modern immunology, states that immune systems continually self-regulate by a process of natural selection. More speculatively, Nobel Prize-winning immunologist Gerald Edelman has proposed a similar explanation of brain dynamics, his theory of "neuronal group selection" or **Neural Darwinism**. In this view, the modification of connections between neuronal groups is a form of evolution.

The very origin of life is thought to have been a process of molecular evolution. Kenneth Boulding, among many others, has used evolution to explain **economic** dynamics. Extending the evolution principle to the realm of culture, Richard Dawkins has defined a "meme" as an idea which replicates itself effectively and thus survives over time. Using this language, we may say that natural selection itself has been a very powerful meme. Most recently, the natural selection meme has invaded computing, yielding the idea of **evolutionary computation**, most commonly referred to by the phrase "genetic algorithms."

These diverse applications inspire a view of evolution as a special kind of **dynamic**. The evolutionary dynamic is particularly useful for modeling extremely complex systems -- biological, sociological and cultural and psychological systems. Evolutionary dynamics has its own properties, different from other dynamics. All systems which embody adaptive evolution will display some of the characteristic properties of evolutionary dynamics -- along with the characteristics of other dynamics, such as the structure-preserving dynamic of "autopoiesis" or "ecology."

Evolutionary Computing

A particular focus will be laid here on evolutionary computing -- not only because evolutionary computer programs are interesting in their own right, but because of the light they shed on evolutionary dynamics in general. Computational models of evolution represent the evolutionary process stripped bare, with the intriguing but distracting intricacies of the biological world stripped away. They allow us to get at the **essence** of the evolutionary process in a way that is sometimes quite striking.

From a practical, computational perspective, the exciting thing about evolutionary computation is that it seems to work by "magic." The recipe is so simple. First, choose an appropriate representation from the standard repertoire -- i.e., represent the set of possible solutions to one's problem as a collection of binary strings, or floating-point vectors, or LISP programs. Next, formulate a fitness function: a function which assigns each potential solution a numerical measure of "quality." Supplied with standard default routines for crossover, mutation and selection, the genetic algorithm will do the rest. Starting from a random initial population, it lets the population reproduce differentially with respect to fitness for a number of generations. Then -

- presto! out pops an answer. With luck, the answer is a good one: a maximizer or near-maximizer of the fitness function.

The detailed dynamics of an evolutionary program are for all practical purposes incomprehensible. Except in trivial cases, there is no workable way to get at "what the program is doing," on a step-by- step basis. One just has to watch, wait and see; and if the result is bad, one adjusts the parameters, the fitness criterion or the representation, and begins again. This mystery as to internal dynamics can be frustrating to those scientists who are accustomed to a greater degree of control. But the key point is that the frustration and the magic of evolutionary computation are **exactly parallel** to the frustration and the magic of real, non-computational evolution. Both real-world and simulated evolution produce a lot of duds, often for hard-to-understand reasons. And both produce a fair number of successes, sometimes slowly, sometimes quickly, and with the occasional appearance of miracle.

From a psychological point of view, the main interest of evolutionary computing is as a general conceptual and computational model of **problem-solving** and **creativity**. Psychological theorists going back to William James and Charles Peirce have viewed the process of learning and invention in evolutionary terms. Evolutionary computing is a concrete instantiation of their insight: it presents systems which actually do learn by evolution. Here we will primarily work with genetic algorithms, but with an eye always on the more general class of evolutionary learning dynamics.

Our interest in genetic algorithms, in this book, has multiple dimensions. First, we will be interested in the **dynamics** of evolutionary learning systems, and in exploring whether there is any commonality between these dynamics and the dynamics of human learning. We will be interested in the use of genetic algorithms to generate complex forms, rather than merely solving optimization problems. We will be interested in evolutionary computing systems as **autopoietic** systems, in the role which self-organization and self-production plays in the dynamics of evolution. Finally, we will be interested in whether it is possible to extend such systems as genetic algorithms, which are obviously very limited models of human creativity, into more complete and psychologically realistic models.

The Simple GA

Perhaps the most common form of evolutionary computation is the Simple Genetic Algorithm, or **simple GA**. The simple GA is an extremely crude model, both biologically and computationally, but as the name suggests, it has the virtue of simplicity. It leaves out a number of features that are absolutely essential to biological evolution -- most notably, spatial distribution, and ecology. Even so, however, it has provided us with a great deal of insight into the evolutionary process.

The simple GA, as I mean it here, consists of a population of fixed size N, each element of which is a binary string of length M. It begins with a random initial population and then, to get from one generation to the next, proceeds as follows:

1) Evaluate the objective "fitness" function value f(x) of

each population element x. Sum these values to obtain the quantity SUMFITNESS. The probability of selecting x will then be given by f(x)/SUMFITNESS.

2) Select two elements of the population, call them x and y, and compute the crossover product C(x,y). Then mutate each bit of this crossover product with a certain independent probability (mutation rate). Repeat this step until M new binary strings have been created.

3) Replace the old population with the new population created in Step 2, and return to Step 1.

What is this thing called a "crossover operator"? On an abstract level, one might make the following definitions:

A **crossover operator** is any map C(,) from the space of pairs of length n bit strings to the space of probability distributions on length n bit strings.

A **true crossover operator** is any crossover operator which possesses the following property: if neither s nor t possess a B in the i'th position, then the probability that C(s,t) possesses a B in the i'th position must be zero (here B is either 0 or 1). In other words, a true crossover operator picks from among the bits provided by the "parents" it is crossing; it does not introduce anything new, but only recombines.

But this level of generality is almost never utilized in practice. Usually in practical and theoretical work with GA's, a bitwise or multi-point cut-and-paste operator is assumed.

Most simply, one may use a single point cut-and-paste operator, according to which, in order to cross $a_1...a_M$ with $b_1...b_M$, one chooses at random a cut-point r between 1 and M, and then forms $a_1...a_r b_{r+1}...b_M$. Mutation then proceeds through the string bit by bit: supposing the mutation rate is .01, then each bit is changes to its negation independently with probability .01.

For instance, suppose one has

00001111101

as the first parent, and

10101010101

as the second parent. If the cut-point is randomly chosen as 5, then the cut occurs before the fifth position, yielding

0000 | 1111101 parent 1

1010 | 1010101 parent 2

———————

0000 1010101 "pure crossover" child

and finally, perhaps

00001011101

as the mutated child.

The specific mechanics of crossover are not essential to the function of the GA; any crossover operator that provides a broad-based search of the space of genotypes will do. The point is that crossover is a very flexible, "intelligent" kind of search. In the language of search algorithms, instead of merely searching from one point to other nearby points, it does a wide-ranging but subtly constrained stochastic search, using the information available in **pairs** of points and **populations** of points.

Beyond the Simple GA

The simple GA is ideal for mathematical analysis, but it can be cumbersome for practical GA work. One common modification is to use **real vectors** instead of bit strings as one's population elements. This is more natural for many applications, particularly since most modern programming languages offer a floating point data type. Numerous examples of the floating point GA are given in Michaelewicz's recent book.

Crossover of two real vectors is easy to define. In the one cut-point model, for instance, one might cross (.33,.55,.44) with (.55,.34,.17) to obtain any of the vectors:

(.33,.55,.44)

(.33,.55,.17)

(.33,.34,.17)

(.55,.34,.17)

(note the noncommutativity of the crossover operation, here as in the bit string model).

Mutation of two real vectors is a little trickier: it is not as simple as merely flipping between 0 and 1 in the bit string case. Instead, to mutate an entry x in a vector, one selects a number from some probability distribution centered around x. So, if (.33,.55,.44) were crossed with (.55,.34,.17) to obtain the unmutated offsprint (.33,.55,.17), the final, **mutated** offspring might look something like (.36,.50,.18). One can see from this that mutation plays a somewhat larger role in the floating point GA than in the bit string GA.

The role that the "mutation probability" plays in the bit string model is here played by the **standard deviation** of this probability distribution. A Gaussian distribution is the standard choice here, but there are also arguments for using something like a Cauchy distribution, which

gives more weight to outlying values. Sometimes it is useful to gradually decrease this standard deviation from one generation to the next.

Substituting vectors and their genetic operations for bit strings and their genetic operations, one obtains a "simple floating-point GA." The simple GA and simple floating-point GA are really very similar as evolutionary implementations go. Much more radical is, for instance, John Koza's "genetic programming" implementation, in which the representation is neither strings nor vectors but simple LISP programs. The LISP programs are represented as trees labeled with commands. Mutating a program involves replacing one command with another; crossing over two programs means taking a subtree from one program tree and putting it in place of some subtree from the other program tree. As will be briefly argued later, this kind of graph-based evolutionary computation would seem to have a deep relevance to cognitive science. For practical purposes, however, it is at the present time rather cumbersome.

The Dynamics of Genetic Optimization

What, finally, is the **psychological** meaning of the GA? This is a topic for later chapters, but a hint may be given here.

In Chapter Six we will study the dynamics of the genetic algorithm in the case of very large population size, using tools from standard dynamical systems theory. This investigation will lead to a number of interesting results, including a mathematical parallel between genetic algorithms and the dynamics human and animal learning. Crossover, it seems, provides the kind of learning power that biological systems have. Mutation does not.

Also, we will look at the possibility of embedding the genetic algorithm in a spatially-extended, ecological model. This leads to much more interesting dynamics. The spatial organization displays typical properties of self-organizing systems, and one sees the emergence of an hierarchy of complex spatiotemporal attractors. This kind of genetic algorithm, it will be argued, has many interesting analogues with the psynet model of mind, and may well be able to serve as a computational substrate for the emergent structures described in the psynet model.

Finally, in the last chapter, the genetic algorithm will re-emerge as a kind of computational metaphor for a certain phase of human creativity. Typically, crossover of mental structures is strongly contrained by autopoietic thought-systems. But in the highly creative portions of the mind, this constrained is lessened, and one has a free-flowing recombination and mutation process quite similar to the GA.

1.4 MAGICIAN SYSTEMS

Beyond continuous dynamical systems and genetic algorithms, we will require one further kind of complex systems model, the **magician system**. Philosophically, magician systems are largely equivalent to Maturana and Varela's "autopoietic systems," and Kampis's "component-systems" (see Varela, 1978; Kampis, 1992) However, the focus is different, and the precise mathematical formulation (as will be elaborated in later chapters) is different.

What are autopoietic and component systems? An autopoietic system, as defined by Maturana and Varela, is a system which produces itself. It is a system of components which interact according to some network of interrelationships; and so that the network of interrelationships is produced by the system components themselves. The paradigm case is the biological organism. Autopoietic systems are generally dissipative, i.e. they do not conserve energy. What they do conserve, however, is **structure**. A body is an open system, creating entropy, but what it uses its energy doing is maintaining its own structure.

A component-system, as defined by Kampis, is something only a little different: it is simply a collection of components which are capable of transforming each other, and of coming together to form larger components. In its emphasis on transformations and compounds, this vision of complex dynamics reflects Kampis's background as a chemist. However, Kampis does not give a formal mathematical definition of component-systems; he claims this is not possible using current mathematics.

The relation between these abstract, system-theoretic models and **computation** is somewhat problematic. Kampis has proclaimed component-systems to be fundamentally non-computational; Varela has made similar remarks about autopoietic systems. However, both of these scientists have studied their own system-theoretic models using computer simulations. Kampis presents a semi-rigorous "proof" that component-systems are uncomputable; however, the proof only applies to deterministic Turing machines, not stochastic computers or quantum computers (Deutsch, 1985). I have argued in *CL* that component-systems are stochatically computable; and I have also pointed out, in *SI*, that there is no way to tell stochastically computable systems from deterministically computable systems. In essence, "This system is stochastic" just means "This system has aspects which appear to me patternless." So, in my view, the computability issue turns out to be a non-issue.

In *CL*, I have given a mathematical formulation in terms of component-systems in terms of hypersets, a concept from transfinite set theory; and I have argued that, for practical purposes, these hyperset component-systems can be effectively modeled in terms of computational systems. However, Kampis (personal communication) does not fully accept this reduction, and so in *CL* I introduce a new system-theoretic model called the Self-Generating System or, more compactly, **magician system**. The magician system captures largely the same insight as Kampis's component-systems, but in a more precisely specified way.

A magician system consists, quite simply, of a collection of entities called "magicians" which, by casting spells on each other, have the power to create and destroy magicians. Magician A can "cast a spell" transforming magician B into magician C; or magicians C and D can cooperatively "cast a spell" creating an anti-A magician, which annihilates magician A.

The existence of "antimagicians," magicians which have the power to annihilate other magicians, is necessary in order for magician systems to have the full computational power of Turing machines. These antimagicians are much like antiparticles in physics: when magician A and magician anti-A come into contact with each other, both are destroyed. In Chapter Seven magician systems will be formalized in terms of hypercomplex numbers and other abstract

algebras, and the algebraic role of antimagicians will be clarified. For now, however, an informal discussion will suffice.

At each time step, the dynamics of a magician system consists of two stages. First the magicians in the current population act on one another, casting their spells on each other, producing a provisional new population. Then the magician/antimagician pairs in the provisional new population annihilate one another. The survivors are the new population at the next time step.

What I call an **autopoietic subsystem** of a magician system, then, is simply a magician system which is an **attractor** for this dynamic of inter-creation. This usage of the word "autopoiesis" may not accord precisely with the intentions of Maturana and Varela, but the meaning is very close. Both of us are concerned with systems that produce themselves. It seems preferable to use the term "autopoiesis" with a slightly different shade of meaning, rather to creating yet another new word.

For psychological modeling, the case where the magicians cast spells involving pattern-recognition processes is particularly interesting. In *CL*, special terminology is introduced to deal with this case: the magician system iteration is called the **cognitive equation**, and autopoietic subsystems are called **structural conspiracies**. A structural conspiracy is a set of patterns that produces itself by mutual pattern-recognition; an autopoietic subsystem is any set of magicians that produces itself by mutual transformation or "spell-casting."

If a system is a fixed-point attractor under the magician dynamic then it, very simply, **produces itself**. More generally, one may look at collections of magician systems which are limit cycles or strange attractors of the magician dynamic: these too, in a practical context, may be considered autopoietic systems.

There are many ways to vary the basic magician system dynamic; for instance, there is the question of **which** magicians get to act on one another at each time step. The pairs may be selected at random with certain probabilities (the "well-mixed" approximation), or else some spatial structure may be imposed, with each magician acting on its neighbors. Also, the restriction to **pairs** is somewhat arbitrary and unnecessary. One may have magicians which create other magicians all by themselves, or magicians which create by acting on groups of magicians. The product of a magician's action need not be a **single** magician; it may be a group of magicians. Finally, there may be a "filtering" operation which acts on the magician population as a whole, before it is turned into the next generation. However, these variations do not affect the basic ideas.

Needless to say, there are very many concrete instantiations of magician systems. For instance, one might point to enzymes and other substances in biochemistry: these substances act on each other to create each other. Or, more generally, one might speak of the whole array of cellular processes existing in the body. Here, however, we will focus on interpreting the dynamical system of **mental processes** making up an individual's mind as a magician system.

To apply the magician systems formalism to any particular situation, one must make specific commitments about the space of magicians. In the context of the psynet model, abstract

discussions can be made without reference to the nature of this space. When one goes to explore the model's instantiation in specific physical systems, however, one can no longer avoid the question of the nature of the individual magicians. If the reader desires to have a concrete model in mind while reading the present chapter, it is convenient to define magician action by reference to a fixed universal Turing machine which takes **two tapes** instead of one: one tape for "program" and one tape for "data" (this construction was inspired by the work of Moshe Koppel; see (Koppel, 1987)). The product action of A on B denotes the result of running the Turing machine with program A and data B.

For example, in the case of a pattern recognition system, the only programs A to be permitted are those which recognize patterns in their data B. Psychologically, we will think of our mental magicians as abstract pattern/processes, the "programs" of which serve to recognize patterns in their "data," and/or to create patterns in their environments (the "data" of other magicians).

The dynamics of magician systems are even less accessible to mathematical analysis than those of the genetic algorithm. For indeed, as will be shown later on, the genetic algorithm may be viewed as a special kind of magician. The magician elements are genotypes (bit strings, or whatever), and the magician action is crossover plus mutation. Magician systems may thus be viewed as a kind of "generalized genetic algorithm," where the standard crossover operator is replaced by a flexible, individualized crossover operator. This kind of dynamical system is difficult to analyze, difficult to simulate, and difficult to understand. However, I will argue in the following chapter that this is also **precisely** the type of dynamical system we need to use to model the brain/mind.

1.5 DYNAMICS AND PATTERN

The relation between structure and dynamics is a question that lies at the very heart of science. Typically, physics has focused on dynamics, and has understood structures in a dynamical light; while chemistry and biology have understood structures as basic entities in their own right. Psychology, more than any other science, has exalted structure over dynamics, and has, except in the context of developmental psychology, almost entirely ignored processes of change.

On a more fundamental level, structure and dynamics are two different ways of looking at the world, two philosophical perspectives. From the one perspective, the world is composed of fixed entities, with aspects that change. From the other perspective, everything is always changing, and these world's many processes of change give rise to semi-permanent, stable "attractors" that we perceive as definite structures.

These general considerations suggest that it is very important for complexity science to come to grips with the relation between structure and dynamics -- and, most essentially, with the emergence of structure out of dynamics. Indeed, this is just a rephrasing of the title of this book, "Complexity to Creativity." How, out of the intricate, chaotic complexity of micro-level dynamics, do we get the creation of beautiful, complicated macro-level structures?

In order to address this kind of question, I suggest, one has to shift the focus of attention from dynamics itself to **emergent patterns** -- emergent patterns in structures, and emergent patterns in

dynamics. From the pattern-theoretic point of view, structure and dynamics are just different ways of getting to **patterns**.

This perspective has particular psychological importance, given that, in *SI*, I have argued that the mind is **the collection of patterns in the brain**. The relation between neurodynamics and mind is the relation between a substrate and the patterns emergent from it. This is a particularly natural philosophy of mind, which has roots in the pragmatist philosophy of Charles S. Peirce (1935).

A Formal Theory of Pattern

There are many ways to formalize the notion of "pattern." For example, algorithmic information theory (Chaitin, 1987) gives us a convenient way of studying pattern using the theory of universal Turing machines. However, the concept of pattern is arguably more basic than the theory of universal computation. In previous works I have given a very simple mathematical definition of "pattern," and used it to model numerous psychological and biological processes. Namely, one may define a pattern as "a representation as something simpler." In symbols, this means, roughly speaking, that a process p is a **pattern** in an entity e if: 1) the result of p is a good approximation of e, and 2) p is simpler than e.

More rigorously, let s be a "simplicity function" mapping the union of the space of entities and the space of processes into the nonnegative real numbers; and let d be a metric on the space of entities, scaled so that $d(r_p,e)/s(e) = 1/c$ represents an unacceptably large degree of similarity. Then one reasonable definition of the intensity with which p is a pattern in e is given by the formula

$$[1 - c\, d(r_p,e)/s(e)]\, [s(e) - s(p)] / s(e) \quad (4)$$

The term $[s(e)-s(p)]/s(e) = 1-s(p)/s(e)$ gauges the amount of simplification or "compression" provided by using p to represent e. If p provides no compression, this yields 0; in the limit where p is entirely simple (s(p)=0), the term yields its maximum value of 1 (100% compression). if, say, s(p) = .5s(e) then one has 50% compression. Next, the term $1 - c\, d(r_p,e)/s(e)$ allows for approximate matching or "lossy compression": it has a maximum of 1 when r_p, the result of carrying out process p, is exactly the same as e. The maximum intensity of a pattern, according to this formula, will be 1; and anything with an intensity greater than zero may be considered to be a pattern. The set of all processes p that are patterns in e is called the **structure** of e; it is denoted St(e) and is a fuzzy set with degrees of membership determined by Equation 1.

The simplest way to express pattern computationally is to introduce a fixed universal Turing machine which takes **two tapes** instead of the usual one: one tape for "program" and one tape for "data." In this Turing machine model, the "entities" involved in the definition of pattern are binary sequences representing data, and the "processes" are binary sequences representing programs. The simplest course in this case is to define the simplicity of a binary sequence as its length. However, this is not the only possibility. There are other factors such as the program's running time, and the "crypticity" of the program (the difficulty of discovering the sequence in the first place). Many though not all of these possibilities are discussed in *The Structure of Intelligence*.

The reader schooled in modern theoretical computer science may be curious as to the relation between this general theory of pattern and algorithmic information theory (Chaitin, 1987). In fact, this relation is not at all difficult to determine. The key is to restrict attention to the special case where the metric d is defined so that d(x,y) is infinite unless x = y. Also, in algorithmic information theory, it is conventional to assume that all computer programs are "self-delimiting," i.e. contain a segment specifying their own length. Given these assumptions, if one defines the simplicity of a binary sequence as its length, one concludes that the algorithmic information I x is the simplicity of the simplest pattern in x. A straightforward example is the binary sequence

x = 100100100100100100100100100100100 ... 100100100100100100100100100100100

consisting of 1000 repetitions of the string "100". Then we have s(p) = 200, while s(e) = 1000, and so the intensity of p as a pattern in e comes out to [1000 - 200]/1000 = .8.

It is crucial not to overlook the presence of the metric d in this equation. Algorithmic information theory assumes d away but, while this simplifies the mathematics, it is not adequate for real-world pattern recognition. Consider, for instance, the example of **grammatical categories**, which is a sort of paradigm case for categorization in general. Imagine a process that assigns words to categories, thus transforming an input stream of words into an input stream of grammatical tags (e.g. "I snipped the fly into two pieces" into "PRO V DET N PREP ADJ N"). This process definitely loses some of the structure of the input stream: namely, it loses semantics, phonology and pragmatics, retaining only grammar. But it allows one to make predictions regarding the sequence. For instance, having seen "PRO V DET N PREP" one can predict that the next word in the input stream is very unlikely to be a V, a verb. This predictive power is a consequence of the fact that the tag sequence **approximates** the input sequence, to within a certain degree of precision. The tag sequence is a potential pattern in the input sequence, in the sense that it produces something simpler than the input sequence, which approximates the input sequence. Whether the degree of approximation is adequate to render the tag sequence a pattern depends on the definition of the metric d. In the human mind the metric d gauges semantic and pragmatic similarity; it is defined implicitly by a constellation of other linguistic and nonlinguistic processes.

This kind of approximate pattern has wide relevance beyond linguistics. For example, the Chaos Language Algorithm, to be described below, recognizes grammatical patterns underlying the trajectories of dynamical systems. The patterns which it recognizes are not patterns in the strict sense of algorithmic information, but they are valid and useful patterns nonetheless.

Static, Dynamic and Static/Dynamic Patterns

In general, when studying any kind of system, one is interested in patterns which are recognizable in the tuple

Hist(S) = (S(t),S(t+1),...,S(r)) (5)

These might be called "static/dynamic patterns"; they are patterns which incorporate information about both the "static" structure of a system at a given time, and the "dynamic" structure of a system's temporal trajectory.

Next, in order to understand the **complexity** of systems, it pays to introduce the further concept of **emergence**. Let e#f denote some kind of "join" of the two entities e and f (such as, if e and f are two physical entities, the composite entity obtained by placing e and f next to each other; or if e and f are two bit strings, the result of placing e and f end to end). Then a process p is an emergent pattern between e and f to the extent

$$IN(p|e\#f) - [IN(p|e) + IN(p|f)] \quad (6)$$

Now, suppose the system S(t) is composed of a collection of "component parts," $\{S_i(t), i=1,...,N\}$. Each component part leads to its own tuple $Hist(S_i)$, and hence to its own static/dynamic patterns. A complex system is one in which a great number of emergent patterns arise as a consequence of interactions between the parts. In other words, the complexity of a system should be measured in terms of the size of

$$St(Hist(S)) - [St(Hist(S_1)) + ... + St(Hist(S_N))] \quad (7)$$

The measurement of the size of a fuzzy collection of patterns is a matter of some difficulty; one must subtract off for overlap among different patterns, and there is no entirely "fair" way to do so. This issue is discussed in *SI*.

It will also be useful to define more specific types of patterns in systems. What I call a "purely static pattern" is, quite simply, a pattern in the state S(t) of a system S at some given time t. A "dynamic pattern" is, on the other hand, a pattern which is observable in the way a system **changes**. This pattern need have nothing to do with the actual structure of the system at any given time; it must emerge solely from observations of the way in which the system changes from one time to the next. Symbolic dynamics, a mathematical tool to be introduced in Chapter Five, deals with purely dynamic pattern. Much of traditional psychology deals only with static pattern: changes are hardly considered at all. The most interesting patterns of all, however, are the static/dynamic patterns, the ones that combine change over space with change over time. And some of the most interesting static/dynamic patterns, as it turns out, are the ones that serve to relate purely static and purely dynamic patterns. It is patterns in this class which allow us to observe the state of a system at a given time, and thus predict how the system will change; or else to observe the way a system changes, and thus make reasonable estimates of the nature of the system's internal structure.

The Structure-Dynamics Principle which is proposed below is a pattern of this type: it states, quite simply, that in many very complex systems there is a large overlap between the set of purely static patterns and the set of purely dynamic patterns. In other words the Principle says that Being and Becoming overlap; the reality of pattern and structure is deeper than the reality of space and time. As we probe the nature of structure and dynamics using tools from mathematics and computing, we will find the essential philosophical nature of these concepts is never distant, and never irrelevant.

CHAPTER TWO

THE PSYNET MODEL

2.1 INTRODUCTION

There is no question that the mind/brain is extremely complex, and therefore falls within the purvey of complexity science. The real question is whether complexity science, in its current state, is up to the challenge of modelling a system as complex as the mind/brain.

I believe that the answer to this question is a guarded **yes**. In a series of publications over the past half-decade, I have constructed a novel complex systems model of mind, which I now call the **psynet model** (most notably *The Structure of Intelligence* (Goertzel, 1993; *SI*), *The Evolving Mind* (Goertzel, 1993a; *EM*) and *Chaotic Logic* (Goertzel, 1994); *CL*) . In this chapter I will review some of the most significant aspects of the psynet model, and discuss the place of the model within complexity science.

The psynet model is a simple construction, but it is fundamentally different from previous complex system models. Before embarking on a detailed description, it may therefore be useful to enumerate the key principles of the model in a concise form, bearing in mind that many of the terms involved have not been explained yet:

2. Minds are **magician systems** residing on graphs

2. The magicians involved are **pattern/process** magicians

3. Thoughts, feelings and other mental entities are

   "structural conspiracies," i.e. **autopoietic systems** within the mind magician system

4. The structural conspiracies of the mind join together

   in a complex network of attractors, meta-attractors, etc.

5. This network of attractors approximates a fractal

   structure called the **dual network**, which is structured according to at least two principles: associativity and hierarchy.

Later chapters will apply the model to particular cases, and will explore the relations between the psynet model and other complex systems models. For now, however, the emphasis will be on giving a clear and general statement of the model itself, by explicating the psychological meaning of these abstract principles.

2.3 THE DUAL NETWORK

Recall that a magician system consists, quite simply, of a collection of entities called "magicians" which, by acting on one another, have the power to cooperatively create new magicians. Certain magicians are paired with "antimagicians," magicianswhich have the power to annihilate them.

According to the psynet model, mind is a pattern/process magician system. It is a magician system whose magicians are concerned mainly with recognizing and forming patterns.

Such a system, as I have described it, may at first sound like an absolute, formless chaos. Just a bunch of magicians acting on each other, recognizing patterns in each other -- where's the structure? Where's the sense in it all?

But, of course, this glib analysis ignores something essential -- the phenomenon of mutual intercreation, or autopoiesis. Systems of magicians can interproduce. For instance, a can produce a, while b produces a. Or a and b can combine to produce c, while b and c combine to produce a, and a and c combine to produce b. The number of possible systems of this sort is truly incomprehensible. But the point is that, if a system of magicians is mutually interproducing in this way, then it is likely to **survive** the continual flux of magician interaction dynamics. Even though each magician will quickly perish, it will just as quickly be re-created by its co-conspirators. Autopoiesis creates self-perpetuating order amidst flux.

Some autopoietic systems of magicians might be unstable; they might fall apart as soon as some external magicians start to interfere with them. But others will be robust; they will survive in spite of external perturbations. These robust magician systems are what I call **autopoietic systems**, a term whose formal definition is given in the Appendix. This leads up to the next crucial idea of the psynet model: that thoughts, feelings and beliefs are autopoietic. They are stable systems of interproducing pattern/processes. In *CL*, autopoietic pattern/process magician systems are called **structural conspiracies**, a term which reflects the mutual, conspiratorial nature of autopoiesis, and also the basis of psychological autopoiesis in pattern (i.e. structure) recognition. A structural conspiracy is an autopoietic magician system whose component processes are pattern/processes.

But structural conspiracy is not the end of the story. The really remarkable thing is that, in psychological systems, there seems to be a global order to these autopoietic subsystems. The central claim of the psynet model is that, in order to form a functional mind, these structures must spontaneously self-organize into larger **autopoietic superstructures**. And perhaps the most important such superstructure is a sort of "monster attractor" called the **dual network**.

The dual network, as its name suggests, is a network of pattern/processes that is simultaneously structured in two ways. The first kind of structure is **hierarchical**. Simple structures build up to form more complex structures, which build up to form yet more complex structures, and so forth; and the more complex structures explicitly or implicitly govern the formation of their component structures. The second kind of structure is **heterarchical**: different structures connect to those other structures which are **related** to them by a sufficient number of

pattern/processes. Psychologically speaking, as will be elaborated in the following section, the hierarchical network may be identified with command-structured perception/control, and theheterarchical network may be identified with associatively structured memory.

While the dual network is, intuitively speaking, a fairly simple thing, to give a rigorous definition requires some complex constructions and arbitrary decisions. One approach among many is described in an Appendix to this chapter.

A **psynet**, then, is a magician system which has evolved into a dual network structure. Or, to place the emphasis on structure rather than dynamics, it is a dual network whose component processes are magicians. The central idea of the psynet model is that the psynet is necessary and sufficient for mind. And this idea rests on the crucial assumption that the dual network is autopoietic for pattern/process magician dynamics.

Psychological Interpretation

At first glance the dual network may seem an extremely abstract structure, unrelated to concrete psychological facts. But a bit of reflection reveals that the hierarchical and heterarchical networks are ubiquitous in theoretical psychology and cognitive neuroscience.

For instance, the whole vast theory of visual perception is a study in hierarchy: in how line processing structures build up to yield shape processing structures which build up to yield scene processing structures, and so forth. The same is true of the study of motor control: a general idea of throwing a ball translates into specific plans of motion for different body parts, which translates into detailed commands for individual muscles. It seems quite clear that there is a **perceptual/motor hierarchy** in action in the human brain. And those researchers concerned with artificial intelligence and robotics have not found any other way to structure their perceiving and moving systems: they also use, by and large, perceptual-motor hierarchies. Perhaps the best example of this is idea of subsumption architecture in robotics, pioneered by Rodney Brooks at MIT (Brooks, 1989). In this approach, one begins by constructing low-level modules that can carry simple perceptual and motor tasks, and only then constructs modules residing on the next level up in the hierarchy, which loosely regulate the actions of the low-level modules. The perceptual-motor hierarchy is created from the bottom up. Recent researchers (Churchland et al, 1994) have pointed out the importance of top-down as well as bottom-up information transmission within the visual system, and the existence of connections at all levels to regions of the non-visual brain. But these observations do not detract from the fundamental hierarchical structure of perception and action; rather, they elaborate its place in the ecology of mind.

On the other hand, the **heterarchical** structure is seen most vividly in the study of memory. The associativity of human long-term memory is well-demonstrated (Kohonen, 1988), and has been simulated by many different mathematical models. The various associative links between items stored in memory form a kind of sprawling network. The **kinds** of associations involved are extremely various, but what can be said in general is that, if two things are associated in the memory, then there is some other mental process which sees a pattern connecting them. This is theprinciple of the heterarchical, associative network.

The key idea of the dual network is that the network of memory associations (heterarchical network) is also used for perception and control (hierarchical network). As a first approximation, one may say that perception involves primarily the passing of information **up** the hierarchy, action involves primarily the passing of information **down** the hierarchy, and memory access involves primarily exploiting the associative links, i.e. the heterarchical network. But this is **only** a first approximation, and in reality every process involves every aspect of the network.

In order that an associative, heterarchical network can be so closely aligned with an hierarchical network, it is necessary that the associative network be structured into different levels **clusters** -- clusters of processes, clusters of clusters of processes, and so on. This is what I have, in *EM*, called the "fractal structure of mind." If one knew the statistics of the tree defining the hierarchical network, the fractal dimension of this cluster hierarchy could be accurately estimated (Barnsley, 1988). Alexander (1995) has argued for the neurobiological relevance of this type of fractal structure, and has constructed a number of interesting neural network simulations using this type of network geometry.

Finally, it must be emphasized that neither the hierarchical network nor the heterarchical network is a static entity; both are constantly evolving within themselves, and the two are constantly coevolving together. One of the key points of the dual network model is that the **structural** alignment of these two networks implies the necessity for a **dynamical** alignment as well. In other words, whatever the heterarchical network does to keep itself well-adjusted must fit in nicely with what the hierarchical network does to keep itself adjusted (and obviously vice versa); otherwise the two networks would be constantly at odds. It stands to reason that the two networks might be **occasionally** at odds, but without at least a basic foundation for harmonious interaction between the two, a working dual network would never be able to evolve.

2.4 EVOLUTION AND AUTOPOIESIS IN THE DUAL NETWORK

The dynamics of the dual network may be understood as a balance of two forces. There is the evolutionary force, which creates new forms, and moves things into new locations. And there is the autopoietic force, which retains things in their present form. If either one of the two forces is allowed to become overly dominant, the dual network will break down, and become excessively unstable, or excessively static and unresponsive.

Of course, each of these two "forces" is just a different way of looking at the basic magician system dynamic. Autopoiesis is implicit in all attractors of magician dynamics, and evolutionary dynamics is a special case of magician dynamics, which involves long transients before convergence, and the possibility of complex strange attractors.

Memory Reorganization as Evolution

Many theorists have expressed the opinion that, in some sense, ideas in the mind evolve by natural selection. Perhaps the most eloquent exposition of this idea was given by Gregory Bateson in his *Mind and Nature* (1980). The psynet model provides, for the first time, a rigorous analysis of the evolution/thought connection.

The largest obstacle that must be overcome in order to apply evolution theory to the mind is the problem of the definition of **fitness**. Natural selection is, in Herbert Spencer's well-worn phrase, "survival of the fittest." When considering specific cases, biologists gauge fitnesses with their own common sense. If animal A runs faster than its predator, but animal B does not, then all else equal animal A is fitter -- no one needs a formal definition to tell them that. The problem is getting a handle on **fitness in general**. As the saying goes, if one cannot define fitness in any way besides reproductive success, then what one has is just survival of the survivors. And, more to the point, if one cannot define fitness in any way besides case-by-case special pleading, then what one has is a very inelegant theory that cannot be easily generalized to other contexts.

One way around this problem, I have proposed, is to measure fitness in terms of **emergent pattern**. In *EM*, I define the **structural fitness** of an organism as the size of the set of patterns which synergetically emerge when the organism and its environment are considered **jointly**. If there are patterns arising through the **combination** of the organism with its environment, which are not patterns in the organism or the environment individually, then the structural fitness is large. Perhaps the easiest illustration is camouflage -- there the appearance of the organism resembles the appearance of the environment, generating the simplest possible kind of emergent pattern: repetition. But symbiosis is an even more convincing example. The functions of two symbiotic organisms match each other so effectively that it is easy to predict the nature of either one from the nature of the other.

The claim is not that structural fitness is all there is to biological fitness; it is merely that structural fitness is an important component of biological fitness. Suppose one says that a system "evolves by natural selection" if, among the individuals who make it up, **reproductive success** is positively correlated with **fitness**. Then, if one accepts the claim that structural fitness is an important component of fitness, one way to show this is to show this is to show that reproductive success is positively correlated with structural fitness.

Using this approach, it is easy to see that ecosystems and immune systems both evolve by natural selection (see *EM*). And, according to the principles outlined above, it is clear that psynets do as well. Consider: the "environment" of a process in the psynet is simply its neighbors in the network. So the structural fitness of a process in the psynet is the amount of pattern that emerges between itself and its neighbors. But, at any given time, the probability of a process **not being moved** in the network is positively correlated with its degree of "fit" in the associative memory. This degree of fit is precisely the thestructural fitness! So, survival in current position is correlated with structural fitness with respect to immediate environment; and thus, according to the definitions given, the psynet evolves by natural selection.

According to this argument, the "individuals" which are surviving differentially based on fitness are, at the lowest level, individual magicians, individual mental processes. By the same logic, clusters of magicians may also be understood to evolve by natural selection. This observation leads up to a sense in which the psynet's evolutionary logic is different from that which one sees in ecosystems or immune systems. Namely, in the psynet, every time a process or cluster is moved in accordance with natural selection, certain processes on higher levels are being crossed over and/or mutated.

In ecosystem evolution, the existence of "group selection" -- evolution of populations, species or higher taxa -- is a matter of contention. In psynet evolution, because of the presence of the hierarchical network, there is no cause for controversy. Higher-order individuals can explicitly represent and control groups, so that the distinction between groups and individuals is broken down. Group selection is a form of individual selection. In this sense, it would seem that the psynet uses natural selection much more efficiently than other evolving systems, such as ecosystems or immune systems. While ecosystems can, at best, carry out higher-order evolution on a very slow scale, psynets can carry out low-order and higher-order evolution almost simultaneously. This striking conclusion cries out for mathematical and computational investigation.

Evolution and Creativity

We have been discussing evolution as a means for maintaining the structure of the associative memory network. However, evolution also has different psychological function. Namely, it comes up in regard to the **creativity** of mental process networks. This is where the computational experiments to be described Chapter Six are intuitively relevant. They show the ability of the genetic algorithm (a computational instantiation of evolution) to produce interesting new forms.

The genetic algorithm consists of a population of entities, which repeatedly mutate and cross over with each other to produce new entities. Those entities which are "fitter" are selected for reproduction; thus the population as a whole tends to assume forms determined by the fitness criterion being used. Typical genetic algorithm experiments are aimed at finding the one correct answer to some mathematical problem. In most of the experiments in Chapter Six, however, the goal is to use the creative potential of **whole populations**, rather than merely using a population as a means to get to some "optimal" guess. This is precisely what, in the psynet model, is done by the mind's intrinsic "evolution." The complex forms created by evolving mental processes are vastly more complex than the simple pictures and melodies evolved in my experiments; on an abstract level, however, the principle is the same.

The genetic algorithm, in a psychological context, must be understood as an approximation of the activity of **subnetworks** ofthe dual network. Subnetworks are constantly mutating as their component processes change. And they are constantly "crossing over," as individual component interactions change in such a way as to cause sub-subnetworks to shift their allegiance from one subnetwork to another. This dynamic has been discussed in detail in *The Evolving Mind*.

And what is the relation between this genetic-algorithm-type creativity, in the hierarchical network, and the evolutionary reorganization of the heterarchical network, the associative memory? The answer is very simple: they are the same! When memory items move around from one place to another, seeking a "fitter" home, they are automatically reorganizing the hierarchical network -- causing subnetworks (mental "programs") to cross over and mutate. On the other hand, when processes switch their allegiance from one subnetwork to another, in a crossover-type process, their changing pattern of interaction consitutes a changing environment, which changes their fitness within the heterarchical network. Because the two networks are one, the two kinds of evolution are one. GA-style evolution and ecology are bound together very tightly, much more tightly than in the case of the evolution of species.

Autopoiesis and Thought

But evolution is not the only kind of dynamics in the dual network. In order to achieve the full psynet model, one must envision the dual network, not simply as an hierarchy/heterarchy of mental processes, but also as an hierarchy/heterarchy of evolving **autopoietic process systems**, where each such systems is considered to consist of a "cluster" of associatively related ideas/processes. Each system may relate to each other system in one of three different ways: it may contain that other system, it may be contained in that other system, or it may coexist side-by-side with that other system. The dual network itself is the "grand-dad" of all these autopoietic systems.

Autopoiesis is then seen to play an essential role in the dynamics of the dual network, in that it permits thoughts (beliefs, memories, feelings, etc.) to persist even when the original stimulus which elicited them is gone. Thus a collection of thoughts may survive in the dual network for two reasons:

-- a usefulness relative to the hierarchical control structure, i.e. a usefulness for the current goals of the organism;

-- autopoiesis

As is shown in *Chaotic Logic*, this line of reasoning may be used to arrive at many specific conclusions regarding systems of thought, particularly belief systems. For purposes of illustration, two such conclusions may be worth mention here:

-- that many belief systems considered "poor" or "irrational" have the property that they are sustained primarily by the latter method. On the other hand, many very useful and sensible belief systems are forced to sustain themselves by autopoiesis for certain periods of time as well. System theory clarifies but does not solve the problem of distinguishing "good" from "poor" belief systems.

-- that one of the key roles of autopoietic systems in the dual network is to serve as a "psychological immune system," protecting the upper levels of the dual network from the numerous queries sent up from the lower levels.

Stability means that a system is able to "absorb" the pressure put on it by lower levels, instead of constantly passing things along to the levels above it. Strong parallels exist between the dynamics of antibody classes in the immune system and the dynamics of beliefs in an autopoietic system, but we will not explore these parallels here.

So, in the end, after thinking about the dual network as an emergent **structure**, one inevitably returns to the dynamic point of view. One sees that the whole dual network is just another autopoietic system which survives by the same two methods: structural conspiracy and external utility. Even if one begins with a fairly standard information-processing picture of mind, such as the master network, one eventually winds up with an "anything-goes" autopoietic-systems

viewpoint, in which a successful mind, like a successful thought system, is one which perpetually and usefully **creates itself**.

2.5 LANGUAGE AND LOGIC

Next, shifting gears somewhat, let us turn from evolution to **language**. Language is the focal point of much modern philosophy and cognitive science. It is commonly cited as a distinguishing feature of intelligent systems. What does the psynet model have to say about the linguistic mind?

A language, as conceived in modern linguistics, is a **transformation system**. It is a collection of transformations, each equipped with its own set of rules regarding which sorts of entities it can be applied to in which situations. By applying these rules, one after the other after the other, to elements of the "deep structure" of thought, sentences are produced. In terms of the dual network, each of these transformation rules is a process with a certain position in the network; and sentences are the low-level result of a chain of information transmission beginning with a high-level structure or "idea."

In the case of the language of mathematics, the transformation rules are very well understood; this is the achievement of the past 150 years of formal logic. In the case of natural languages, our understanding is not yet complete; but we do know a handful of general transformation rules (e.g. Chomsky's famous "move-alpha"), as well as dozens upon dozens of special-case rules.

But this formal syntactic point of view is not enough. A set of transformation rules generates an incredible number of possible sentences, and in any given situation, only a miniscule fraction of these are appropriate. A system of transformation rules is only useful if it is amenable to reasoning by analogy -- if, given a reasonable set of constraints, the mind can -- by implementing analogical reasoning -- use the system to generate something satisfying those constraints. In other words, roughly speaking, a transformation system is only useful if **structurallysimilar sentences have similar derivations**. This "principle of continuous compositionality" is a generalization of Frege's (1893) famous principle of compositionality. It appears to hold true for natural languages, as well as for those branches of mathematics which we have studied to date.

This has immediate implications for the theory of **formal semantics**. When one hears the phrase "the mathematics of meaning," one automatically thinks of the formal-semantic possible-worlds approach (Montague, 1974). But though it was brilliant in its time, it may well be that this approach has outlived its usefulness. The theory of computation suggests a more human and intuitive approach to meaning: the meaning of an entity is the fuzzy set of patterns that are "related" to it by other patterns. If one accepts this view of meaning, then the connection between syntax and semantics becomes very simple. A useful transformation system is one in which structurally similar sentences have similar derivations, and two sentences which are structurally similar will have similar meanings. So a useful transformation system is one in which sentences with similar meanings have similar derivations.

It is not hard to see that continuous compositionality is exactly what is required to make a language naturally representable and usable by the dual network. The key point is that, by definition, statements with similar meanings are related by common patterns, and should thus be

stored near one another in the memory network. So if a transformation system is "useful" in the sense of displaying continuous compositionality, it follows that statements stored near each other in the memory network will tend to have similar derivations. This means that the same "derivation process," using the same collection of strategies, can be used for deriving a whole group of nearby processes within the network. In other words, it means that useful transformation systems are tailor-made for the superposition of an associative memory network with an hierarchical control network containing "proof processes." So, continuous compositionality makes languages naturally representable and learnable in the dual network. It is what distinguishes natural languages from arbitrary formal languages. As briefly argued in *Chaotic Logic*, this analysis has deep implications for the study of language learning. Language acquisition researchers are conveniently divided into two camps: those who believe that inborn knowledge is essential to the process of language learning, and those who believe that children learn language "on their own," without significant hereditary input. The psynet model does not resolve this dispute, but it does give a relevant new perspective on the processes of language learning.

In the language acquisition literature there is much talk of the "bootstrapping problem," which essentially consists of the fact that the different aspects of language are all interconnected, so that one cannot learn any particular part until one has learned of all the other parts. For instance, one cannot learn the rules of sentence structure until one has learned the parts of speech; but how does one learn the parts of speech, except by studying the positions of words in sentences? From the psynet perspective the bootstrapping problem is noproblem whatsoever; it is simply a recognition of the autopoietic nature of linguistic systems. Language acquisition is yet another example of convergence to a structural conspiracy, an autopoietic system, a strange attractor of pattern/process magician dynamics.

The question of innateness is thus reformulated as a question of the size of the basin of the language attractor. If the basin is large enough then no innate information is needed. If the basin is too small then innate information may be needed in order to be certain that the child's learning systems **start off** in the right place. We do not presently know enough about language learning to estimate the basin size and shape; this exercise has not even been carried out for formal languages, let alone natural languages.

Overcoming the "Paradoxes" of Logic

It is interesting to apply this analysis of language to the very simple language known as Boolean logic. When Leibniz invented what is now called "Boolean logic" -- the logic of **and**, **or** and **not** -- he intended it to be a sort of language of thought. Mill, Russell, and many recent thinkers in the field of artificial intelligence have pursued the same intuition that much thought is just the solution of Boolean equations. But many problems stand in the way of this initially attractive idea.

For example, there are the "paradoxes of implication." According to Boolean logic, "A implies B" just means "either B is false, or A is true." But this has two unsavory consequences: a false

statement implies everything, and a true statement is implied by everything. This does not accord very well with our intuitive idea of implication.

And there is Hempel's paradox of confirmation. According to Boolean logic, "All ravens are black" is equivalent to "All non-black entities are non-ravens." But then every piece of evidence in favor of the statement "All non-black entities are non-ravens" is also a piece of evidence in favor of the statement "All ravens are black." But this means that when we observe a white goose, we are obtaining a piece of evidence in support of the idea that all ravens are black -- which is ridiculous!

All of these paradoxes are easily avoided if, rather than just hypothesizing that the mind uses Boolean logic, one hypothesizes that the mind uses Boolean logic **within the context of the dual network**. As an example, let us consider one of the paradoxes of implication: how is it that a false statement implies everything? Suppose one is convinced of the truth both of A and of the negation of A, call it not-A. How can one prove an arbitrary statement B? It's simple. The truth of A implies that either A is true, or B is true. But the truth of not-A then implies the truth of both not-A, **and** either A or B. But on the other hand, if not-A is true, and either A or B is true, then certainly B must be true.

To put it less symbolically, suppose I love mom and I hate mom. Then surely either I love mom or cats can fly -- after all, I love mom. But I hate mom, so if either I love mom or cats can fly, then obviously cats can fly.

So what, exactly, is the problem here? This paradox datesback to the Scholastic philosophers, and it hasn't obstructed the development of mathematical logic in the slightest degree. But from the point of view of psychology, the situation is absurd and unacceptable. Of course a person can both love and hate their mother without reasoning that cats can fly.

The trick to avoiding the paradox is to recognize that the psynet is primary, and that logic is only a tool. The key step in the deduction of B from "A and not-A" is the formation of the phrase "A or B." The dual network, using the linguistic system of Boolean logic in the manner outlined above, simply will not tend to form "A or B" unless **A and B are related by some pattern**. No one ever thinks "either I love mom or cars can fly," any more than they think "either I love mom or planes can fly." So the dual network, using Boolean logic in its natural way, will have a strong tendency not to follow chains of reasoning like those required to reason from a contradiction to an arbitrary statement.

But what if some process within the dual network, on an off chance, **does** reason that way? Then what? Will this contradiction-sensitivity poison the entire dual network, paralyze its reasoning functions? No. For a process that judges every statement valid will be **very poor at recognizing patterns**. It will have no clue what patterns to look for. Therefore, according to the natural dynamics of the multilevel network, it will rapidly be eliminated. This is natural selection at work!

This is a very partial view of the position of logic in the dual network -- to complete the picture we would have to consider the other paradoxes mentioned above, as well as certain other

matters. But the basic idea should be clear. The paradoxes of Boolean logic are fatal only to Boolean logic as an isolated reasoning tool, not to Boolean logic as a device implemented in the context of the psynet. In proper context, the species of linguistic system called logic is of immense psychological value.

2.6 PSYNET AI

The psynet model was originally conceived as a kind of "abstract AI" -- an AI theory based, not on the speed and memory limitations of current computers, nor on the mathematical tools of formal logic, but on the introspectively and experimentally observed structures of mind itself. Subsequent development of the model has taken a more philosophical and psychological turn, but the model is still computational at the core. Given this background, it is hardly surprising that the psynet model should have significant implications for AI.

In fact, the model's AI implications are more radical than might be perceived at first glance. While the typical pattern in AI is for cognitive theory to be driven by computational experiment, the psynet model represents an attempt to move in exactly the opposite direction, from theory to experiment: to begin with a general, well fleshed-out model of computational psychology, and then formulate computational experiments based on this model. The SEE model briefly discussed above is a simple example of this approach.

The psynet approach to AI might be called a "scaling-down"approach, as opposed to the standard "scaling-up" approach which assumes that one can construct simple computational models and then **scale up** to obtain a complex, intelligent system. To fully appreciate the radical nature of the scaling-down approach, a little historical background is needed.

Early AI researchers, as has been amply documented (Dreyfus, 1993), vastly overestimated the ease of generalization. They produced programs which displayed intelligence in very limited domains -- e.g. programs which were good at playing chess, or recognizing letters, or doing calculus problems, or moving blocks around in a room. In this way, they believed, they were constructing intelligent algorithms, which would then, with a little tinkering, be able to turn their intelligence to other problems. This is not an unreasonable notion; after all, teaching a person chess or calculus improves their general powers of thought; why shouldn't the same be true of a computer? But in fact these classic AI programs were idiot savants. The programs worked because they embodied rules for dealing with specific situations, but they never achieved the ability to come into a new situation and infer the appropriate rules. The assumption was that reasoning ability would **scale up** from micro-worlds to the real macro-world in which we live, or at least to artificially constructed macro-worlds containing numerous intersecting sub-environments. But this assumption proved disastrously wrong.

Over the last ten to fifteen years, the connectionist paradigm has breathed new life into artificial intelligence. Even more recently, the theory of genetic algorithms has begun to direct AI research down a new path: the achievement of intelligence by simulating **evolutionary** processes rather than brain processes or reasoning processes. But it is not hard to see that, exciting as they are, these new ideas fall prey to the same basic fallacy as old-style AI. The programs are designed to perform well on toy problems; it is then assumed that whatever works

on the toy problems will "scale up" to deal with the real problems confronted by actual intelligent systems. But this assumption of **scale-uppability** contradicts the available evidence. Anyone who has worked with neural nets or genetic algorithms knows that, to get any practical use out of these constructs, a great deal of intelligent human planning and intervention is needed. One must first figure out how to best **represent** the data in order to present it to the program. There is a huge gap between generalization ability in simple, appropriately represented domains and generalization ability across a variety of complex, un-preprocessed domains. The former does not "scale up" to yield the latter.

Classical AI depends on scaling up from rule-based learning in a "microworld" to rule-based learning in the macroworld. Connectionist and GA-based AI depend on scaling up from localized generalization to reality-wide generalization. But according to the psynet model, the most important aspect of an intelligent system is precisely that aspect which cannot be inferred by the "scaling-up" method: the **overall structure**. To get psychologically meaningful AI applications, one must scale down from the appropriate overall structure, instead of blindly counting on scaling-up.

Developing Psynet Applications

What sorts of questions are involved in actually developing AI applications based on the the psynet model? There are two important decisions to be made: the degree of static structure to be built in, and the nature of the component pattern/ processes.

The first question returns us to the dynamic and static views of the psynet. The static approach to psynet AI begins with a dual network data structure and populates this structure with appropriate pattern/processes. Each node of the network is provided with a "governor" processor that determines the necessity for mutation and swap operations, based on the success of the processes residing at that node. If the processes have done what was expected of them by the higher-level processes which guide them, then little mutation of subprocesses, and little swapping of subprocess graphs, will be required. But if the processes have left a lot of higher-level expectations unfulfilled (i.e., according to the ideas given above, if they have generated a large amount of emotion), then mutation and swapping will be rampant.

The dynamic approach, on the other hand, begins with pattern/processes interconnected in an arbitrary way, and, instead of imposing a dual network structure on these processes, relies on autopoietic attraction to allow the dual network to emerge. The distinction between these two approaches is not a rigid one; one may begin with more or less "dual network like" graphs. At this point, there is no way of determining the point on this continuum which will lead to the most interesting results.

The second decision, how to implement the pattern/processes, leads to an even greater variety of possibilities. Perhaps the simplest viable options are bit string processes (a la the GA), Boolean processes, and **repetition** processes, processes which recognize repeated patterns in external input and in one another. The latter, one suspects, might be useful in text processing applications.

These simple pattern/process options all lead to architectures that are fairly regimented, in the sense that all the processes have the same basic form. However, this kind of regimentation is not in any way implicit in the psynet model itself. The only substantial restrictions imposed by the model are that: 1) processes must be able to recognize a wide variety of patterns, and 2) processes must be able to act on each other with few limitations (i.e. they must form a "typeless domain," or a good approximation thereof).

In the long term, it will certainly be interesting to construct psynets involving a similar variety of pattern/processes. But we may also be able to do a great deal with regimented architectures; at present this is difficult to predict. Interestingly, even the degree of regimentation of the human brain is a matter of debate. On the one hand, the brain contains all sorts of specialized pattern-recognition processes; the processes relating to visual perception have been studied in particular detail. On the other hand, as mentioned above, Edelman (1988) has argued that these complex processes are allbuilt as different combinations of a fairly small number of repeated, functionally equivalent neuronal groups.

Psynets and the Darwin Machine Project

One possible route toward psynet AI is the Simple Evolving Ecology (SEE) model, which will be discussed in a later chapter, once more background on the genetic algorithm has been provided. Another possibility involves the "Darwin Machine" project, currently being carried out by the Evolutionary Systems Department at ATR Human Information Processing Research Laboratories in Japan, under the supervision of Katsunori Shimohara. Shimohara (1994) describes a three-phase research programme, the ultimate result of which is intended to be the construction of an artificial brain. The philosophy of this programme is not to simulate the precise workings of the brain, but rather to use methods from evolutionary computation to construct a device that works **better** than the brain. The first phase is an exploration of techniques for evolution of software and hardware. The second phase is a construction of a "Darwin Chip" and "Darwin Machine" which will incorporate these techniques, thus moving evolutionary learning from the software level to the hardware level. The third phase, still almost entirely speculative, is a re-implementation of the Darwin Machine using nanotechnology. This third phase, it is felt, will produce a Darwin Machine of sufficient complexity to support true intelligence, an "Artificial Brain System."

The Darwin Machine project would seem to be an excellent testing ground for the psynet model. If the Psynet Conjecture is correct, then the construction of an artificial brain **cannot** succeed unless the psynet structure is adhered to. On the other hand, the evolutionary methodology which Shimohara advocates is ideally suited for the psynet model -- a fact which should be fairly clear from the brief discussion of evolution in the psynet given above, and even clearer from the more detailed discussion of evolutionary computation in the dual network given in *EM*.

Perhaps the most interesting immediate goal of Shimohara's group is the CAM-Brain Project (a CAM, or Cellular Automata Machine, is a special kind of hardware designed for parallel simulation of cellular automata):

The aim of the CAM-Brain Project is to build (i.e. grow/evolve) an artificial brain by the year 2002. This artificial brain should initially contain thousands of interconnected artificial neural network modules, and be capable of controlling approximately 1000 "behaviors" in a "robot kitten." Using a family of CAM's, each with its own processor to measure the performance quality or fitness of the evolved neural circuits, will allow the neural modules and their interconnections to be grown and evolved at electronic speeds.

The psynet model makes a very specific suggestion about this project. Namely, it suggests that the project will succeed in producing a reasonably intelligent kitten if and only if:

-- many of the neural network modules are used as pattern-recognition processes (pattern/processes)

-- the network of modules is arranged in a dual network structure

Because of the relative simplicity of a robot kitten, as opposed to, say, a human brain, one cannot call these suggestions "predictions." The absence of a psynet structure in a robot kitten would not constitute a disproof of the psynet model. But on the other hand, if it were empirically demonstrated that a psynet **is** necessary for intelligence in a robot kitten, this would certainly constitute a strong piece of evidence in favor of the psynet model.

In fact, depending on the level of intelligence of this kitten, many of the more specific phenomena discussed above can be expected to show up. For instance, the psynet resolution of the paradoxes of logic should be apparent when the kitten learns causative relationships. A very simple version of continuous compositionality may be observable in the way the kitten responds to combinations of stimuli in its environment. The evolution by natural selection of subnetworks should be obvious every time the kitten confronts new phenomena in its environment. The autopoiesis of belief systems should be apparent as the kitten retains a system for reacting to a certain situation even once the situation has long since disappeared. Even the emergence of dissociated self-systems, as will be discussed in later chapters, could probably be induced by presenting the kitten with fundamentally different environments, say, on different days of the week.

2.7 CONCLUSION

The complexity of the mind, I have argued, does not prohibit us from obtaining a unified understanding. Rather than interpreting the mind's complexity as an obstacle, one may take it as a challenge to the imagination, and as an invitation to utilize ideas from complex systems science. One may seek to provide a theory of sufficient simplicity, flexibility and content to meet the complexity of psychological systems head on.

The psynet model is not at bottom a "synthesis"; it is a simple entity with its own basic conceptual coherence. However, because of its focus on interdependences, the model is

particularly amenable to study using the various overlapping tools of complex systems science. In particular it leads one to think of psychological phenomena in terms of:

-- evolution by natural selection as a general mechanism of form creation

-- universal computation, as a foundation for the study of information and pattern

-- dynamical systems theory -- the concept of attractors and convergence thereto

-- autopoiesis, or self-organization combined with self-production

-- agent-based modeling, as a way of bridging the gap between connectionist and rule-based modeling

Summing up, we may ask: What is the status of the psynet model, as a model of mind? To make a fair assessment, one must consider the model in three different lights: as mathematics, as theoretical neuroscience, and as psychology.

Mathematically, one can argue convincingly that the dual network is indeed an attractor of the cognitive equation, of pattern/process magician dynamics. The proof is inductive: one shows that if a given set S of processes are all attractors of the cognitive equation, and one arranges the elements of S in a flat heterarchical network M, with process systems recognizing emergent patterns between elements of S in a flat heterarchical network N supervising over M, then the combined network of M and N is still an attractor of the cognitive equation. According to this theorem, if one undertakes to model the mind as a magician system (or, more generally, agent system), then the dual network is one configuration that the mind might get itself into. The possibility of other, non-dual-network attractors has not been mathematically ruled out; this is an important open question.

In the next chapter I will deal with the applications of the psynet model to neuroscience. The psynet-brain connection has been mentioned repeatedly in previous publications, but has never before been systematically pursued. It turns out that the psynet model matches up quite naturally with what is known about the structure of the cortex, and provides a handy platform for exploring various questions in cognitive neuroscience.

Finally, in the area of psychology, a great deal of work has been done attempting to demonstrate the ability of the psynet model to account for human mentality. In *SI*, induction, deduction, analogy and associative memory are analyzed in detail as phenomena of pattern/process dynamics. In *EM*, the parallels between Neural Darwinism, evolutionary ecology, and the psynet model are illustrated, the point being to demonstrate that the psynet model is capable of accounting for the evolutionary and creative nature of human thinking. In *CL*, a psynet-theoretic account of language is given, and personal and scientific belief systems are analyzed in terms of autopoietic attraction in mental process systems. Of course, all the phenomena touched on in this work may be explained in many different ways. The point is that the psynet model gives a simple, unified explanation for a wide variety of psychological

phenomena, in terms of complexity-science notions such as algorithmic pattern, attractors, agent systems and adaptive evolution.

The task of developing and validating such a complex model is bound to be challenging. But this difficulty must be weighed against the immense potential utility of a unified theory of biological and computational intelligence. The explorations summarized here and in the companion paper indicate that, at very least, the psynet model shows promise in this regard.

APPENDIX 1: THE PSYNET MODEL AS MATHEMATICS

The task of this Appendix is to formalize the statement that the psynet model is an accurate model of mind, thus turning the psynet model into a purely mathematical hypothesis. The firststep toward this end is to define what I mean by "mind." Having done this, it is not difficult to give various rigorous formulations of the statement that the psynet models mind.

In *SI* a simple "working definition" of mind is given: a mind is the structure of an intelligent system (i.e. the fuzzy set of patterns in an intelligent system). One may also give more sophisticated definitions, e.g. by weighting the degree of membership of each pattern in the mind according to some measure of its relevance to intelligent behavior. Intelligence is then defined as "the ability to achieve complex goals in difficult-to-predict environments." A mind is, therefore, the structure of a system that can achieve complex goals in unpredictable environments. These definitions obviously do not solve the numerous philosophical and psychological problems associated with mind and intelligence. But, in the spirit of all mathematical definitions, they do give us something to go on.

The terms in this definition of intelligence may all be defined precisely. For instance, put simply, an environment is said to be difficult to predict if it couples unpredictability regarding precise details (e.g. high Liapunov exponent) with relative predictability regarding algorithmic patterns (i.e. a moderate correlation between patterns inferrable prior to time t and patterns inferrable after time t; and between patterns inferrable at one spatial location and patterns inferrable at another spatial location). Similarly, a goal, formalized as a function from environment states to some partially ordered space representing outcome qualities, is said to be complex if it couples unpredictability regarding precise details (e.g. high Lipschitz constant) with relative predictability regarding algorithmic patterns (i.e. moderate correlations between patterns inferrable from one part of the functions's graph and patterns inferrable from another).

Given this definition of intelligence, one may give the following formalization of the statement that the psynet models mind (this is a rephrasing of an hypothesis given in the final chapter of *SI*):

-- Psynet Hypothesis: A system displays general intelligence if and only if it displays the psynet as a prominent algorithmic pattern (where intelligence and pattern are defined according to the same model of computation).

In other words, what this says is that a system is intelligent only if it has the psynet as a substantial part of its mind. A proof (or disproof!) of this conjecture has proved difficult to come

by. However, it is possible to unravel the conjecture into simpler conjectures in a way that provides at least a small amount of insight. For instance, suppose one narrows the focus somewhat, and instead of general systems considers only appropriate magician systems. Then the Psynet Hypothesis suggests the following, somewhat more approachable, hypothesis:

-- Probable Producibility Hypothesis: a large dual network is a wide-basined attractor of the cognitive equation.

If one removes the word "wide-basined" from this conjecture, one obtains the statement that a large dual network is an autopoietic system under the cognitive equation; a slightly weaker conjecture which in *CL* is called the ProducibilityHypothesis. The conjecture formulated here claims not only that the dual network stably produces itself, but also that, if one starts a magician system from an arbitrary initial condition, it is reasonably likely to self-organize into a dual network.

These hypotheses give rise to a variety of possibilities. If the Producibility Hypothesis were to be proved false, then the psynet model would have to be abandoned as fundamentally unsound. On the other hand, if the Producibility Hypothesis were to be proved **true**, then the problem of validating the psynet model as a model of human and non-human intelligence would still remain. But at least the internal consistency of the model would not be in question. The psynet would be a demonstrably viable cognitive structure.

If the Probable Producibility Hypothesis were proved false, while the weaker Producibility Hypothesis were proved true, this would validate the dual network as a viable cognitive structure, but would raise a serious problem regarding the **evolution** of the dual network. One would have to assume that the evolution of the dual network had begun from a very special initial condition; and, in working out AI applications, one would have to be certain to choose one's initial condition carefully.

Finally, if the Probable Producibility Hypothesis were proved true, then this would validate the dual network as a viable cognitive model, and would also verify the ease of arriving at a dual network structure from an arbitrary adaptive magician system.

These statements give a rigorous way of approaching the claim that the psynet model is a valid psychological model. However, they do not address the question of **other** cognitive models. This question is dealt with by the following

-- Exclusivity Hypothesis: There are no abstract structures besides the dual network for which the Probable Producibility Hypothesis is true.

Howver, a proof of this appears at least as difficult as a proof of the Psynet Hypothesis.

APPENDIX 2: FORMAL DEFINITION OF THE DUAL NETWORK

While the concept of the dual network is intuitively quite simple, it supports no simple formalization. There are many different ways to formalize the same basic idea. What follows is one approach that seems particularly reasonable.

Define a geometric magician system, or **GMS**, as a graph labeled with a collection of magicians at each node. The concept of a dual network may then be formalized in terms of the notion of a **fuzzy subset** of the space of geometric magician systems. Let k be a small integer, to be used for gauging proximity in a GMS.

Where G is a GMS, define the heterarchicality of G, het G, as the average over all nodes N of G of the quantity v/|G|, where v is the amount of emergent pattern recognized by the k-neighbors of the residents of N, between the residents of node N and the residents of k-neighboring nodes.

This gauges the degree to which G is "associative" in thesense of instantiating patterns in its structure.

Next, define a stratified geometric magician system or **SGMS** as a GMS in which each node has been assigned a certain integer "level." Where H is an SGMS, define the hierarchality of H, hier H, as the product wx/|H|*2*, where

-- w is the total weight of those magician interactions appearing in F[R[H]] which involve a magician on level i acting on a magician of level i-1 to produce a magician on level i-2.

-- x is the total amount of emergent pattern recognized by magicians on some level i amongst magicians on level i-2.

The quantity w gauges the degree to which H instantiates the principle of hierchical control, the quantity x gauges the degree to which H demonstrates the hierarchical emergence of increasingly complex structures, and the quantity hier H thus measures the degree to which H is a valid "hierarchical network."

The degree to which a geometric magician system G is a dual network may then be defined as the product of the hierarchality and heterarchality of G.

All this formalism speaks of GMS's. To transfer it to ordinary magician systems, we must define a GMS G to be **consistent** with a magician system M if the magician population of G is the same as the magician population of M, and the interactions permitted by the magician dynamic of M do not include any interactions that woudl be forbidden in G. The degree to which a **magician** system M is a dual network is the maximum, over all graphs with which M is consistent, of the GMS obtained by associating M with that graph.

What this exceedingly unsightly definition says is, quite simply, that the degree to which a magician system is a dual network is the degree to which this system may be understood as a combination of hierarchical and heterarchical geometric magician systems. The difficulty of expressing this idea mathematically is an indication of the unsuitability of our current mathematical language for expressing psychologically natural ideas.

CHAPTER THREE

A THEORY OF CORTICAL DYNAMICS

3.1 INTRODUCTION

The psynet model, as outlined above, is an abstract model of the dynamics of **mind**. It is not a model of the dynamics of the **brain**. This is an important distinction, both methodologically and conceptually. Mind is not brain; mind is, rather, a collection of patterns emergent in the brain. Mind is a mathematical entity; a collection of relations, not an actual physical entity.

It is clear that, in modern mind theory, psychology and neuroscience must proceed together. But even so, given the tremendous rate of change of ideas in neuroscience, it seems foolish to allow one's psychological models to be dictated too precisely by the latest discoveries about the brain. Rather, one must be guided by the intrinsic and elegant structure of thought itself, and allow the discoveries of neuroscience to guide one's particular ideas within this context.

An outstanding example of this point is neural network modelling. Neural network models take an idea from neuroscience -- the network of neurons exchanging charge through synapses -- and elevate it to the status of a governing psychological principle. Many notable successes have been obtained in this way, both psychologically and from an engineering point of view. However, the problem of connecting the states and dynamics of neural networks with **mental** states and dynamics has never really been solved. Neural networks remain a loose, formal model of the brain, with an uncertain, intuitive connection to the mind itself.

The advantage of neural network models is that they allow one to import the vocabulary of dynamical systems theory into the study of the brain. One can talk about attractors of various kinds, attractor basins, and so forth, in a rigorous and detailed way. The idea of thoughts, memories and percepts as attractors is given a concrete form. However, in a sense, the representation is **too** concrete. One is forced to understand deep, fuzzy, nebulous patterns of mind in terms of huge vectors and matrices of neural activations and synaptic weights.

In this chapter I will present an alternative approach to brain modelling, based on the psynet model. Instead of looking at the brain and abstracting a model from brain structure, I will look at the brain through the lense of the psynet model and ask: What structures and dynamics in the brain seem to most naturally give rise to the structures posited in the psynet model? The resulting theory is quite different from standard neural network models in its focus on **overall structure**. And it is different from standard, non-connectionist AI in its focus on self-organization and emergence. Rather than specifying the update rules of the neurons and synapses, one is specifying the overall emergent structure of the autopoietic system.

In particular, what I will present here is a theory of **cortical structure**. The brain is a highly complex system -- it is complex on many different levels, from overall architecture to neuronal wiring to biochemical dynamics. However, it is possible to single out certain aspects of neural

complexity as being more important than others. What differentiates the human brain from the primate brain is, above all, our greatly enlarged neocortex. Elucidation of the workings of the cortex would thus seem to be a particularly important task.

Thanks to recent advances in neurobiology, we now know a great deal about the structure and function of the cortex. What we do not know, however, is **what cortex computes**. Or, to put it in different terminology, we do not know the **dynamics** of the cortex. On a very crude level, it is clear that the cortex is largely responsible for such things as high-level pattern recognition, abstract thought, and creative inspiration. But **how** this particular configuration of brain cells accomplishes these fantastic things -- this is the sixty-four million dollar question. This is the question that I will attempt to answer here.

3.2 NEURONS AND NEURAL ASSEMBLIES

In this section I will present a few basic facts about the structure and function of the brain. The goal is to give the minimum background information required to understand the model of cortex to be presented. For an adequate review of neuroanatomy and neurochemistry, the reader is referred elsewhere. An excellent, if dated, overview of the brain may be found in (Braitenberg, 1978); a more modern and thorough treatment may be found in any number of textbooks. Finally, the comprehensive reference on cognitive neuroscience is (Gazzaniga, 1995).

First, the **neuron** is generally considered the basic unit of brain structure. Neurons are fibrous cells, but unlike the fibrous cells in other body tissues such as muscles or tendons, they have a tendency to ramify and branch out. The outer cell membrane of a neuron is shaped into extensive branches called **dendrites**, which receive electrical input from other neurons; and into a structure called an **axon** which, along its main stalk and its collateral ramifications, sends electrical output to other neurons. The gap between the dendrite of one neuron and the axon of another is called the **synapse**: signals are carried across synapses by a variety of chemicals called **neurotransmitters**. There is also a certain amount of diffusion of charge through the cellular matrix. The dynamics of the individual neuron are quite complex, but may be approximated by a mathematical "threshold law," whereby the neuron sums up its inputs and then gives an output which rapidly increases from minimal to maximal as its total input exceeds a certain threshold level.

By passing signals from one neuron to another in complex circuits, the brain creates and stores information. Unlike nerve cells in the skin or the retina, which transmit information about the external world, neurons in the brain mostly trade information around among themselves. One may thus say with somejustification that the brain is a sense organ which senses itself.

The neuron is in itself a complex dynamical system. However, many theorists have found it useful to take a simplified view of the neuron, to think of it as an odd sort of electrical switch, which takes charge in through certain "input wires" and puts charge out through certain "output wires." These wires are the biologist's synapses. Some of the wires give **positive** charge -- these are "excitatory" connections. Some turn positive charge into **negative** charge -- these are "inhibitory." Each wire has a certain conductance, which regulates the percentage of charge that gets through. But, as indicated above, the trick is that, until enough charge has built up in the

neuron, it doesn't fire at all. When the magic "threshold" value of charge is reached, all of a sudden it shoots its load.

This "electrical switch" view of the neuron is the basis of most computer models of the brain. One takes a bunch of these neuron-like switches and connects them up to each other, thus obtaining a vaguely brain-like system which displays remarkable learning and memory properties. But it is important to remember what is being left out in such models. First of all, in the brain, the passage of charge from one neuron to the other is mediated by chemicals called neurotransmitters. Which neurotransmitters a given neuron sends out or receives can make a big difference. Secondly, in the brain, there are many different types of neurons with different properties, and the arrangement of these types of neurons in particular large-scale patterns is of great importance. Each part of the brain has different concentrations of the different neurotransmitters, and a different characteristic structure. Here we will be concerned in particular with the cortex, which poses a serious problem for the brain theorist, as it has much less of an obvious architecture than such areas as the hindbrain or the cerebellum.

Cell Assemblies

In the late 1940's, in his book *The Organization of Behavior*, Donald Hebb (Hebb, 1949) proposed a neuronal theory of high-level brain function. He hypothesized that learning takes place by the adaptive adjustment of the conductances of the connections between neurons. And he argued that thoughts, ideas and feelings arose in the brain as neural assemblies -- groups of neurons that mutually stimulate each other, and in this way maintain a collective dynamical behavior. While crude and clearly in need of biochemical, neuroanatomical and mathematical elaboration, Hebb's conceptual framework is still the best guide we have to understanding the emergence of mind from brain. It has inspired many current theorists, most notably Edelman (1987) and Palm (1982), and it underlies the ideas to be presented here.

In dynamical systems terms, we may recast Hebb's model by saying that mental entities are activation patterns of neural networks, which may arise in two ways: either as relatively ephemeral patterns of charge passing through networks, or as persistent **attractors** of subnetworks of the brain's neural network. The process of learning is then a process of adaptive modification of neuronal connections, so as to form networks withdesired attractors and transient activation patterns. The transient case corresponds to formal neural network models of perceptual and motor function, principally feedforward network models (Rumelhart et al, 1986). In these models the goal is to modify synaptic conductances so that the network will compute a desired function. The attractor case, on the other hand, corresponds to formal neural network models of associative memory (see Serra and Zanarini, 1991). In these models, the goal is to modify synaptic conductances so as to endow a network with a given array of attractors. The network may then remain constantly in a certain attractor state, or, alternately, it may possess a variety of different attractor states. A given attractor state may be elicited into by placing the network in another state which is in the **basin** of the desired state.

This modernized Hebbian view of neural network learning is the basis of many formal neural network models, and in this sense it is known to be mathematically plausible. Biologically, it cannot be regarded as proven, but the evidence is nevertheless fairly convincing. On the one

hand, researchers are beginning to document the existence of complex periodic and chaotic attractor states in the cortex -- the best example is Freeman's (1992) work on the olfactory cortex. And, on the other hand, the search for biochemical mechanisms of synaptic modification has turned up two main candidates: the number of vesicles on the presynaptic side of the synapse and the thickness of the spine on the postsynaptic side of those synapses involving dendritic spines (Braitenberg and Schuz, 1994 and references therein).

The neural assembly model is quite general. What it does not give us is a clear picture of how simple assemblies build up to form more complex assemblies, and how the dynamics of simple assemblies relate to the dynamics of the more complex assemblies of which they are parts. To get at issues such as this, one needs to look at the architecture of particular regions of the brain, in this case the cortex.

3.3 THE STRUCTURE OF THE CORTEX

The cortex is a thin, membrane-like tissue, about two millimeters thick. It is folded into the brain in a very complex way, and is generally understood to be structured in two orthogonal directions. First it has a laminar structure, a structure of layers upon layers upon layers. The consensus is that there are six fairly distinct layers, although in some areas these six may blend with each other, and in others some of these six may subdivide into distinct sublayers. Then, perpendicular to these six layers, there are large neurons called **pyramidal** neurons, which connect one layer with another. These pyramidal neurons are surrounded by smaller neurons, most notably the **interneurons**, and form the basis for **cortical columns**, which extend across layers.

Pyramidal cells comprise about 85% of the neurons in the cortex, and tend to feed into each other with excitatory connections; there is good reason to consider the network of pyramidal cells as the "skeleton" of cortical organization. Pyramidal cells are distinguished by the possession of two setsof dendrites: basal dendrites close to the main body of the cell, and apical dendrites distant from the main cell body, connected by a narrow shaft-like membrane formation. Pyramidal cells in the cortex transmit signals mainly from the top down. They may receive input from thousands of other neurons -- less than the 10,000 inputs of a Purkinje cell, but far more than the few hundred inputs of the smallest neurons. Pyramidal neurons can transmit signals over centimeters, spanning different layers of the cortex. Lateral connections between pyramidal cells can also occur, with a maximum range of 2-3 millimeters, either directly through the collateral branches of the axons, or indirectly through small intervening interneurons. In many cases, there is a pattern of "on-center, off-surround," in which pyramidal neurons stimulate their near neighbors, but inhibit their medium-distance neighbors.

The columnar structure imposed by pyramidal neurons is particularly vivid in the visual cortex, where it is well-established that all cells lying on a line perpendicular to the cortical layers will respond in a similar way. A column of, say, 100 microns in width might correspond to line segments of a certain orientation in the visual field. In general, it is clear that neurons of the same functional class, in the same cortical layer, and separated by several hundred microns or less, share almost the same potential synaptic inputs. The inputs become more and more similar as the cell bodies get closer together. What this suggests is that the brain uses redundancy to

overcome inaccuracy. Each neuron is unreliable, but the average over 100 or 1000 neurons may yet be reliable. For instance, in the case of motion detection neurons, each individual neuron may display an error of up to 80% or 90% in estimating the direction of motion; yet the population average may be exquisitely accurate.

The visual cortex highlights a deep mystery of cortical function which has attracted a great deal of attention in the last few years, the so-called **binding problem**. The visual cortex contains a collection of two-dimensional maps of the scene in front of the eyes. Locations within these maps indicate the presence of certain features -- e.g. the presence at a certain location of a line at a certain orientation, or a certain color. The question is, how does the brain know which features correspond to the same object? The different features corresponding to, say, a cat in the visual field may be stored all over the cortex, and will generally be all mixed up with features of other objects in the visual field. How does the cortex "know" which features go with the cat? This is related to the problem of consciousness, in that one of the main functions of consciousness is thought to be the binding of disparate features into coherent perceived objects. The current speculation is that binding is a result of temporal synchronization -- that neurons corresponding to features of the same object will tend to fire at the same time (Singer, 1994). But this has not been conclusively proven; it is the subject of intense current research.

A recent study by Braitenberg, Shuz and others at the Max Planck Institute for Biological Cybernetics sheds much light upon the statistics of cortical structure (Braitenberg and Shuz, 1994). They have done a detailed quantitative study of theneurons and synapses in the mouse cortex, with deeply intriguing results. They find that, in the system of pyramidal cell to pyramidal cell connections, the influence of any single neuron on any other one is very weak. Very few pairs of pyramidal cells are connected by more than one synapse. Instead, each pyramidal cell reaches out to nearly as many other pyramidal cells as it has synapses -- a number which they estimate at 4000. Furthermore, the cells to which a given pyramidal cell reaches can be spread over quite a large distance. The conclusion is that no neuron is more than a few synapses away from any other neuron in the cortex. The cortex "mixes up" information in a most remarkable way.

Braitenberg and Shuz give a clever and convincing explanation for the emergence of columnar structure from this sprawling pyramidal network; they show how patches of lumped inhibitory interneurons, spaced throughout the cortex, could cause the pyramidal neurons inbetween them to behave as columnar feature receptors, in spite of having connections extending at least two or three columns out in any direction. This is fascinating, as it shows how the columnar **structure** fits in naturally with excitatory/inhibitory neuron **dynamics**.

Finally, the manner in which the cortex deals with sensory input and motor output must be noted. Unlike the multilayered feedforward neural networks often studied in cognitive science, which take their inputs from the bottom layer and give their outputs from the top layer, the cortex takes both its input and its outputs from the bottom layers. The top layers help to process the input, but if time is short, their input may be overlooked and processing may proceed on the basis of lower-level neural assemblies.

3.4 A THEORY OF CORTICAL DYNAMICS

Given the highly selective account of neural dynamics and cortical structure which I have presented here, the broad outlines of the relation between the psynet model and the cortex become almost obvious. However, there are still many details to be worked out. In particular, the emergence of abstract symbolic activity from underlying neural dynamics is a question to which I will devote special attention.

Before going into details, it may be useful to cite the eight principles of brain function formulated by Michael Posner and his colleagues (Posner and Raichle, 1994), on the basis of their extensive work with PET brain scanning technology. Every one of these principles fits in neatly with the psynet view of brain/mind:

Elementary mental operations are located in discrete

neural areas...

Cognitive tasks are performed by a network of widely

distributed neural systems...

Computations in a network interact by means of "re-

entrant" processes...

Hierarchical control is a property of network operation...

Activating a computation from sensory input (bottom-

up) and from attention (top-down) involves many of the same neurons...

Activation of a computation produces a temporary

reduction in the threshold for its reactivation...

When a computation is repreated its reduced threshold

is accompanied by reduced effort and less attention...

Practice in the performance of any computation will

decrease the neural networks necessary to perform it...

Posner's principles emphasize pattern recognition, hierarchical structure, distributed processing and self-organization (re-entrant processes) -- qualities which the psynet model ties up in a neat and synergetic bundle. What they do not give, however -- what does not come out of brain imaging studies at all, at the current level of technology -- is an explanation of how these processes and structures emerge from underlying neurodynamics. In order to probe this issue,

one must delve deeper, and try to match up particular properties of the cortex with particular aspects of mental function.

There are many different ways to map the psynet model onto the structure of the cortex. The course taken here is to look at the most straightforward and natural correspondence, which can be summarized in four principles:

Proposed Psynet-Cortex Correspondence

1. Neural assemblies may be viewed as "magicians" which

transform each other

3. What assemblies of cortical pyramidal neurons do is to

recognize patterns in their inputs

3. The multiple layers of the cortex correspond to the

hierarchical network of the dual network

4. The pyramidal cells based in each level of the cortex

are organized into attractors that take the form of two-dimensional, heterarchical networks, in which cells represent emergent patterns among neighboring cells

The first principle is essentially a reinterpretation of the cell assembly theory. If one accepts that cell assemblies have persistent attractors, and if one accepts that synapses are modified by patterns of use, then it follows that cell assemblies can, by interacting with each other, modify each other's synapses. Thus cell assemblies transform each other.

The second principle, that neural processes recognize patterns, is also more programmatic than empirical, since virtually any process can, with a stretch of the imagination, be interpreted as recognizing a pattern. The real question is whether it is in any way **useful** to look at neural processes as pattern-recognizers.

The pattern-recognition view is clearly useful in the visual cortex -- feature detectors are naturally understood as patternrecognizers. And I believe that it is also useful in a more general context. Perhaps the best way to make this point is to cite the last three of Posner's principles, given above. These state, in essence, that what the brain does is to **recognize patterns**. The two principles before the last state that components of the brain are more receptive to stimuli similar to those they have received in the recent past -- a fact which fact can be observed in PET scans as reduced blood flow and reduced activation of attention systems in the presence of habituated stimuli. And the final principle, in particular, provides a satisfying connection between neuroscience and algorithmic information theory. For what it says is that, once the brain has recognized something as a repeated pattern, it will use less **energy** to do that thing. Thus, where

the brain is concerned, energy becomes approximately proportional to subjective complexity. Roughly speaking, one may gauge the neural complexity of a behavior by the amount of energy that the brain requires to do it.

Turning to the third principle of the proposed psynet-cortex correspondence (that the multiple layers of the cortex are layers of more and more abstract patterns, ascending upwards), one may once again say that this is the story told by the visual cortex, in which higher levels correspond to more and more abstract features of a scene, composed hierarchically from the simpler features recognized on lower levels. Similar stories emerge for the olfactory and auditory regions of the cortex, and for the motor cortex. Numerous connections have been identified between perceptual and motor regions, on both lower and higher levels in the hierarchy (Churchland et al, 1995), thus bolstering the view that the lower levels of the cortex form a unified "perceptual-motor hierarchy." It would be an exaggeration to say that the layers of cortex have been conclusively **proved** to function as a processing hierarchy. However, there are many pieces of evidence in favor of this view, and, so far as I know, none contradicting it.

The final principle, the correspondence between the heterarchical network and the organization of attractors in single layer of the cortex, is the least obvious of the four. In the case of the visual cortex, one can make the stronger hypothesis that the columnar organization corresponds to the heterarchical network. In this case the organization of the heterarchical network is based on the organization of the visual scene. Feature detectors reside near other feature detectors which are "related" to them in the sense of responding to the same type of feature at a location nearby in the visual field. This organization makes perfect sense as a network of emergence, in that each small region of a scene can be approximately determined by the small regions of the scene immediately surrounding it. The dynamic and inventive nature of this network representation of the visual world is hinted at by the abundance of perceptual illusions, which can often be generated by lateral inhibition effects in neural representations of scenes.

In order for the fourth principle to hold, what is required is that other regions of the cortex contain maps like those of the visual cortex -- but not based on the structure of physical space, based rather on more general notions of relatedness. Thisis not an original idea; it has been explored in detail by Teuvo Kohonen (1988), who has shown that simple, biologically plausible two-dimensional formal neural networks can be used to create "self-organizing feature maps" of various conceptual spaces. All the formal neurons in one of his feature map networks receive common input from the same collection of formal neurons on an hypothetical "lower level"; each formal neuron also exchanges signals with the other formal neurons in the feature map network, within a certain radius.

This is a crude approximation to the behavior of pyramidal cells within a cortical layer, but it is not outrageously unrealistic. What happens is that, after a number of iterations, the feature map network settles into a state where each formal neuron is maximally receptive to a certain type of input. The two-dimensional network then mirrors the topological structure of the high-dimensional state space of the collection of inputs, in the sense that nearby formal neurons correspond to similar types of input.

Kohonen's feature map networks are not, intrinsically, networks of emergence; the notion of "relatedness" which they embody is simply proximity in the high-dimensional space of lower-level inputs. However, these feature maps do provide a vivid illustration of the spontaneous formation of two-dimensional neural maps of conceptual space. What Kohonen's work suggests is a restatement of Principle 4 of the hypothesized psynet/cortex correspondence:

4'. Each cortical layer consists of a network of pyramidal neurons organized into Kohonen-style feature maps, whose topology is based on a **structural** notion of relatedness between nearby pyramidal neurons.

In order for this restated principle to hold true, it is sufficient that two properties should hold. These criteria lie at the borderline of mathematics and biology. It is possible that they could be proved true mathematically, in such a general sense that they would have to hold true in the cortex. On the other hand, it seems more likely that their mathematical validity relies on a number of conditions, the applicability of which to the cortex is a matter of empirical fact.

The first property is that pyramidal neurons and neuronal groups which are close to each other, and thus have similar inputs from lower level neurons, should, on average, recognize similar patterns. Of course, there will be cases in which this does not hold: one can well have two neurons with almost the same inputs but entirely different synaptic conductances, or with different intervening interneurons turning excitatory connections to inhibitory ones. But this is not generally the case in the visual cortex -- there what we see is quite consistent with the idea of a continuity of level k+1 feature detectors corresponding to the continuity of their level k input.

The second property is that higher-level patterns formed from patterns involved in a network of emergence should themselves naturally form into a network of emergence. We have seen that the associative-memory "network of emergence" structureis an attractor for networks of pattern recognition processes; what the fulfillment of this criterion hinges on is the basin of the network of emergence structure being sufficiently large that patterns recognized among level k attractors will gradually organize themselves into a level k+1 network of emergence.

3.5 EVOLUTION AND AUTOPOIESIS IN THE BRAIN

I have delineated the basic structural correspondence between the psynet model and the cortex. In essence, the idea is that the two orthogonal structures of the cortex correspond to the two principal subnetworks of the dual network. The next natural question is: what about **dynamics**? The psynet model comes equipped with its own dynamics; how do these correspond to the dynamics of brain function?

Recall that, in the previous chapter, a distinction was drawn between two types of dynamics in a dual network: evolution and autopoiesis. This is to some extent an artificial distinction, but it nevertheless useful in a neurobiological context. It is essentially the same as the distinction, in biology, between evolution and ecology. On a more basic, philosophical level, it is a distinction between a force of **change** and a force of **preservation**.

Evolution

First, what about neural evolution? On the simplest level, one may say that the reinforcement of useful pathways between neural assemblies is a form of evolution. Edelman (1987) has called this view "neuronal group selection," or "Neural Darwinism." Essentially, in Neural Darwinism, one has survival of the fittest connections. Chaos and randomness in the neural circuits provide mutation, and long-term potentiation provides differential selection based on fitness. As in the dual network model, the progressive modification of synapses affects both associative memory (within a layer) and hierarchical perception/control (across layers).

In this simplest model of neural evolution, there is no reproduction -- and also no crossover. Edelman argues that the lack of reproduction is compensated for by the vast redundancy of the cortex. For, in a sense, one doesn't need to reproduce connections, because almost every connection one might wish for is already there. There may not be many multiple connections between the same pair of pyramidal neurons, but pyramidal neurons tend to have similar connections to their neighbors, so there will be plenty of multiple connections from one cluster of similar pyramidal neurons to another.

Edelman does not even mention the lack of crossover. From his perspective, mutation alone is a perfectly valid evolution strategy. From another point of view, however, one might argue for the necessity of neural crossover. As will be argued in Chapter Six, crossover is demonstrably a more powerful learning technique than mere mutation. Furthermore, if one considers the two methods as learning algorithms, crossover gives the power-law learning curve so familiar from psychology, while mutation givesa straight-line learning curve. Finally, intuition and introspection indicate that human creativity involves some form of combination or crossing-over of ideas.

As I have argued in *EM*, synaptic modification, in the context of an hierarchical processing network, can provide a kind of reproduction by crossover. By appropriate strengthening and weakening of synapses, one can take two trees of neural assemblies and swap subtrees between them. This is very close to the kind of crossover studied by John Koza (1992) in his "genetic programming paradigm." The difference is that, instead of trees of neural assemblies, he has trees of LISP functions. There is therefore a sense in which the Neural Darwinist model of neural evolution can provide for crossover.

It is not clear, however, whether this kind of mutation-based crossover is enough. I have proposed as a speculative hypothesis that the brain, in its creative evolution, routinely carries out a more flexible kind of crossover -- that its neural networks are easily able to move assemblies from place to place. Memory reorganization would be more effective, it would seem, if memories were actually able to **move** from one part of the brain to another, rather than merely having the connections between them modified. And, from the hierarchical point of view, actual **moving** of neural assemblies would provide for a much more flexible crossover operation between trees and other systems of assemblies. This hypothesis finds some support both in neuroscience and in formal neural network theory. On the one hand, evidence is emerging that the brain is, in certain circumstances, able to move whole systems of assemblies from one place to another; even from one hemisphere to another (Blakeslee, 1991). And, on the other hand, one may show that simple Kohonen-style neural networks, under appropriate conditions, can give rise to spontaneously mobile activation bubbles (Goertzel, 1996a). It is not possible to draw any

definite conclusions as yet, but the concept of "sexually" reproducing neural assemblies is looking more and more plausible.

Autopoiesis

Autopoiesis has an obvious correlate in neural networks. If neural assemblies are magicians, then structural conspiracies are assemblies of neural assemblies -- neural meta-assemblies. Hebb, in the original statement of cell assembly theory, foresaw that neural assemblies would themselves group into self-organizing systems. Of course, self-organizing systems of neural assemblies need not be static, but may be in a continual process of mutual growth and change.

Autopoiesis, in the psynet model, is asked to carry out a wide variety of functions. Essentially, anything involving the preservation of structure over time must be accomplished by pattern/process autopoiesis. This leads to a variety of intriguing hypotheses. For instance, consider the vexing question of symbolic versus connectionist processing. How do the messy, analogue statistical learning algorithms of the brain give rise to the precise symbolic manipulations needed for language and logic. According to the psynet model, this must come out of pattern/process autopoiesis. Thus, in the current brain theory,it must come out of autopoietic systems of neural assemblies.

But how can symbol processing come out of autopoiesis? Intriguingly, mathematics provides a ready answer. The technique of **symbolic dynamics**, to be discussed in Chapter Five, deals precisely with the emergence of symbol systems and formal languages out of complex dynamical systems. To study a dynamical system using symbolic dynamics, one partitions the state space of the system into N+1 regions, and assigns each region a distinct code number drawn from {0,...,N}. The system's evolution over any fixed period of time may then be represented as a finite series of code numbers, the code number for time t representing the region of state space occupied by the system state S(t). This series of code numbers is called a "symbolic trajectory"; it may be treated as a corpus of text from an unknown language, and grammatical rules may be inferred from it. In particular, systems with complex chaotic dynamics will tend to give rise to interesting languages. Chaos, which involves dynamical unpredictability, does not rule out the presence of significant dynamic patterns. These patterns reveal themselves visually as the structure of the chaotic system's strange attractor, and they reveal themselves numerically as languages emergent from symbolic dynamics.

Cohen and Eichenbaum (1995) have demonstrated that cortical-hippocampal feedback loops play a fundamental role in helping the neocortex to store and access symbolic, declarative information. The hypothesis to which the psynet model of brain leads us is that the cortical-hippocampal feedback loops in fact serve to encode and decode symbolic memories in the structures of the attractors of cortical neural assemblies. In fact, one may show that these encoding and decoding operations can be carried out by biologically plausible methods. This is an intriguing and falsifiable hypothesis which ensues directly from applying the simplicity of the psynet model to the complexity of the brain.

3.6 CONCLUSION

It is important not to fall into the trap of believing neural network models, in particular, to exhaust the applicability of complex systems science to the study of brain function. The brain is a very complex system, and complex systems ideas can be applied to it on many different levels -- from the microtubular level stressed by Stuart Hameroff in *Ultimate Computing* and Roger Penrose in *Shadows of Mind*, up to the abstract mental-process level emphasized here.

I am not a neurobiologist, and the cortical model presented here is plainly not a neurobiologist's model. It has an abstract structure which doubtless reflects my background as a mathematician. But, on the other hand -- and unlike many neural network models -- it is not a mere exercise in mathematical formalism. It is, rather, a conceptual model, an intuitive framework for understanding.

The value of such models lies in their ability to guide thought. In particular, this model was developed not only to guide my own thinking about the brain, but to guide my ownthinking about the learning behavior of human, animals and artificial intelligence systems. My hope is that it may help others to guide their thoughts as well. For, after all, the project of understanding the brain/mind is just barely getting under way -- we need all the ideas we can muster.

CHAPTER FOUR

PERCEPTION AND MINDSPACE CURVATURE

4.1 INTRODUCTION

The psynet model is a general model of mental structure and dynamics. It does not pretend, however, to **exhaust** the structure and dynamics of mind. Unlike, say, theories of fundamental physics, it does not pretend to be a **complete** theory of the domain with which it deals.

First of all, particular types of minds will have their own peculiar structures and dynamics -- e.g., the human mind inherits a whole host of peculiarities from its specific array of neurotransmitters, its bilaterally symmetric structure, and so forth. And secondly, there may also be other **general laws** of mental dynamics which are not captured in the basic ideas of the psynet model. At the end of this chapter I will propose one such general law -- the law of "form-enhancing distance distortion," or **mindspace curvature**.

The idea of mindspace curvature is a generalization of the idea of visual space curvature, proposed by A.S. Watson in 1978 as a way of explaining the simple geometric illusions. In a recent paper, Mike Kalish and I have shown how this visual space curvature could arise from underlying self-organizing dynamics in mental process space. The picture that emerges is one of subjective visual space undergoing a dynamic, self-organizing autopoietic process -- one which enhances and creates patterns and forms, in the process creating certain logical inconsistencies, that we call "illusions."

And this notion of visual space creating itself, which arises in the analysis of illusions, turns out to be essential to the psynet model itself. For, the psynet model claims that the mind is an autopoietic magician system -- that the mind, as as a whole, produces itself. This is easy to see in the higher reaches of mind: beliefs support each other, and lower-level percepts lead to higher-level concepts, etc. But it is more difficult to see in the context of the lowest-level mental processes corresponding to the perceptual world. These lowest-level processes, it might seem, come from **outside** the mind: they are not produced by the rest of the mental system.

A careful analysis, however, shows that this is not the case. Even the lowest levels of the mind, dealing with raw perceptual information, can be understood as autopoietic systems, producing themselves with the assistance of the immediately higher levels in the perceptual-motor hierarchy. This autopoiesis, which is the cause of perceptual illusions, would seem to have valuable lessons to teach us about the nature of autopoiesis **throughout the mind**.

4.2 PRODUCIBILITY AND PERCEPTION

The psynet model states that mental entities are autopoietic magician systems; it also states that mind itself is an autopoietic magician system. This means that, in some sense, each element of the mind is producible by other elements of the mind. This statement is, in *Chaotic Logic*, called the **producibility hypothesis**.

It is not immediately obvious, however that this "producibility hypothesis" is a tenable one. The level at which it would seem most vulnerable to criticism is the perceptual level. Supposing one takes an hierarchical point of view; then it is hardly surprising that higher-level patterns should be emergent from, i.e. producible out of, lower-level ones. But where are the lowest-level processes to come from? Low-level motor pattern/processes are reasonably seen to be produced by higher-level motor pattern/processes. But what about processes carrying low-level perceptual information? On the face of it, this would seem to violate the producibility hypothesis.

The violation, however, is only apparent. I will show here that the lowest-level perceptual processes in the dual network can, with the aid of higher-level processes, **mutually produce each other**. Perceptual illusions, I will argue in the following section, are a consequence of quirks in this dynamic of lower-level self-construction -- quirks which manifest themselves as locally-varying curvature of the visual field.

The Construction of Space

Just as visual forms emerge out of space, so space emerges out of visual forms. This observation is important both philosophically and practically. It is obvious in the case of three-dimensional space, and less obvious, but no less important, in the case of two dimensions.

The two-dimensional structure of visual space would appear to follow almost directly from the two-dimensional structure of the retina, and the cortical sheets to which the retina maps. There are, admittedly, many other structures to which the 2-D structure of the retina could be mapped besides 2-D lattices. But at any rate, the 2-D lattice is presented to the brain in an immediate and natural ways.

The step to 3-D vision, on the other hand, requires a far greater degree of abstraction. Somehow the brain posits an additional dimension of lattice structure which is not prima facie there in its input. This additional dimension is an emergent pattern in its binocular input. In this sense, the statement that mind constructs space is not controversial but, rather, almost obvious.

To understand the emergence of space from visual forms, let L denote a process which produces an abstract spatial lattice structure. I.e., given a collection I of stimuli, L arranges these inputs in a lattice with some specified dimensions. Let P denote the perceptual processes that act on the lattice L*I, producing an higher-level abstracted version of L*I, which we may call Q*I. The idea is that Q*I has already had some segmentation operations done on it -- it is fit to be presented to some of thelower-level cognitive processes.

In general, we must allow that L and P interact with each other, i.e., that the lattice arrangement of stimuli may depend on the patterns recognized in the lattice. Thus we find $Q*I = P_n * (L_n * I)$, where the higher-order lattice and perceptual processes are defined in terms of an iterative system $L_{n+1} = L_n * P_{n-1}$, $P_{n+1} = P_n * L_{n-1}$, where the action of L_n on P_{n-1} is a process of top-down parameter adjustment rather than bottom-up information transmission.

Now, Q*I is not identical to L*I, nor need it be possible to go backwards from Q*I to L*I. But in most cases Q*I will be at least as valuable to the remainder of the mind as L*I. In fact, it will often be more valuable: we need the general picture more than we need the micro-level details. In addition, Q*I will generally be simpler than L*I (a conceptual sketch of a scene being much more compact than a detailed representation of the scene at the pixel level). Thus, the operation Q will, as a rule, be a pattern in the stimulus collection I.

The Construction of Stimuli

If stimuli can produce space, can space in turn produce stimuli? Not exactly. What can happen, however, is that space and stimuli can cooperate to produce other stimuli. This is the key to understanding the applicability of the producibility hypothesis to low-level perception.

To make this idea concrete, suppose that one subtracts a particular part of the input collection I, obtaining I' = I - J. Then it will often be the case that the "missing piece" J provides useful information for interpreting the remainder of the image, I'. In other words, at some stage in the production of Q*I', it becomes useful to postulate something similar to J, call it J'. This may occur in the construction of the lattice L, or at a higher level, in the recognition of complex forms in the lattice image. Thus we have, in general, a situation where each part of I is a pattern in Q*I', while Q itself is a pattern in I or I'. In other words, we have a collection of patterns with more structural integrity and resilience than a living organism: chop off any one part, and the other parts regenerate.

One example of this kind of filling-in, at a very low level, is the maximum entropy principle, or MAXENT, discussed extensively in *The Structure of Intelligence*. This principle states that, in a lattice L*I' with a missing sublattice L' corresponding to the missing data J, the missing sublattice L' should be filled in with the data that provides the maximum possible entropy. This is a briefly stated and fast-running algorithm for filling-in missing regions, and in many cases it

provides very accurate results. To understand MAXENT more concretely, suppose one has a photograph P, and obtains two discretizations of it, D_{100} which uses a 100x100 pixel lattice, and D_{400} which uses a 400x400 pixel lattice. If one uses MAXENT to extrapolate from D_{100} to the 400x400 level of accuracy, one finds that the ensuing image MAXENT(D_{100}) is visually very similar to D_{400}. Viewed the other way around, what this means is that, if a single pixel of D_{400} is removed, the other three pixelscorresponding to a single pixel of D_{100} can be used to approximate the missing pixel with reasonable accuracy. Of course, using more pixels in the surrounding region could produce an even better approximation.

Perceptual Autopoiesis and Structural Conspiracy

I have argued, in this section, for a kind of autopoiesis in perceptual systems, by means of which lower-level perceptual processes are produced by each other, in combination with higher-level perceptual processes. Before summarizing the argument, however, a few ancillary points must be clarified.

First, it is certainly not true that all possible input stimuli I will be structured so as to lead to this kind of autopoiesis. However, if a collection of stimuli lacking the properties required for autopoiesis is received, it will quickly be transformed into one which does not suffer this lack. This is one of the many ways in which we construct our own world. We grant the world a coherence over and above whatever coherence is intrinsic to it -- and we do this, not only on the abstract conceptual level, but on the level of the processing of perceptual information.

Next, it should be noted that this same process might occur on a number of hierarchical levels. The output of Q might be taken as inputs to another process Q of similar form, and thus arranged in a lattice. This would appear to be the case in the visual cortex, in which we have lattices of features feeding into lattices of more complex features, and so forth.

Finally, it is clear that the reconstruction process by which a region I' is produced by its neighbors is not exact. Thus a collection of perceptual processes, in this set-up, does not reproduce itself exactly. Rather, it reproduces itself approximately. In other words, it lies within a stochastic strange attractor in its state space: it can vary, but it varies according to certain hidden regularities, and it is unlikely to vary a great deal.

The Appendix to this chapter indicates how one may go about elaborating these arguments in more mathematical detail. The main point, however, has already been made. What these arguments show is that the producibility hypothesis is indeed a plausible one. The emergence of higher-level patterns from lower-level ones is plain; and, according to these arguments, the emergence of lower-level processes from higher-level processes and each other is equally unproblematic. The possibility of self-producing mental systems is confirmed. More formally, we are led towards the following conclusion:

PROPOSITION: There exist hierarchical pattern/process systems, involving lattice structure on the lowest perceptual level, which are structural conspiracies (autopoietic, attractive magician systems).

Admittedly, such an abstract proposition is of rather limited interest, in and of itself. Merely to show that such systems exist, in a mathematical sense, does not say anything about whether such systems are plausible in a scientific sense,i.e. as models of biological, psychological or computational systems. In previous publications, however, I have expended much effort to show that "self-producible" systems -- autopoietic magician systems -- do indeed make scientific sense. The formal notions presented here are, obviously, to be understood in this context.

4.3 THE GEOMETRY OF VISUAL ILLUSIONS

According to the psynet model, the perceptual world is not just perceived but produced -- and largely **self**-produced. But what are the **consequences** of this self-production? Does it make any difference whether we perceive the world or produce it? One obvious consequence of the production of the perceptual world is the existence of perceptual illusions.

It is worth pausing for a moment to consider the concept "illusion." An illusion is not just a "false perception" -- for there is no absolute, non-perceptual reality to which one might meaningfully compare one's perceptions. An illusion is, rather, an instance where perceptions obtained in different ways **disagree**. An illusion is a contradiction in the perceptual world.

For instance, looking at two lines, one judges that line A is longer than line B; but measuring the two with a ruler, one judges that they are the same. Two methods of perceiving the world disagree. The world, as subjectively perceived, is logically inconsistent. This is the essence of illusion. Perceptual illusions are an instance of mental self-organization taking precedence over rational, deductive logic.

Theories of Illusion

Early perceptual psychologists understood the philosophical power of illusions. Existing illusions were hotly discussed; and the discovery of new illusions was greeted with fanfare. These simple line figures, it was thought, were "windows on the mind." If one could understand the mechanisms of illusions, one would understand the dynamics of perception, and perhaps of mind in general.

After 150 years of study, however, this early hope has not paid off -- it lost its steam long ago. Today the study of illusions is an obscure corner of perceptual psychology, considered to have little or no general importance. However, despite dozens, perhaps hundreds of theories, the illusions are still not understood.

The various theories may be, loosely speaking, divided into three categories: physiological, descriptive, and process-oriented. Each of these categories includes a wide variety of mutually contradictory theories. For instance, the physiological theories have covered everything from optical diffraction and eye movements to lateral inhibition in neural networks (Von Bekesy, 1967; Coren et al, 1988). And the descriptive theories have been even more diverse, though they mostly been restricted to single illusions or small classes ofillusions. For instance, it was observed in the middle of the last century that acute angles tend to be underestimated, while obtuse angles tend to be overestimated. Helson (1964), Anderson (1981, 1990) and others have

proposed descriptive models based on contrast effects; while theorists from Muller-Lyer (1889) to Fellows (1968) and Pressey (1971) have used the concept of "assimilation" to explain wide classes of illusions.

Finally, theories of the **processes** underlying the geometric illusions vary from the simplistic and qualitative to the sophisticated and quantitative. For instance, Hoffman has used a sophisticated Lie algebra model to derive a theory of illusions, the empirical implications of which are, however, essentially equivalent the much older theory of angle distortion. More intuitively, Eriksson (1970) proposed a field-theoretic model of illusions, based on the concept that the lines in the visual field repel and attract one another by means of abstract **force fields**.

Cladavetscher (1992) has proposed a general theory of illusions based on information integration theory, which attributes illusions to a weighted sum of contrast and assimilation processes (which, however, are themselves only vaguely defined).

The Curvature of Visual Space

Out of the vast range of illusion theories, the most concordant with the psynet model is the one proposed by A.S. Watson in 1978, which states that the geometric illusions result from the curvature of visual space. This theory is primarily descriptive, but it also hints at underlying processes. In the end, of course, a thorough understanding of the illusions will have to span the physiology, description and process levels.

The concept of curved visual space is most often encountered in the theory of binocular perception. In these models, however, the spaces involved usually possess constant, negative Gaussian curvature, which hypothesizes a non-Riemannian space of fixed geometric structure). Along similar lines, Drosler (1978) has explained aspects of the psychophysics of visual extent by hypothesizing a constant positive Gaussian curvature for monocular visual space. Watson's theory is different from all these in that the proposed curvature of visual space is not constant: it varies locally based on the distribution of objects in visual space, according to a particular equation.

In its use of locally varying curvature, Watson's theory is a conceptual relative of Einstein's General Theory of Relativity, in which massive objects are considered to curve the spacetime continuum. However, it may also be reformulated as a force field theory, bringing it closer to Gestaltist notions of perception. In this interpretation, rather than curving visual space, objects are understood to attract other objects toward themselves by a certain complex, distance-dependent function.

What does the curvature of visual space **mean**? Think about it this way. Nearly everyone can recognize a triangle. And, as Helmholtz discovered, nearly everyone, if asked to estimate the individual angles that make up a triangle, will respond with numbers that do not sum to 180 degrees. This seems paradoxical;but the paradox is removed if one drops the implicit assumption that visual space is **Euclidean**. The key insight of Watson's (1978) theory of illusions is that all the simple geometric illusions can be dealt with in a similar manner: by introducing the notion of curved monocular visual space.

According to Watson, each object present in the visual field curves the region of visual space surrounding it. This curving action is carried out in such a way that pairs of small objects near the curving object are perceived as closer together than they actually are, while pairs of small objects at a moderate distance from the curving object are perceived as further apart than they actually are. Pairs of small objects at a sufficiently great distance from the curving object are unaffected. This particular way of deriving curvature from objects is quite different from anything found in relativity or other branches of physics.

Watson's theory was inspired partly by Eriksson's (1970) unsuccessful field-theoretic model of illusions; it is not surprising, therefore, that it may be recast as a force field model. In particular, Watson's equation for metric distortion is mathematically equivalent to the hypothesis of an attractive force field emanating from an object, with a maximum amount of attraction at some finite nonzero distance.

Some Specific Examples.

Using his equation for local visual space distortion, Watson (1978) gives detailed explanations for a variety of geometric illusions, including the Poggendorff, Muller-Lyer, Enclosure, Ponzo, Wundt, Hering, Orbison, Zollner and Delboeuf illusions. Extension to other illusions besides these seems unproblematic. Three examples will serve to illustrate the character of the explanations provided: the Poggendorff, Delbouef and Ebbinghaus illusions (see Figure 2).

The Poggendorff illusion consists of two parallel lines intersected by a diagonal line. The portion of the diagonal line between the two parallel lines is deleted. The remaining segments of the diagonal line then do not appear to be collinear. Watson derives equations for the curvature of visual space induced by the parallel lines. His psychological hypothesis is then that, when trying to continue one of the segments of the diagonal line, a person follows a **geodesic** in the curved space rather than a straight line (which would be a geodesic in flat, Euclidean space).

A geodesic is a path with the property that it is the shortest path between any two points that lie on it. It is also the only kind of curve along which one can do parallel propagation of a unit vector. Examples are lines in Euclidean space, and great circles on spheres. There is considerable evidence that geodesics are psychologically natural; for instance, when intuitively estimating the path an object has taken from one point to another, a person naturally assume that the object has followed a geodesic.

To understand the Poggendorff illusion in terms of curved visual space, suppose a person is trying to continue the left segment of the diagonal line. The geodesic which begins at theintersection of that segment with the left parallel line, will intersect the right parallel line at a point **above** the intersection of the right segment of the diagonal line with the right parallel line. Thus, according to the theory, the person will perceive the right segment of the diagonal line to begin higher than it really does, which is the effect observed experimentally. The parameters of the curvature equations may be set in such a way as to fit the empirical data regarding the amount of displacement.

Next, the Delbouef illusion consists of two concentric circles (though there are many variants, e.g. concentric squares). The size of the outer circle is underestimated, while the size of the inner circle is overestimated. Here the explanation is more straightforward. One need only assert that the two circles exert an attractive force on each other. The particular form of Watson's force field predicts that, as the two diameters become more and more disparate, the attraction should increases, until a certain point is reached, after which it should begin to decrease. This phenomenon has been observed empirically; it is called the "distance paradox" (Ikeda and Obonai, 1955). A variant on the Delbouef illusion, not considered by Watson, is the Ebbinghaus illusion (Figure 3). This illusion gives particular insight into the nature of Watson's force field. The figure consists of a circle surrounded by four symmetrically positioned context circles. Large context circles make the center circle look smaller than do small context circles.

This illusion was at one point (Massaro and Anderson, 1971) interpreted as a simple contrast effect; however, subsequent experiments reported by Cladavetscher (1990) showed this explanation to be inadequate. Cladavetscher demonstrated a U- shaped curve relating the estimated size of the center circle to the distance between the center circle and the context circles. In order to explain this, he invokes an "information integration" theory, according to which context effects are dominant for large distances, but assimilation effects mute their effect for small distances. This U-shaped curve is taken as evidence for a two- process theory of illusions; Cladavetscher (1990) writes that "no single process seems to provide a satisfactory account of the data." In fact, however, Watson's theory, published 12 years previously, accounts for the U-shaped curve in a very simple fashion. When the center circle is close to the context circles, the center circle is not yet within the range of maximum attraction of the context circles. When the context circles are moved out a little further, the far side of the center circle is within the range of maximum attraction of each context circle, and so the sides of the context circle are pulled closer together. When the context circles are moved even further out, then, the range of maximum attraction of the context circles does not contain the center circle at all, and the perceived size gradually increases again.

Watson's theory accounts for a wide variety of illusions in a simple, unified way, using only a small number of numerical parameters. A word should be said, however, about the possible interactions between the process identified by Watson and other aspects of vision processing. There are many other processeswhich alter the structure of visual space, and one may very plausibly suppose that some of these act simultaneously with Watson's curvature process. It is possible that the interactions between the curvature process and other processes might, in some cases, confuse the practical application of the theory. In the Ebbinghaus illusion, for instance, Watson's theory leads to the hypothesis that the center circle should actually be perceived as a sort of "diamandoid" shape, due to the non-isotropic nature of the curvature induced by the four context circles. One possibility is that this is actually the case; another possibility is that a higher-level propensity to perceive shapes as circles intervenes here. Similarly, in the Poggendorff illusion, Watson's theory predicts that the diagonals should not actually be perceived as straight line segments, but should be pulled toward the parallel lines by an amount which varies with distance. Again, it is possible that this is the case, and it is also possible that a higher-level "propensity for straightness" intervenes, thus restoring a semblance of Euclidean structure. These questions are interesting and important ones; however, we will not pursue them further here.

Illusion and Self-Organization

Watson's theory of illusions is descriptive: it says that visual space is curved, but doesn't say why or how. In accordance with the ideas of the previous section, we would like to say that visual space curvature is a consequence of low-level **autopoiesis**; that it is a quirk in the way the mental processes responsible for visual space construct themselves. In order to do this, however, it is necessary to connect the abstract notion of visual space curvature with some kind of lower-level self-organizing dynamic.

This has been done in a recent paper by Mike Kalish and myself (Goertzel and Kalish, 1996). We have demonstrated the emergence of the force field equivalent of Watson's equation for visual space curvature from an underlying process of spreading activation dynamics and nonlinear filtering. The spreading activation process may be understood in the context of a continuum model of visual space; or it may, with equal facility, be interpreted discretely, in terms of neural network or magician-type models.

One considers a two-dimensional lattice of processing elements, each corresponding a small region of visual space. The activation of each element corresponds to the brightness of the corresponding region of space. The network then undergoes a two-phase dynamic: first a phase of spreading activation, in which bright elements pass their brightness to other elements in their neighborhood, and then a postprocessing phase, in which the extra activation introduced by the spreading is removed. These phases are repeated, in order, until the desired amount of motion has been obtained. The result of this two-phase dynamic is a self-organization of visual space; i.e., the active elements move around in a way which is determined by their distribution.

The dynamics of the motion of the active elements can be modified by adjusting the spreading activation function. Inparticular, as we show, one can construct a spreading activation function that causes active elements to move as if they were obeying a force field equivalent to Watson's equation for visual space curvature. In this way, the geometric illusions are obtained as a consequence of the **self-organization of visual space**.

The Utility of Illusions

Finally, an intriguing open question is the **utility** of the processes leading to illusions. Are the illusions just the result of a shoddily designed visual system, or are they a difficult-to-avoid consequence of some process that is useful for other reasons?

The spreading activation model of (Goertzel and Kalish, 1996) does not resolve this long-standing question, but it does lead to an highly suggestive hypothesis. For spreading activation and subsequent nonlinear filtering is a dynamic of great potential use in early vision processing. It filters out missing segments in lines, and in general "fills out" shapes. Thus it seems quite plausible to conjecture that the illusions are caused by a spreading activation process which evolved for another purpose entirely: not for distorting things metrically, but rather for constructing comprehensible shapes.

The mathematical construction of (Goertzel and Kalish, 1996) can be naturally interpreted as a biological model, following the example of Kohonen (1988), who takes two-dimensional lattices of formal neurons to correspond roughly to portions of cortical sheet. Or it can be understood purely abstractly, in the spirit of the previous section: as a particular dynamic amongst lower-level pattern/process magicians. The two-phase dynamic we have identified is a plausible mechanism for the self-organization of that part of the dual network dealing with visual space. It carries out a consolidation and solidification of forms -- and as a consequence, it distorts forms as well. The illusions arise because, in the self-organization of the dual network, autopoeisis and pattern come first, and deductive, rational order comes after.

Finally, what about perceptual illusions as windows on the mind? Do the processes identified here as being essential to visual illusions have any more general importance? Or are illusions just another example of autopoiesis and self-organization? This question will be addressed in the following section. As I will argue there, it seems quite plausible that the process of **local curvature distortion** might have implications beyond the study of visual illusions. It may well be a part of the formation of mental structure in general. If this is true, then the conclusion is that the early visual psychologists were right: illusions really are "windows on the mind." They really do embody deep processes of more general importance.

4.4 MINDSPACE CURVATURE

I believe that the phenomenon of "inner space curvature" isnot unique to visual perception, or to perception at all, but rather represents a general principle influencing the structure and dynamics of the mind. In particular, the concept of mindspace curvature fits in very nicely with the psynet model. It is not a core concept of the psynet model, but it does have the potential to help explain an important question within the model itself: the **emergence of the dual network**.

Form-Enhancing Distance Distortion

A less poetic but more precise name for the kind of "mindspace curvature" I am talking about here is **FEDD**, or "form-enhancing distance distortion." This is what is seen in visual illusions, and what I am proposing to apply more generally throughout the mind.

FEDD applies wherever one has a mental process P that is responsible for judging the degrees by which other mental processes differ. This process P may be distributed or localized. P's judgement of the difference between x and y will be denoted $d_P(x,y)$. Note that this function d_P need not necessarily satisfy all the mathematical definition of a "metric" -- neither symmetry, nor the triangle inequality.

FEDD states that, over time, process P's judgement of distance will change in a specific way -- in the same way observed in visual illusions, as discussed in the previous section. For each fixed x, the distances $d_P(x,y)$ will all be **decreased**, but the amount of decrease will be greatest at a certain finite distance $d_P(x,y) = F$.

In other words, according to FEDD, each element x pulls the other elements y closer to it, but the maximum amount of attraction is experienced at distance F. This process of adjustment does

not continue forever, but will dwindle after a finite period of time. When new entities are introduced to P, however, the process will be applied to the new entities.

What effect does the pull of x have on the distances between two entities y and z? It is not hard to see that, if y and z are sufficiently close to x, then their mutual distance is decreased; while, if y and z are at a moderate distance from x, their mutual distance is increased. The fabric of the space judged by P is thus distorted in a somewhat unusual way, similar to the way mass distorts space in General Relativity Theory.

What is the purpose of mental-space-distorting force, this FEDD? As the name suggests, it gives additional **form** to the processes judged by P. In particular, it enhances clusters. If a large number of processes are all close to each other, the distance adjustment process will bring them even closer. On the other hand, processes surrounding the cluster will tend to be pushed further away, thus making an "empty space" around the cluster. The net effect is to create much more clearly differentiated clusters than were there before. This may be verified mathematically and/or by computer simulations.

FEDD will not create significant structure where there was no structure before. However, it will magnify small structures into larger ones, and will "clean up" fuzzy structures. Furthermore, it acts in a form-preserving manner. For instance,if there is, instead of a circular cluster, a cluster in the shape of a line, then FEDD will enhance this cluster while retaining its line shape.

FEDD causes patterns to emerge from collections of processes. Thus it is both a pattern recognition process and a pattern formation process. To obtain more concrete characterizations of what FEDD does, it is necessary to look at the different areas of psychological function separately.

Applications of Mindspace Curvature

I have already mentioned the relevance of FEDD to visual perception. A FEDD-like process acting on monocular visual space can explain all the simple geometric illusions (Muller-Lyer, Poggendorff, Enclosure, etc. etc. etc.). My own simulations indicate that FEDD can also do some simple pre-processing operations in vision processing. E.g., it fills in gaps in lines and shapes, and generally makes forms clearer and more distinguishable.

Another likely application is to social psychology. Lewin's theory of force fields in social psychology is well-known. Dirk Helbing, in his book "Quantitative Sociodynamics" (1995) has given a mathematical formulation of Lewin's theory in terms of diffusion equations. Lewin and Helbing's force fields do not necessarily satisfy the form of FEDD. However, it would be quite possible to seek evidence that social fields do, in reality, satisfy this form.

For instance, I would hypothesize that individuals tend to:

1) overestimate the similarity of individuals in

their own social group to one another

2) underestimate the similarity between individuals

in their group and individuals "just outside"

their group

If this is true, it provides evidence that individuals' measures of interpersonal distance obey the FEDD principle. Intuitively speaking, FEDD would seem to provide a very neat explanation for the formation of social groups, meta-groups, and so on.

Finally, there would seem to fairly strong evidence for something like FEDD in the areas of concept formation and categorization. A "concept" is just a cluster in mental-form space; a collection of mental forms that are all similar to each other. Things which we have grouped together as a concept, we tend to feel are more similar than they "really" are. Differences between entities in the concept grouping, and entities outside, are overestimated. This is FEDD. As in the case of perception, here FEDD serves to create structure, and also to cause errors. Mark Randell (personal communication) has suggested that cognitive errors created in this way should be called "cognitive illusions," by analogy to perceptual illusions.

Another possible relationship is with our tendency to make systematic errors in reasoning by induction. Human beings tend to jump to conclusions: as a rule, we are overly confident that trends will continue. This may be seen as a consequence of bunching previously seen situations overly close together, andputting new situations in the bunch.

FEDD and the Emergence of the Dual Network

Now let us see what FEDD has to tell us about the psynet model. The dual network, I have said, is a network of processes which is simultaneously structured according to associative memory (similar things stored "near" each other) and hierarchical control (elements obtaining information from, and passing instructions to, lower-level elements). The only way to combine hierarchy and associativity is to have a "fractal" or recursively modular structure of clusters within clusters within clusters.... Each cluster, on each level, is a collection of related entities, which is govered implicitly by its own global attractor. It is this global attractor which carries out the hierarchical control of the smaller-scale attractors of the entities making up the cluster.

FEDD is a method for inducing and enhancing clusters; it explains how the dual network might emerge in the first place. Each pattern/process P in the hypothesized mind magician system induces its own distance measure: namely, $d_P(x,y)$ is the amount of structure that P recognizes in x but not y, or in y but not x. Here structure can be defined in a number of ways, e.g. by algorithmic information theory. The emergence of the dual network in the mind magician system is explained by the assumption that these distance measures are updated in the manner prescribed by FEDD.

This idea has interesting implications for human memory. Not only does FEDD explain the nature of concept formation within memory systems but, via the theory of the dual network, it explains the formation of memory systems themselves. What psychologists call separate memory systems are just different clusters in the overall dual network. The clusters all have smaller clusters within them, which is why, having begun to divide memory into separate systems, memory researchers keep coming up with more and more and more different subsystems, subsubsystems, etc. It may be that some of this clustering is there from birth (e.g. episodic versus semantic memory, perhaps). But FEDD could enhance this innate clustering, and help along the more refined clustering that comes along with experience.

Conclusion

In sum, I contend that the early psychologists, who believed visual illusions to contain deep secrets of mental process, were not far wrong. The illusions give vivid pictorial examples of how we distort things, how we move things around in our minds to make more attractive forms -- to **create patterns**, in short. We create patterns because we need patterns; because pattern is the stuff of mind.

Mindspace curvature, distance distortion, is a powerful way of creating new patterns and solidifying old ones. It is a very simple dynamic, and a very general one: it applies wherever there is judgement of distances, i.e., wherever there is evaluation of similarity, or relative complexity. Like gravitational spacetime curvature in general relativity, it is only one force among many. But it is an important force, deserving of much further study.

APPENDIX: PERCEPTUAL PRODUCIBILITY

In section 2 we have argued that lower-level perceptual processes produce each other, with the collaboration of higher-level processes. Using the notation of that section, I will now pursue this same line of argument in a more specific context.

Suppose, for concreteness, that I is drawn from the set S_n of collections of n^2 real numbers (representing, say, brightnesses perceived at different points of the retina). Let M_n denote the set of nxn real matrices; this will be, also for concreteness, our set of possible lattices, the range of the function L. The lattice process L will be taken as a fixed element of the class of computable, bijective functions that map S_n onto M_n.

Let $M_{n;i,j}$ denote the equivalence class containing all matrices M_n which are equal to each other except possibly in the (i,j) entry; let M' denote the set of all such equivalence classes. Let F denote a "filling-in" algorithm, i.e. a computable function that maps M' into M_n. Where J is an element of I, let a(J) denote the corresponding entry (i,j) of L*I. As with M, let I' = I'(J) = I - J. Finally, let P map M_n into a space F of "higher-order features," the nature of which is not important.

Now, let G be the subset of S_n with the property that, whenever I is in G,

$$[1+d(I'',I)]\,[s(Q*I'') + s(K) + C] < s(Q*I'),$$

where K is an approximation of J, I" is the member of M_n obtained by inserting K where J should have been, and C is a constant representing the complexity of this insertion. In other words, this class G of input vectors contains only input vectors whose parts are all integral to the whole, in that it is simpler to assume the part has something close to its correct value, than just to ignore the part altogether.

Clearly, not all inputs I will fall into category G. The nature of category G depends on the process P, the simplicity measure s, the lattice process L, and the metric d. What is required, however, is merely that G is nonempty. In practice one wishes G to be widely distributed throughout the space S_n. The input I may not initially be a member of G, but by a process of successive iteration, the system will eventually **produce** an input vector I which does belong to G.

CHAPTER FIVE

DYNAMICS AND PATTERN

5.1 INTRODUCTION

In this chapter I will return to the question of structure versus dynamics, as discussed at the end of Chapter One. In a psychological context, the psynet model goes a certain distance toward resolving this question: it shows how mental structures might arise as a consequence of mental dynamics. But here I will confront the question in a different way, by describing some formal tools for thinking about and measuring the algorithmic patterns emergent from system dynamics.

I will begin by introducing symbolic dynamics as a means for inferring formal languages describing attractor structure. The Chaos Language Algorithm, a new computational technique for inferring algorithmic patterns from system dynamics, will be described; and some general hypotheses regarding the appearance of emergent patterns in complex systems will be presented.

The relevance of these ideas for the psynet model will be discussed. It will be argued that symbolic dynamics provides a natural mechanism by which a psynet might store symbolic information in its attractor structure.

Finally, in the last section, I will turn to physics, and explore the connections between dynamics, algorithmic pattern, and physical **entropy**.

5.2 SYMBOLIC DYNAMICS

Attractors are patterns in dynamics -- **geometric** patterns. It is interesting and important to find ways to represent these geometric patterns **symbolically**. This leads us to what is perhaps the most interesting tool in the dynamical systems theory toolkit: **symbolic dynamics**. Symbolic dynamics relates the continuous mathematics of iterations on real and complex spaces to the discrete mathematics of computation and formal linguistics.

Only in the past few decades has symbolic dynamics emerged as a useful mathematical tool. Its conceptual roots, however, go back much further, at least to Leibniz [Leibniz, 1679; see Loemker, 1969]. Leibniz proposed a "universal character for 'objects of imagination'"; he argued that

a kind of alphabet of human thoughts can be worked out and that everything can be discovered and judged by a comparison of the letters of this alphabet and an analysis of the words made from them

And this systematization of knowledge, he claimed, would lead to an appreciation of subtle underlying regularities in the mind andworld:

[T]here is some kind of relation or order in the characters which is also in things... there is in them a kind of complex mutual relation or order which fits things; whether we apply one character or another, the products will be the same.

In modern language, the universal characteristic was intended to provide for the mathematical description of complex systems like minds, thoughts and bodies, and also to lead to the recognition of *robust emergent properties* in these systems, properties common to wide classes of complex systems. These emergent properties were to appear as *linguistic regularities*.

Leibniz didn't get very far with this idea. He developed the language of formal logic (what is now called "Boolean logic"), but, like the logic-oriented cognitive scientists of the 1960's-1980's, he was unable to build up from the simple formulas of propositional logic to the complex, self-organizing systems that make up the everyday world. Today, with dynamical systems theory and other aspects of complex systems science, we have been able to approach much closer to realizing his ambitions. (One might argue, however, that in the intervening centuries, Leibniz's work contributed to the partial fulfillment of his programme. The formal logic which he developed eventually blossomed into modern logic and computer science. And it is computer power, above all, which has enabled us to understand what little we do about the structure and dynamics of complex systems.)

Abraham and Combs [1995] point out the remarkable similarity between Leibniz's Universal Characteristic and the modern idea of a unified theory of complex system behavior; they single out the concepts of attractors, bifurcations, Lyapunov exponents and fractal dimensions as being important elements of the emerging "universal characteristic." It is *symbolic dynamics*, however, which provides by far the most explicit and striking parallel between Leibniz's ideas and modern complex systems science. Symbolic dynamics does precisely what Leibniz prognosticated: it constructs formal alphabets, leading to formal "words" and "sentences" which reveal the hidden

regularities of dynamical systems. As yet it does not quite live up to Leibniz's lofty aspirations -- but there is reason to be optimistic.

Symbolic Dynamics

The basic idea of symbolic dynamics is to divide the state space of a system into subsets called **cells**, numbered 1 through N. Usually the cells are assumed disjoint, and are assumed to completely fill the state space of the system. Every possible state of the system is thus assigned some code number. One then charts each trajectory of the dynamical system as an itinerary of cells: "this trajectory goes from cell 14 to cell 3 to cell 21 to cell 9, and so forth...." The collection of **code sequences** obtained in this way can tell one a great deal about the nature of the dynamics. In essence, one is translating the dynamics of the system into a collection of abstract **words**!

Symbolic dynamics is particularly useful where the system involved is **chaotic**. Chaos, which involves dynamical unpredictability, does not rule out the presence of significant purely dynamic patterns. These patterns reveal themselves as the structure of the chaotic system's strange attractor. Examples will be given in the following section.

For instance, consider the Baker map

$$x_{n+1} = 2x_n \bmod 1 \quad (1)$$

Using the two "cells" [0,1/2] and (1/2,1], one finds that the dynamics of the Baker map lead to the following map on code sequences

$$s(a_1a_2a_3...) = a_2a_3a_4... \quad (2)$$

Specifically, the code sequence corresponding to a certain initial value x_0 is precisely the binary expansion of the number x_0. Defining a topology on code sequence space by

$$d(a_1a_2a_3..., b_1b_2b_3...) = 2^{-k}|a_k - b_k| \quad (3)$$

it is easy to see that s is chaotic on code sequence space: it is sensitively dependent, topologically transitive, and has dense repelling periodic points.

Probabilistic Symbolic Dynamics

Unfortunately, at present most of the mathematical theory of symbolic dynamics is only useful for analyzing "toy iterations." But for computational purposes, the probabilistic version of symbolic dynamics, often called **Markov analysis**, is

of extremely general utility. While a little less elegant than its deterministic counterpart, it is far more versatile. In Markov analysis, instead of charting which cell-to-cell transitions are **possible**, one assigns a **probability** to each cell-to-cell transition. Thus the dynamics on the attractor are

represented by a matrix of probabilities p_{mn}, representing the chance of a point in cell m moving to cell n at the next iteration step.

There is a beautiful result called **Pesin's Theorem** which connects Markov analysis with Liapunov exponents. It says that, under appropriate conditions (e.g. if f is a diffeomorphism), the sum of the Liapunov exponents is equal to a quantity called the **metric entropy**. If the Liapunov exponents are not constant, this sum must be integrated over the attractor; see Pesin, 1977). But the the intriguing thing is that metric entropy is **also** intimately connected with thermodynamics. The metric entropy involves the **entropy** of the collection of first-order Markov probabilities -- i.e.

$$- p_{mn} \log p_{mn} \quad (4)$$

One divides this entropy by , and then takes the **supremum** (i.e., the least upper bound) of this quotient over all possiblepartitions into cells. The metric entropy thus defined is a qualitative, information-theoretic measure of the degree of "information-scrambling" implicit in a dynamical system. It is very satisfying indeed that the metric entropy should turn out to connect so simply with the Liapunov exponents, which are defined in terms of calculus rather than information.

Dynamical Systems that "Know" English

As a rule, it is is very difficult to construct the symbolic dynamics underlying a given iteration, even a "toy" iteration. But there are exceptions. One of these is the class of "piecewise linear" interval maps -- maps which, for some division of the interval into finitely many subintervals, are linear on each subinterval. The Baker map is piecewise linear, but it is nonrepresentatively simple. Piecewise linear maps are extremely flexible in structure. For instance, it is amusing to observe that one can construct a piecewise linear map which "knows" English.

What I mean by "knowing" English is this. Suppose that one partitions the interval [0,1] into a collection of subintervals, and assigns a subinterval to each English word. Then, if one produces a trajectory of the map from an arbitrary starting point, the symbolic dynamics of this trajectory will be a series of grammatical English sentences.

The key to this construction is the notion of **probabilistic grammar**. A probabilistic grammar is a finite set of rewrite rules, involving elements of a finite list of words, each one tagged with a certain probability. Here we will be directly concerned only with formal grammars, but of course, these formal grammars would be of little interest if not for their connection with the grammars of natural languages. Numerous linguists, most notably Zellig Harris (1988), have argued that the syntax of natural language can be expressed in terms of probabilistic grammars. Harris gives a fairly complete treatment of English syntax using tables indicating the probability of a given "operator" carrying a given "argument," combined with a handful of simple transformation rules.

It is easy to construct a dynamical system whose state space represents a given grammar. First, one simply takes each of the N words involved, and assigns to them, in alphabetical order, the N subintervals (i/N,(i+1)/N), where i = 0,...,N-1. Let I_w denote the subinterval assigned to word w.

Divide each subinterval I_w into M equally sized subintervals (subsubintervals) $I_{w,r}$, r = 1,...,M. Here the number M should be very large; for instance, M might represent the size of the corpus from which the grammar was constructed, and should be divisible by N. The function f is defined to be linear over each subinterval $I_{w,r}$. Each word x gets a number of subintervals $I_{w,r}$ proportional to the probability of the transition from w to x. Over each subinterval $I_{w,r}$ which is assigned to x, the function f is defined to be the linear function mapping the left endpoint of $I_{w,r}$ into the left endpoint of I_x, and the right endpoint of $I_{w,r}$ into the right endpoint of I_x.

Now, given this construction, suppose one constructs symbolic dynamics on the partition defined by the I_w, for all thewords w. One finds that the probabilistic grammar obtained in this way is precisely the probabilistic grammar from which the dynamical system was constructed. For models incorporating k'th order transitions, where k exceeds one, the standard trick suffices: instead of looking at intervals corresponding to words, look at intervals corresponding to k-tuples of words. If the probabilistic grammar in question is, say, a grammar of English, then the dynamical system knows English! According to Harris's analysis of English, all English sentences can be obtained by taking the results of this dynamical system and transforming them by a handful of simple transformation rules.

One may well doubt whether there is any purpose in constructing dynamical systems which know English! The answer, of course, is that while there is no inherent value in such a highly contrived system as the one given here, the **possibility** of encapsulating language in the attracting behavior of a dynamical system is rife with potential applications. If, instead of a piecewise linear map, one had a dynamical system of some relevance to neural or psychological function, then one would have quite an interesting result! It is not clear exactly how complexity of grammar generation corresponds with dynamical complexity; but this would seem to be a promising area for research.

5.3 GENERALIZED BAKER MAPS

To understand the complexity that can arise from even the simplest iterations, general considerations will not suffice; one must look at specific examples. An example which I have found interesting is the generalized Baker map

$$x_{n+1} = Phi(x_n) = (beta\ x_n + alpha) \bmod 1 \qquad (5)$$

$$1 < beta < 2$$

A detailed study of this map is quite rewarding. A good special case to think about is *beta* = 3/2, *alpha* = 0. The iteration function for this case is as shown in Figure 3.

If one does symbolic dynamics on the generalized Baker map, using the natural cells [0,1/*beta*] and (1/*beta*,1], one finds that each point gives rise to a code sequence which is its **expansion in base beta** (see e.g. Blanchard, 1989). Expansions in noninteger bases being somewhat subtler than binary expansions, one runs into certain complications: the "code

sequence space" in question no longer contains all binary sequences, but only those special binary sequences that correspond to expansions in base beta. But the argument essentially works the same; the iteration can be proved chaotic by reference to the chaoticity of a shift map on this strange code space (Goertzel, Bowman and Baker, 1993).

Taking the case beta = 3/2, let us consider some sample expansions. We have, for instance,

.100... = 2/3

.0100... = 4/9

.00100... = 8/27

.10100... = 2/3 + 8/27 = 26/27

All these are unproblematic. Things get trickier, however, when one looks at something like

.1100... = 2/3 + 4/9 = 10/9.

How can a "decimal," a number with zero on the left hand side of the "decimal point," be greater than one? Clearly something is wrong here. What is wrong is, specifically, that one is not dealing with an integer base!

Since a double one is illegal in the leading position, it is also illegal everywhere else. After all, it is equally unsettling to look at

.01100... = 10/27 > 2/3 = .100...

And double ones are not the only forbidden sequences. Consider, for instance,

.1010100... = 2/3 + 8/27 + 32/243 = 266/243

The pattern 10101 is therefore disallowed in any base 3/2 expansion. On the other hand, a simple computation shows that

.100100100...

is just fine -- skipping two zeros does the trick!

With a little creativity one can construct a huge variety of disallowed sequences. This gives rise to the question: precisely how many sequences **are** allowed? As the sequence length n tends to infinity, what percentage of the 2*n* possible sequences are permitted? There is only one answer which makes any intuitive sense: approximately (3/2)*n* sequences should be allowed. In the general case, approximately *betan* sequences should be allowed. Leo Flatto (personal communication) has proved this result using number-theoretic techniques; however, his proof works only for the case alpha = 0. Working independently, Harold Bowman and I (Goertzel et al,

1993) proved the same theorem for the case of general alpha, in a way which relates naturally with the **dynamics** of the map. Both of the proofs, however, involve fairly advanced mathematics, and therefore neither will be given here.

The map *Phi* does not have an "attractor" in the strict sense; what is has is an "invariant measure." This means that, no matter what value of x_0 one begins with, the values of x_i will wander all over the interval [0,1], but there will be a **pattern** to their distribution. If one charts the probabilities with which the x_i fall into different subintervals, one will obtain an approximation to the invariant measure of the map. The measure is invariant because these probabilities are the same for almost any initial value.

However, the distinction between attractors and invariant measures is really not philosophically important. If one constructs the "Frobenius-Perron" operator corresponding to *Phi*,an operator which tells how *Phi* maps probability distributions onto probability distributions, then the invariant measure re-emerges as an **attractor** of this operator.

The motion of the iterates of *Phi* through the invariant measure are chaotic. It is not difficult to see why. Conjugacy to the shift map follows in the same way that it does for the ordinary Baker map; the difference is that not all sequences are allowable. To show topological transitivity, one must therefore be a little cleverer. Let A denote the set of all finite allowable sequences, and suppose that the elements of A are ordered $\{a_1,a_2,...\}$. Then the sequence

$$.a_100a_200a_300a_4... \quad (6)$$

is allowable, and is topologically transitive, because running the shift dynamics on it will bring one arbitrarily close to any infinite allowable sequence. It is worth pausing a moment to realize why this sequence must be allowable. The reason is that, as observed above, .1001001001... is allowable, and because all of the a_i are allowable, none of the expressions $.a_i$ can exceed 1 in value.

Finally it is interesting to think about the **periodic** points of the generalized Baker map. How many periodic points of period n are there? Quite elegantly, it turns out that the answer is: on the order of *betan*. The argument is as follows. Suppose one takes an allowable sequence of length n-2 and sticks two zeros on the end of it; then this sequence can be repeated over and over again forever, to give a periodic allowable sequence of period n. On the other hand, not every periodic sequence of period n is an **allowable** periodic sequence of period n. Thus the number of allowable periodic sequences of period n is greater than the number of allowable sequences of length n-2, and less than the number of allowable sequences of length n. But the number of allowable sequences of length n-2 is about *betan-2*, and the number of allowable sequences of length n is about *betan*. So the number of periodic points of period n is on the order of *betan*.

All this is fairly elementary. The remarkable thing is the elegance and subtlety that arises from considering this very simple iteration. One may well wonder, if this much structure underlies an iteration constructed from drawing two lines above the unit interval, we can possibly hope to understand what is going on with **real-world** dynamical systems? The answer is that we will have to ask different questions. Counting the periodic points of a given order will be out of the

question; even proving chaoticity will usually be too difficult. We will have to learn how to ask questions reflecting on **qualitative** behavior; otherwise it is unlikely to be possible to go beyond simple "toy

iterations" like the Baker map and its generalizations.

5.4 THE CHAOS LANGUAGE ALGORITHM

We have discussed symbolic dynamics, and then placed the concept of system pattern in a more general context. In thissection I will describe a computational algorithm for extracting patterns from systems, which is an extension of symbolic dynamics. The algorithm is still in a developmental stage, but it has already yielded some interesting results. A little later we will describe an application of the algorithm to the exploratory analysis of mood cycle data.

The algorithm, called the Chaos Language Algorithm (CLA), was developed collaboratively by the author and Gwen Goertzel. In its current implementation, the CLA is being used for the recognition of purely dynamic patterns, and it recognizes only a narrow class of patterns: patterns based on repetition. But this type of pattern is very common in practice, thus even this simplistic implementation is of definite use in the analysis of real systems. Above all, the CLA serves to put some "meat on the bones" of the abstract concept of algorithmic information in complex systems, by giving an explicit account of an important class of algorithmic patterns -- context-free grammars.

The concept at the heart of the CLA is the use of formal languages to represent patterns in the trajectories of a dynamical systems. In its simplest version, the CLA is a three-stage process, consisting of

1. discretization of trajectories of a dynamical system by symbolic dynamics. One divides the potential state space of the system into a finite number of disjoint regions, and assigns each region a code symbols, thus mapping trajectories of the system into series of code symbols. This series of code symbols may be interpreted as a lengthy text in an unknown formal language; the problem of the CLA then being to perform an Harris-style linguistic analysis of this text.

2. tagging of symbolic trajectory sequences using a self-organizing tagging algorithm. One takes the symbols derived in the first step and assigns each one to a certain "category" based on the idea that symbols which play similar roles in the trajectory should be in the same category. The "tag" of a code symbol is a number indicating the category to which it belongs; The code sequence is thus transformed into a "tag sequence." In linguistic terms, one is assigning words to grammatical categories based purely on frequency information.

3. inference of a grammar from the tag sequence produced in the previous step, and iterative improvement of the tagging in order to maximize grammar quality.

The end result of this algorithmic process is a formal language which captures some of the statistical and algorithmic structure of the attractor of the dynamical system. Each of the three

stages may be carried out in a variety of different ways; thus the CLA is as much a "meta-algorithm" as it is an algorithm in itself.

Step 1, in the current implementation, is carried out using the standard method of symbolic dynamics. This is straightforward for the case where the system's state space is a subset of R*n*, as occurs, for example, when one's data regarding the system takes the form of a table of numbers.

Step 2 is carried out by computing the "successor structure" corresponding to each code symbol. This structure is a compact way of representing the various Markov probabilities associated with a given code symbol. The idea is that the relationship of a symbol a with other symbols may be quantified in terms of the collection of subsequences containing a. For each symbol a, one may construct a vector $Succ_1(a)$ of the form

$$Succ_1(a) = (count(a0),count(a1),...,count(a(N-1)) \quad (14)$$

where count(ab) gives the number of times that the subsequence ab occurs in the symbol sequence. And, in a similar fashion, one may build a vector $Succ_2(a)$ of all two-symbol sequences that occur immediately after a at any point in the symbol sequence, and so forth. One thus derives, for each symbol a, a "successor structure"

$$Succ(a) = (Succ_1(a),...,Succ_k(a)) \quad (7)$$

where k is the **order** of the structure. By omitting all sequences of zero count and representing Succ(a) as a tree, one saves memory and arrives at very efficient insertion and retrieval operations.

The distance between two successor structures Succ(a) and Succ(b) may be defined in various ways, the simplest of which is:

$$d(Succ(a),Succ(b)) =$$

$$d(Succ_1(a),Succ_1(b)) + ... d(Succ_k(a),Succ_k(b)) \quad (8)$$

$$d(Succ_i(a),Succ_i(b)) =$$

$$w_1 [Succ_i(a)_1 - Succ_i(b)_1] + ... + w_M [Succ_i(a)_M - Succ_i(b)_M]$$

where $(w_1,...,w_M)$ is a vector of weights and M=N*k*. This distance is a crude quantitative measure of the syntactic similarity of a and b, of how similarly a and b relate to the other symbols in the sequence s. The weight vector determines the statistical model underlying the distance measure.

The goal of Step 2 is a collection of categories with the property that each category consists of symbols which are, in an appropriate sense, mutually similar in their relationship with other symbols. This self-referential characterization of categories leads to a computational problem: if the membership of symbol a in category C_j can only be gauged by a's syntactic similarity with other symbols in category C_j, then where does one start? This problem is solved by the idea of a

self-organizing tagging algorithm. One constructs a dynamical iteration on the space of categorizations, the fixed points of which are satisfactory categorizations. Then, beginning from a random categorization, one iterates until one reaches the fixed point. In this way, the tagging is allowed to progressively construct itself.

The idea of self-organizing tagging has deep connections with the cognitive models of (Goertzel, 1994); perhaps surprisingly, it also lies at the root of one of the standardcategorization methods from statistics, the k-means method. Thus, the current implementation of the CLA carries out the crucial categorization step by k-means clustering on the space of successor structures. The k-means method has numerous shortcomings, but it is a "rough and ready" self-organizing tagging method which is eminently suitable for an initial implementation.

Finally, the third step of the CLA is the one which is most "wide open" in the sense that it involves a somewhat arbitrary measure of grammar quality. In the experiments with mood cycle data, to be reported in Chapter Eight, this step has been foregone, and the results of k-means clustering on successor structures have simply been used to derive grammatical rules. This approach has the advantage of simplicity. In general, however, it is not adequate. In (Goertzel and Goertzel, 1996) it is shown that without iterative improvement in Step 3 of the CLA, the CLA is incapable of inferring the natural partitions underlying simple mathematical iterations on the unit interval. The "three-halves" map discussed here is an example of this phenomenon.

Here we have focused on purely dynamic pattern; however, similar methods may be used for purely static or static/dynamic pattern. The only difference is the dimensionality of the entity from which repeated structures are extracted. Instead of a sequence of code numbers, one will have, in general, an n-dimensional lattice of code numbers. Instead of pairs, triples, etc. of code numbers, one will have n-cubes of code numbers of diameter 2, 3, etc. The same fundamental method may be used, but the computational complexity will obviously be higher. Purely dynamic pattern is the simplest case, and hence a natural starting point. Crutchfield (1991) has already applied his related epsilon-machine algorithm to the static/dynamic patterns traced out by two-dimensional cellular automata over time, with only partial success but very interesting results.

The CLA and Chaos Theory

To see what the CLA means in **chaos theory** terms, recall that attractors are typically divided into categories, such as fixed points, limit cycles, strange attractors, or chaotic attractors. The technical definition of the latter terms is a subject of some minor technical debate; my preference, followed here, is to call any attractor that is neither a fixed point nor a limit cycle a strange attractor. Strange attractors are often though not invariably associated with chaotic behavior. "Chaotic behavior" also admits several definitions. The one quality that is invariably associated with chaos, however, is sensitive dependence on initial conditions.

Chaos implies the unpredictability of the precise numerical state of a dynamical system. But this kind of unpredictability does not imply a similar unpredictability on the level of algorithmic structure. Walter Freeman, for example, has written about "attractors with wings" -- attractors with a pronounced sectional structure, each section corresponding to a certain type of system

state (Freeman, 1991). Even though the system is chaotic, if one knows which section of the attractor the systemis in, one may make meaningful statistical predictions regarding the section of the attractor the system will visit next. This is a simple example of the kind of attractor structure that can be captured by a formal language.

The Generalized Baker Map Revisited

To illustrate the application of the CLA, let us return to the "three-halves map" (Goertzel et al, 1993):

$x_{n+1} = (1.5\ x_n) \bmod 1$ (9)

This iteration, begun from almost any x_0 in the unit interval, leads to a chaotic trajectory on the interval. But this trajectory, though chaotic, is not devoid of dynamic pattern. If one divides the unit interval into 10 equally sized nonoverlapping subintervals, a trajectory of the system is encoded as a series of integers from 0 to 9, where the tag i represents the subinterval [i/10,(i+1)/10]. The CLA, with appropriately chosen parameters, is able to form the optimal categories

Category 0:

0 1 2 3 4 5 6

Category 1:

7 8 9

These categories give rise to natural grammatical rules. For instance, one finds that

00

01

10

are all grammatical constructions, but "11" is forbidden. Just as "V V" is forbidden in English (one does not have two consecutive verbs, though there are apparent exceptions, e.g. gerunds), "11" is forbidden in the language of the three-halves map. Similarly,

000

010

100

001

are all permitted, but "101" is not.

These grammatical rules are (approximate) patterns in the system. They allow one to make predictions about the system: if at time t the system is in a state that falls into Category 1, then at time t+1 it will definitely **not** be in a state that falls into Category 1. If at time t it is in Category 0, and at time t-1 it was in Category 1, then at time t+1 it **must remain** in Category 0. Furthermore there are statistical predictions, e.g.: if in Category 0 at time t, there is a 2/3 chance of remaining in Category 0 at time t+1. This kind of prediction can be madedespite the underlying chaos of the dynamical system. The key is in choosing a categorization which leads to a suitably restrictive grammar. This is what the CLA attempts to do. In the case of the three-halves map, the CLA can find the correct partition only if grammatical information is given enough weight in the tagging process. Purely statistical information is in this case misleading and the most effective tactic is to explicitly search for the tagging that leads to the most informative grammar.

5.5 ORDER AND CHAOS IN MOOD FLUCTUATIONS

In this section I will describe a simple attempt to use the CLA to analyze actual psychological data (rather than data generated by a mathematical function such as the Baker map). The focus is more on the methodology than on the actual results, which are quite tentative. The data in question concern human mood fluctuations.

Moods are complex things; and, as we all know from personal experience, they are difficult to predict. They are influenced by multiple, interlinked factors. Most obviously there are biological rhythms, and events in the external environment. There are also the possibilities of purely psychological rhythms, not directly tied to any specific chemical processes; and cooperative rhythms, generated by the coupling of the moods of two different people (e.g. husband and wife).

Combs et al (1994) reports an empirical study of mood fluctuations. In his experiments, each participant recorded their mood half-hourly during all waking hours of the day, for a total of 550-750 mood reports per participant. Each mood report consisted of markings on two Likert scales: one indicating excitement vs. relaxation, the other indicating happiness vs. sadness. These mood reports may be understood as points in the plane, and the data set for each participant may then be interpreted as the trajectory of a dynamical system.

Based on this data, Combs et al (1994) concludes that mood fluctuations display the characteristics of **chaotic** dynamics (i.e., dynamics which are neither stable, periodic, nor random). However, the identification of chaos here cannot be considered conclusive. The intrinsic fuzziness of the data, together with the short length of the time series and the breaks in the time series representing periods of sleep, all combine to render a truly rigorous numerical analysis impossible. All that can be said with certainty is that the data could plausibly be the result of noisy chaotic dynamics. On the other hand, they could also be the result of simpler underlying dynamics polluted by noise.

In this section I will reinterpret the Combs data using the tools of symbolic dynamics, rather than numerical dynamical systems theory. In the symbolic dynamics approach, as pursued here,

one does not need to ask whether the underlying dynamics are chaotic or merely periodic, and one does not need to question the statistical properties of the environmental perturbations or "noise." Instead, the focus is on the emergent statistical/algorithmic patterns in the time series. Thequestion is: what patterns are there that might allow us to make approximate predictions about a person's mood at a given time, based on their moods at previous times? Specifically, a computational method called the Chaos Language Algorithm (CLA) is introduced, which infers probabilistic regular grammars as emergent patterns in discretized versions of time series.

Like Combs et al (1994), this is primarily an exploratory, methodological investigation, rather than an attempt to posit a solid new theory of mood fluctuations. Complexity science is still in an early stage of development, and the psychological applications of complexity are particularly undeveloped. The need at the present time is for the clarification of basic theoretical concepts, and the development of innovative, theoretically-informed techniques for data analysis. The theoretical concept underlying the present investigation is that the analysis of complex psychological data requires a focus on **emergent algorithmic patterns** rather than merely numerical variables. The application of the CLA to mood fluctuation data is presented as a concrete exploration and illustration of this abstract proposition.

The Dynamics of Human Moods

What are the results of the CLA applied to the Combs data? No tremendously striking results have been found. Roughly and intuitively speaking, the essential result is that each individual has a certain "normal mood," represented a region of the state space of their "mood trajectory." The movements away from the normal mood seem to be either random or chaotic. Once the system moves away from normal, into a different part of the state space, it is likely to stay there a while, before wandering back. However, certain parts of the "non-normal" region would seem to have more internal coherence than others, as represented by a longer average residency time. These relatively coherent regions would seem to indicate relatively integral non-normal mood states; they also tend to be fairly similar to the normal mood state. It should be noted that the Combs study did not involve individuals with manic-depressive disorder, multiple personality disorder, or other conditions involving severe mood swings.

For the analysis of the Combs data, the state space of the two-dimensional mood system was divided into 36 equally-sized regions. The CLA was set to divide these 36 regions into 3 categories. The regions are defined by dividing the rectangle formed from the minimum and maximum values on each dimension into 36 equally-sized squares. The numbering for the regions is the natural one, given by

1 2 3 4 5 6

7 8 9 ...

13 14 ...

...

The choice of the numbers 3 and 36 here is somewhat arbitrary; they were arrived at by experimentation. However, since the original data were recorded on a Likert scale, sixcertainly seems a reasonable degree of granularity. Each of the two scales is being divided into six regions. Variation within these six regions may easily be ascribed to "measurement noise." Regarding the number of categories were also tried (2,4, and 5, specifically), but 3 seemed to provide the maximally informative analysis of the Combs data. Given the tendency of the systems involved to cluster in certain narrowly defined regions, multiplication of categories serves only to divide up infrequently-visited regions into categories in pseudo-random ways.

There was moderately significant inter-subject variation in the CLA results. For two subjects, no clear categorizations was found: different runs of the CLA led to a disparate variety of different categorizations. Out of the five subjects considered in the Combs study, only two gave rise to particularly clear results.

Results for one of these subjects were as follows. The bounding rectangular region was:

$4.000000 < x < 30.000000$

$2.000000 < y < 40.000000$

The inferred categories were:

Category 0

17 23 6 7

Category 1

0 1 10 11 14 18

19 2 20 22 24 28

3 5 8

Category 2

12 13

Regions 12 and 13, isolated here as Category 2, represent a normal mood for this subject. They are adjacent regions and thus represent, in essence, a single mood state.

By far the most salient pattern in this subject's data is: Usually in regions 12 and 13, moving back and forth between one and the other, with occasional deviations to other mood states. However, other patterns beyond this are present. The regions isolated in Category 0 display a certain "coherence" beyond that which an arbitrary collection of regions would display. They lead to each other slightly more often than they lead back to the normal mood state. Category 1,

on the other hand, is a "garbage collection" category, a set of apparently unrelated mood states. Thus, categories 6 and 7 represent a weakly defined but identifiable "non-normal mood."

The first-order transition frequencies between the categories are as follows (where "a b # x" indicates that a transition from region a to region b was observed x times in the subject's data set):

0 0 # 52.000000

0 1 # 17.000000

0 2 # 46.000000

1 0 # 16.000000

1 1 # 37.000000

1 2 # 45.000000

2 0 # 47.000000

2 1 # 44.000000

2 2 # 419.000000

The second-order transition frequencies are less revealing, but do contain some surprises. For instance, we have

0 0 0 # 29.000000

0 0 1 # 4.000000

0 0 2 # 19.000000

suggesting that, once the non-normal mood has persisted for two reports in a row, it is particularly likely to be observed in the next mood report. In other words, this mood state has a slight but noticeable tendency toward persistence. This is the sort of tendency that would be more convincingly observable from a longer data set.

The CLA does not always arrive at the same categorization. However, when an appropriate number of categories is used, there usually is only minor variation between different runs. The categorization given above was the most common for the subject in question; the second most common was the following:

Category 0

0 1 10 11 14 17

19 2 20 22 24 28

3 5 8

Category 1

18 23 6 7

Category 2

12 13

The third most frequent state was identical but contained both 17 and 18 in the category along with 23, 6 and 7. The occurence of a variety of different categorizations is not problematic: it only indicates that the CLA, which is itself a complex dynamical system, has a number of attractors clustered close together in its state space. The attractors with larger basins will tend to be found more often. It generally seems to be the case that the higher-quality categorizations have larger basins; but this is not known to be the case in general.

Another subject gave rise to the following results. Based on the bounding rectangular region

4.000000 < x < 34.000000

4.000000 < y < 37.000000

the following categories were obtained:

Category 0

12

Category 1

0 10 16 2 20 22

26 5 6 8 9

Category 2

1 11 13 15 17 18

21 3 7

0 0 # 86.000000

0 1 # 9.000000

0 2 # 36.000000

1 0 # 9.000000

1 1 # 184.000000

1 2 # 65.000000

2 0 # 36.000000

2 1 # 66.000000

2 2 # 196.000000

Here the "normal" mood 12 was not entered into all that often. However, the other two categories display a definite difference in their behavior with respect to Category 0. Category 1 has a very strong tendency not to lead to category 0. It, rather, leads to itself, or else to Category 2. Category 2, on the other hand, has a fairer chance of leading to Category 0. Similarly, Category 0 almost never leads to Category 1. This is a significant grammatical rule. The word 10 is infrequent, almost "ill-formed" in this grammar.

An obvious interpretation of this is that Category 2 consists of predominantly "near-normal" states -- after all, Categories 2 contains three of regions 12's nearest neighbors (11, 6 and 18). However, the key point is that 12 was coherent enough to consistently emerge as its own category, distinct from the "near-normal" states. This person has a definite pattern of moving from normal mood states, to near-normal mood states, then sometimes back to normal mood states, or sometimes to further outlying mood states. This pattern may also be inferred from the first subject, considered above. However, it is not so marked there due to the extreme frequency of the normal mood state.

Another, much less frequently obtained categorization of this subject's data, is as follows. Here 11 and 12 are bunched together as the "normal mood." The other two categories seem on the surface quite different, but a similar qualitative pattern emerges.

Category 0

16 17 6 7 8

Category 1

11 12

Category 2

0 1 10 13 15 18

2 20 21 22 26 3

5 9

0 0 # 266.000000

0 1 # 49.000000

0 2 # 34.000000

1 0 # 44.000000

1 1 # 141.000000

1 2 # 9.000000

2 0 # 38.000000

2 1 # 7.000000

2 2 # 100.000000

Here, once again, we see certain "forbidden expressions": 12 and 21. The composition of Category 2 is somewhat pseudo-random in the sense that moving a low-frequency region from Category 2 to Category 0 is not going to make much difference to the overall quality of the categorization, and may well lead to another attractor of the CLA. However, the same general pattern is observed here as in the more common categorization: one has a group of outliers (Category 2) and a group of near-normal moods (Category 1).

As noted above, there is no theory of statistical significance for the inference of grammatical patterns from symbolic data sets. Thus, these results must be considered as intuitive and exploratory rather than rigorous. In the end, however, most tests of significance are based on rather arbitrary statistical assumptions; and the ultimate evaluation of results is always intuitive. Linear statistics overlooks subtle patterns in complex data, but these patterns are there, and must be taken seriously. The CLA is one technique for eliciting such patterns.

Speculative Extensions

Putting aside for the moment the practical problems of analyzing the Combs et al data, let us now speculate as to what sort of patterns one might expect to find in an hypothetical "really good" set of human mood data. To keep things simple, instead of looking for optimal categories, let us think about only four categories:

A -- happy and excited

B -- happy and relaxed

C -- sad and excited

D -- sad and relaxed

Given this coding, the data set produced by a person over a period of time would look like

AABCDBCADCCBBCDDACCCDCCDCAAABBC...

The idea of symbolic dynamics is that list of letters is going to harbor *implicit grammatical rules*.

What kinds of linguistic rules might arise in these time series? The simplest rules are *first-order constraints*, constraints on what tends to follow what -- these, as discussed above, are the kinds of rules that have been observed in the real mood fluctuation data so far. One might find, for instance, that a person rarely moves from A to D or from D to A -- that the pairs AD and DA rarely appear in the series. This leads to the probabilistic grammatical rule that an A tends to be followed byan A, B or C. Or one might find that a certain person rarely goes from B to C or from C to B. These constraints would be pretty reasonable. Who goes from happy and excited all the way to sad and relaxed, without passing through intermediate stages?

Pushing a little further, one might find rules saying: "A has got to be followed by B or by C." Or, say: "A will nearly always be followed by B or by C." It could even happen, conceivably, that A, B, C and D only occurred in a few combinations, say ABACAB, DABACA, BABA... -- then one could look for patterns in the occurence of these combinations, these emergent "words."

These first-order transition rules may not seem like a very interesting kind of "language." They are what the Chaos Language Algorithm, in its present incarnation, looks for -- but they are only the beginning. For starters, there will be higher-order transition rules, i.e. rules involving sequences of length greater than two. An example would be: "If A is followed by C, then D is **very** likely to occur." In other words, ACD is probable, while ACA is unlikely. According to this rule, having been happy and excited, then sad and excited, one is not that likely to become happy and excited again; it is much more probable that one will get sad and relaxed.

In general, one finds that the simpler rules are more universal in nature, whereas the more complicated rules will tend to be more individual. For instance, while ACD may be a very likely sequence for some people, for others it may be rare. A manic-depressive person might well tend to have excitement stay high while happiness and sadness oscillated back and forth, meaning, in symbolic dynamics terms, a lot of cycles ACACACA..., or perhaps ACDACDA... or ACDCACDCA....

The patterns we have been discussing so far are all **context-free** rules. The next step up, in terms of complexity, is to context-sensitive rules, such as: "A follows B, but only in contexts where A occurs between C and D." Context-sensitive rules can always be summarized as collections of context-free transition rules, but they are more ordered than arbitrary collections of

transition rules. They are higher-order patterns. For instance, one can list the following permissible sequences:

CABB

AABD

CABB

BACC

...

or else, using a context-sensitive rule, one can say something simple like: *XABY is only a rule when X is C and Y is D.*

Examples of context-sensitive rules are **transformation rules**, such as one finds in the transformational grammar approach to human languages. The simplest transformation rules are movement rules, such as "Whenever you have ABCDB, you can replace it with BCDAB, and still have a permissible sequence." In this rule the A is moving from the first to the penultimate position in the word, without disrupting the "meaning."

One can, potentially, have very complicated situations here. For instance, suppose

ADCBD

ACBCC

ADBDD

ABDCB

are all permissible. This would lead one to believe the rule would be

AX__X

But instead, it might well happen that the rule

ABCDB

that fits the pattern isn't a rule; instead one might have something like

BCDAB

as a rule instead. This situation, if it were to occur, would be a symbolic-dynamics analogue of what one finds in human linguistics, where

* I know the person was where

is ungrammatical even though

I know the person was here

I know the person was there

I know the person was in Topeka

I know the person was higher

...

are all permissible.

Based on the limited data that have been analyzed so far, there is no reason to believe that such subtle patterns actually exist in human moods. On the other hand, there is no reason to rule it out either. Mood fluctuations appear chaotic, but we know that chaos is just unpredictability in detail. Statistical and algorithmic predictions can still be made about chaotic systems; and symbolic dynamics is the foremost tool for making such predictions. Perhaps when we "get to know" a person, part of what we are doing is recognizing higher-order linguistic patterns in the time series of their behaviors.

Remarks on Modelling Mood Fluctuations

The model of mood fluctuations suggested by this data, and these speculations, is a simple one. Much as Combs (1996) views states of consciousness as attractors of psychological systems, one is here led to view mood states as psychological attractors. Individuals' mood reports are functions of their mood state attractors. In the case of a normal individual, there is a "normal" mood state attractor, in which the person tends toreside a majority of the time. They also have other mood state attractors, which they move in and out of.

The nature of the underlying psychological system is not really relevant for this qualitative model of mood fluctuations. However, it is well worth observing how well this understanding of moods fits in with the psynet model. The psynet model view mood states as autopoietic magician systems. In this view the attractors observable in numerical data such as that collected by Combs et al are **images**, in the numerical domain, of more fundamental attractors in the **process** domain.

In the laboratory, after all, one cannot directly study mental processes. One can only study observable "properties" of mental processes -- and these are usually **numerical** properties. The

attractors of these numerical data sets reflect but do not entirely capture underlying process dynamics. Some of the noise and "messiness" involved in psychological time series data has to do with data-set limitations such as small size, but some of it also has to do with the inherent inaccuracy involved in the step from largely unknown process dynamics to observable but meaning-obscure numerical dynamics.

Freeman (1995) has observed, in the context of his analysis of EEG data, that the mathematical concept of "attractor" is really too rigid to apply to complex biological and psychological systems. A real psychological attractor, he points out, is only approximately defined: it is not guaranteed to be attracting, and once it has been reached, it may eventually be moved away from. This complaint, in my view, has to do with the fact that mental entities are attractors of magician systems, process systems, rather than numerical systems such as Likert scales or EEG readings.

Symbolic dynamics provides a way of abstracting away some of the peculiarities imposed by numerical measurements. What is meant by an "attractor" in a symbolic dynamics context is simply a stable region of state space: a set of symbols which tend to map into each other with reasonably high probabilities. These are the "attractors" that the CLA finds. The idea is that the **symbols** identified in numerical space will correspond to meaningful **symbols** in mental process space. One does not need the particular numbers involved in the experiment to mean much of anything, individually.

Returning to mood, one may well ask whether the mood state attractors are moved in and out of **spontaneously**, or only in response to external perturbations. This pertains to the question of whether chaos is present. One might hypothesize that each mood state attractor represents a wing of a strange attractor. The chaotic dynamics keeps the mood within that wing for a period of time, but sometimes it may move the mood out of the wing into another one.

Unfortunately, however, the question of spontaneous mood transitions cannot be answered on the basis of the Combs data. Furthermore, this inability does not seem to be a consequence of easily remedied flaws in this particular data set (e.g. short length, small dimensionality of mood reports, etc.). In general, the degree of environmental fluctuation in natural life is such that there will always be "reasons" to make the transition from one mood state to another. Thus, to distinguish natural,intrinsic transitions from extrinsically-induced transitions would be an extremely difficult task. Intuitively, one suspects that, with a similar data set perhaps ten to twenty times longer, one might be able to start to make some headway on this issue. On the other hand, the situation might well be the same. It may be pertinent to observe that, in the course of ordinary life, people often have extraordinary difficulty determining whether their own mood transitions are internally or externally induced.

Conclusion

This section reports an exploratory exercise in data analysis, involving the elicitation of emergent statistical/algorithmic patterns from complex psychological data. The data set in question, the Combs data on mood fluctuations, is not a "high quality" data set: it is very short, very noisy, and is riddled with breaks. Unfortunately, however, it is fairly typical of data in

behavioral science. It is thus of interest to see just how much information can be squeezed out of such data.

On this point, the results here are somewhat equivocal. The number of large-basined attractors of the CLA would seem to depend on the quality of the data. If the data set is insufficiently informative, then there will tend to be a large variety of attractors, and a conclusive categorization and ensuing grammar are difficult to come by. This problem occurred severely with two of Combs' five subjects, and moderately with a third. Only for two of the subjects did the CLA have sufficiently regular behavior to justify detailed analysis of the results.

The results obtained for these two subjects do make intuitive sense. The pattern is one of a pattern of a normal mood with periodic deviations into non-normality. Non-normal moods, similar to the normal mood state, have a degree of persistence on their own. Jumps between normal moods and highly abnormal, outlying moods are unlikely; these unusual moods are generally reached via near-normal moods with their own internal coherence.

But, while these results are sensible, the clearest empirical conclusion to be drawn from this investigation is that, in order to obtain really good grammars for human mood fluctuation, one needs significantly better data than is provided by the Combs study. Since the amount of noise in the data would seem to be essentially outside the experimenter's control, what this means in practice is the collection of longer time series. Longer time series would allow clearer discrimination of various non-normal mood states and the transition between them. With the data used here, it is impossible to make inferences on such transitions with a comfortable degree of reliability.

This moderately pessimistic conclusion does not, however, cast doubt on the value of symbolic, pattern-theoretic methods as opposed to numerical methods of data analysis. What should be clear from the exploratory results presented here is that thetwo methods of analysis provide fundamentally different types of information. The numerical paradigm can tell us, in principle, whether or not chaos is present, how much chaos is present, and so forth; in the case of periodicity, it can tell us the period involved. This is useful information. The pattern paradigm, however, gives us information of more direct and intuitive psychological meaning. It tells us what kinds of states people get into, and how they transition between these states. The key point is that our natural-language understanding of psychological phenomena such as moods is based on the recognition of non-numerical patterns. Non-numerical, pattern-oriented tools for data analysis, such as the CLA, are thus entirely appropriate.

5.6 TWO POSSIBLE PRINCIPLES OF COMPLEX SYSTEMS SCIENCE

In the Introduction, the lack of precise, general laws in complexity science was mentioned. The need for an appropriate complex-systems "philosophy of science" was discussed. Instead of general laws, it was argued, one should expect complex systems science to provide us with **abstract heuristic principles** and **general computational tools**.

The ideas of this chapter follow this philosophy. The Chaos Language Algorithm is an example of a general computational tool. And, in this section, I will give some abstract heuristic

principles which go along with the CLA: the Chaos Language Hypothesis, and the Structure-Dynamics Principle. These "principles" are not at this stage proven; they are merely plausible hypotheses. However, they are eminently falsifiable. Further work on the inference of structure from complex systems will automatically put them to the test.

The Chaos Language Hypothesis, first of all, states that the CLA will tend to produce similar grammars even when applied to very different psychological or social systems. In other words, it claims that psychological and social systems demonstrate a small number of "archetypal" attractor structures, and that the attractor structures observed in real psychological and social systems approximate these archetypal attractor structures. Mathematically speaking, the approximation of these archetypes should reveal itself as a clustering of inferred formal languages in formal language space. Thus one obtains the following formal hypothesis:

Chaos Language Hypothesis: The formal languages implicit in the trajectories of psychological and social dynamical systems show a strong tendency to "cluster" in the space of formal languages.

This hypothesis suggests the following three-stage research programme:

1. By computational analysis of data obtained from empirical studies and mathematical models, try to isolate the archetypal formal languages underlying complex psychological and social systems.

2. Analyze these languages to gain an intuitive and mathematicalunderstanding of their structure

3. Correlate these languages, as far as possible, with qualitative characterizations of the systems involved

If the postulated phenomenon of clustering could be shown, this would be a way of using dynamical systems theory to find precise, useful mathematical structures representing the qualitative properties of complex psychological and social systems. And this would, needless to say, be an extremely significant advance.

The Chaos Language Hypothesis postulates similarity of structure across different systems. The next hypothesis to be discussed, the Structure-Dynamics Principle, also postulates a similarity of structure, but in a somewhat different context: not between different systems, but between the purely static and purely dynamic aspects of the **same** system.

The essential idea of the Structure-Dynamics Principle is that, in many cases, specific **components** of a system are able to place the **entire system** in well defined states. Consider a system S which possesses two properties:

1. The components S_i are each capable of assuming a certain degree of **activation**.

2. There are perceptible **pathways** between certain pairs (S_i,S_j) of system components, indicating a propensity for activation of S_i to lead to activation of S_j

The paradigm case here is obviously the brain: activation has a clear definition on the neural level, in terms of neural firing, and pathways are defined by dendritic connections and synaptic conductance. Similarly, if, following Edelman (1987) and others, one takes the fundamental components to be neuronal groups, one finds that the neuronal definitions of activation and pathways naturally extend up to this level.

Now let Sys_i denote the collection of global system states that immediately follow high-level activation of system component S_i. Then the question is whether the sets Sys_i will emerge as natural categories from application of the CLA (or some more sophisticated, related algorithm) to the trajectory of the overall system S. Suppose this is the case: then what we have is a grammar of overall system states corresponding to the "grammar" of individual system components that indicates the order in which system components will be activated. But the latter grammar is, by the second assumption above, perceptible from purely **structural** information, from the observation of pathways between system components. Thus one has a grammar which is both a purely structural pattern (a pattern emergent in the collection of pathways in the system at a given time) and a purely dynamical pattern (a pattern emergent in the symbolic dynamics of the trajectory of the system). This is a most remarkable thing.

At this stage, there is no hard evidence that the Structure-Dynamics Principle is obeyed by real systems. However, the hypothesis is a plausible one, and, especially in the context of neuroscience and AI, it would go a long way toward solving somevexing problems. A little later, I will briefly consider the problem of knowledge representation from the perspective of the Structure-Dynamics Principle.

5.7 SYMBOLIC DYNAMICS IN NEURAL NETWORKS

With the concept of symbolic dynamics under our belts, we may return to the idea raised at the end of Chapter Two: that symbolic information is encoded in the brain/mind in the form of **attractor structures**. Symbolic dynamics gives a straightforward way by which logical and linguistic information can emerge out of complex, chaotic dynamics. It fits the psynet model like a glove.

But can brain systems really read each others' symbolic dynamics? In the end this is a question for the neuroscientist. As a mathematical psychologist, however, one may still gather valuable, pertinent evidence. For instance, one may ask: Are there biologically plausible ways for neural network architectures to encode and decode information in the symbolic dynamics of other neural networks? This is the question that we will address here.

Regarding encoding, first of all, it is quite possible to envision *training* neural networks to display given symbolic dynamics structures. This may even be done "blindly": it seems that the genetic algorithm may be used as a tool for evolving neural networks with given symbolic dynamics. Using networks of five to ten neurons, each one following the "discrete chaotic neuron" iteration of G. Mayer-Kress (1994; this is not a very simple formal model but a more complex iteration which attempts biological realism), and a population size in the range 20-50, it is possible to evolve neural networks whose strange attractors display various patterns of permissible and impermissible words (e.g., the second-order pattern of the generalized Baker

map, in which 11 and 101 are disallowed). These are only simple, preliminary experiments, but they indicate that evolutionary processes are at least plausible candidates for the "training" of neural networks to produce formal languages. Given the existence of a strong body of evidence for the evolutionary character of neurodynamics and learning (Edelman, 1988; Goertzel, 1993), this is an encouraging datum.

Next, the question of **decoding** brings us back to the Chaos Language Algorithm. For the question is: Are there plausible neural network models that infer grammars from dynamics, in particular, from the dynamics of other neural networks.

The CLA, as currently programmed, is a fully "symbolic" implementation of the Markovian framework for grammar induction. It uses special trees for successor structures, and carries out categorization using the k-means method on tree space. However, there is no escaping the naturalness with which Markovian grammar induction lends itself to neural network implementation. Though relatively inefficient when run on today's serial computers, such neural network designs are of considerable theoretical interest. In this section I will describe one such design, which I call the "Markovian Language Network" or MLN. The MLN gives an affirmative answer to the question posed above. Yes, it says,biologically reasonable neural networks **can** infer grammars.

The Markovian Language Network

Alexander (1994) has designed a formal neural network that

recognizes repeated subsequences in strings. This network is "recursively modular" in structure, meaning that it consists of clusters within clusters within clusters.... This sort of network architecture is particularly interesting insofar it has been proposed to be essential to brain function (Alexander, 1995) (this idea will recur in the final section). Alexander's results show that the construction of successor structures can be carried out in a "biologically natural" fashion. The drawback of Alexander's network is that it requires a large number of iterations to carry out the recognition task, which makes serial simulation extremely slow. Recognizing the subsequences of length up to 4 in a single sentence can take a Sun workstation several days. However, calculations show that, if this algorithm were implemented in the brain, it would operate with acceptable speed.

Categorization, on the other hand, has been studied much more thoroughly by the neural network community. The most elegant and biologically sensible neural network model of categorization is Kohonen's self-organizing feature map (Kohonen, 1988). A feature map can be used to map n-dimensional successor structure space into the 2-dimensional space of a neuronal lattice. Categorization is then carried out by forming word classes from words whose successor structures activate neurons in the same region of the lattice.

Consider, then, the following architecture, consisting of four modules:

Module 1: a recursively modular network computing successor structures (Markov information) from a linear array of inputs representing linguistic units

Module 2: a self-organizing feature map, applied to the output of the Markov-information-gathering network, sorting the successor structures (and hence the original morphemic units) into categories.

Module 3: a simple network mapping the input array into a new array in which two elements have the same state if they correspond to two linguistic units in the same category. The output of this network is the "tag array."

Module 4: a recursively modular network (perhaps the same one that has been called Module 1) computing successor structures from the tag array, to be interpreted as "rules," and, possibly, to be fed into Module 2 again.

This type of neural network is what I call the "Markovian Language Network." The output of Module 4 is a list of the grammatical rules implicit in the input array. More sophisticated variations on the network can also be given, for instance, designs which incorporate a priori information on word categories in various ways.

The MLN requires only the addition of two extra "preprocessing" step to become a neural network implementationof the CLA. Specifically,

Module 5: a self-organizing feature map, used to provide a categorization of a neural system's behavior, mapping the network's states into a collection of categories.

Module 6: a simple network mapping trajectories of the neural system being studied into category labels ("code symbols"), based on the results of the feature map (in the manner of Module 3 of the MLN).

The code symbols are the linguistic units which the Markovian network in Module 1 of the MLN architecture receives in its input array. An MLN can thus be used to infer the grammatical patterns arising from another neural network's dynamics. A neural network is inferring the emergent grammatical patterns overlying the chaotic behavior of another neural network.

This train of thought has, it would seem, immense potential importance for neuroscience. Since the pioneering work of Walter Freeman (1993), it has become increasingly clear that strange attractors exist in neural systems and play an important role. Alexander (1995), generalizing Freeman's observations, has hypothesized that chaos is used by the brain on many different levels, as a general strategy for search and innovation. The MLN fits into this train of thought quite nicely: it provides a mechanism by which one neural system might understand and predict the complex, chaotic dynamics of another.

Pushing this line of reasoning to its natural conclusion, one might propose to understand the the brain as a whole as a system of neural networks mutually inferring languages in one another's behavior. This concept, while admittedly speculative, has deep meaning for cognitive science, as it hints at a possible strategy for bridging the gap between symbolic and neural network ideas. **Chaos**, in this view, may be seen as the link between logical, linguistic formalism and messy, self-organizing neurodynamics. And language is not a specialized, unnatural "add-on" to the

routine processes of brain function, but rather, part and parcel of intermediate-level neural net dynamics.

1.8 DYNAMICS, PATTERN AND ENTROPY

Now, for the final section of this chapter, let us turn from psychology to physics, and look at some of the **thermodynamic** implications of dynamical patterns in system behavior.

My own slant on complex systems science is avowedly computational. I am not a natural scientist but rather a cognitive scientist, with expertise in mathematics, computer science, psychology and philosophy. However, one may also take a quite different approach to the emerging science of complexity: an approach grounded in the **physics** of complex systems. In this section I will deal with the relationship between dynamics and pattern from a physics point of view, by exploring some of the relationships between **entropy**, **chaos**, and **algorithmic information**. As this material will not be explicitly taken up elsewhere in the book, the impatient reader may wish to skip itover.

First I will review a result which I call "Brudno's Equation," which relates the metric entropy of the trajectory of a dynamical system to the **computational** properties of this trajectory, using the notion of algorithmic information. Algorithmic information is an important concept which will arise again in later chapters.

Then I will give some inequalities due to Carlton Caves (1990), related to Brudno's Equation, which connect chaotic dynamics with computational complexity. Finally, I will discuss Baranger's Law (Baranger, 1993), a recent discovery in far-from-equilibrium thermodynamics which explains entropy production in terms of chaotic dynamics. Drawing these various ideas together, I will argue that **physical entropy bounds degree of chaos, which bounds algorithmic information** -- a nice relationship which joins thermodynamics, dynamical systems theory and computation into a single formula.

Algorithmic Information and Entropy

The **algorithmic information** of an entity, introduced above, is defined roughly as the length of the **shortest** program which computes it. If an entity has algorithmic information less than its length, this means it has at least one pattern. If one has an **infinite** entity, say an infinite sequence of numbers like

12345678910111213141516171819202l...

then the algorithmic information is defined as the value of

(algorithmic information of the first N digits)/N

for N very large.

This business of "program length" might seem a little dubious -- doesn't it depend on what computer you use? But this is exactly the crucial point. Suppose one has a huge collection of computers -- say, ten million or so. Some of them are PC's with built-in BASIC compilers, some are PC's without, some are CRAY supercomputers, some are tiny little Timex-Sinclairs from the late 1970's, some are sophisticated massively parallel wristwatch supercomputers from the year 2029,.... Then, if one takes a sequence of letters that is **long enough**, all the computers will basically agree on which programs are patterns in the sequence and which ones are not. The reason is that all of them are **universal computers**, and hence all of them can, in principle, simulate each other. The program for simulating another computer may be long, but there are only ten million computers, so there are at most (ten million) * (ten million) necessary simulation programs. Just take a sequence vastly longer than the longest one of these simulation programs. This fundamental insight is due to Gregory Chaitin.

But what does all this abstract computation theory have to do with entropy and chaos? In 1978, the Russian mathematician A.A. Brudno proved the following truly remarkable equation:

ALGORITHMIC INFORMATION = METRIC ENTROPY (10)

Now, this result is not completely general. It applies only to dynamical systems that are **ergodic** (a technical way of saying that every trajectory of the system must lead to basically the same behavior; time averages must equal ensemble averages). Or, for a dynamical system that is not "ergodic," it holds on the system's "ergodic set" ... the set of trajectories for which the system behaves as if it **were** ergodic. But most of the typical examples of **chaotic** systems are also **ergodic**. So Brudno's result, despite its limited scope, brings chaos theory up to a whole new level. It implies that, for **every single trajectory** of such a dynamical system, the algorithmic information is the same as the amount of chaos.

But the algorithmic information is just a measure of how **difficult** it is to predict the behavior of a system using computer programs. So we reach the following conclusion: for ergodic systems, truly chaotic systems, **chaos is unpredictability**. This is the most intuitive result in the world: this is exactly what chaos is supposed to be ... deterministic unpredictability. But Brudno's Theorem puts some meat on the idea: it explains **why** chaos is unpredictability.

The restriction to "ergodic" systems, though somewhat tiresome, cannot be escaped. These results are too powerful to hold with complete generality. But some recent work by the laser physicist Carlton Caves provides at least a partial generalization of Brudno's Equation. Caves has proved an equation which implies that, in **every case**,

METRIC ENTROPY < AVERAGE ALGORITHMIC INFORMATION <

2 * METRIC ENTROPY

(11)

(The reasoning leading up to this relation was borrowed from two Russians named Zvonkin and Levin, who in 1970 discovered a primitive form of Brudno's Equation).

This is not so precise as Brudno's Equation, but it still says quite a bit. In general, always, regardless of what kind of dynamical system we're dealing with, the average algorithmic information of a trajectory is trapped between the metric entropy and its double. Or in other words: **the average algorithmic complexity of a system is between one and two times the amount of chaos in the system.**

Entropy

Now let us turn from the mathematical notion of metric entropy to the physical concept of entropy. Operationally, in the chemistry lab, entropy measures the amount of **heat** that must be expended to set up a system in a given state, using reversible operations. But heat is really nothing more than **random molecular motion** -- and thus the entropy of a collection of molecules may be regarded as our degree of **ignorance** regarding the positions of the molecules. This what makes entropy so interesting. It is subjective and objective at the same time.

More formally, recall that each possible microscopic state of a system can be represented as a point in an abstract multidimensional space called "phase space." The entropy of a system in which **all states are equally likely** is just the **logarithm** of the total volume in phase space occupied by all the states of the system. And if all states are not equally likely, then one just divides the phase space up into regions R(1),...,R(N), assigns each region a probability p(i), and defines the entropy as the **weighted average**

$$-k[\, p(1)\log p(1) + \ldots + p(N)\log p(N) \,] \quad (12)$$

where k is a number called "Boltzmann's constant." A standard limiting process extends this definition to the case of an infinite number of divisions of phase space; the sum becomes an integral, and the entropy becomes coordinate-dependent.

The First Law of Thermodynamics says that energy is conserved for closed macroscopic systems, just as it is for microscopic systems. An **open** system, one which interacts with its environment, may **appear** to gain or lose energy ... but if one considered the overall system of "open system + environment," one would find energy conserved once again. The Second Law of Thermodynamics, on the other hand, states that entropy must always increase. Entropy can never, under any circumstances, go down; and in an idealized mathematical system it might stay the same, but in a **real** system it will always go up. In other words, we can never get more knowledgeable, only more ignorant. We may gain knowledge about those things which are important to us -- but in doing so we will randomize the motions of **other** molecules and thus create ignorance in **their** regard. Life itself is a process of localized entropy **decrease**: living systems are highly ordered; their very premise is predictability and hence low entropy. But living systems can only **maintain** themselves by creating heat and hence increasing the overall entropy of the universe.

Mathematically, when one puts the Second Law together with the First Law of Thermodynamics, one obtains a contradiction. The First Law, the macroscopic conservation of energy, implies that, as a system evolves, **volume in phase space doesn't change**. This conclusion follows from a result called "Liouville's Theorem." The region of phase space

representing a physical system can change its **shape** ... but its total area must always remain the same.

What gives? The facile answer is the the first law applies only to closed systems, whereas entropy is used to study open systems -- systems that are "dissipative" rather than "conservative." But this answer is actually no answer: every dissipative system is just an approximation to some conservative system. Every open system can be closed: just add in the environment, and consider it part of the system.

The **real** answer is much more interesting. In fact, it is the apparent contradiction between the First and Second Laws that gives entropy its meaning. The trick is that the region of phase space representing a system can **spread out** like a multidimensional octopus -- it can get stringy, and distribute itself very **thinly** over a very large range. This doesn't changeits **mathematical** volume. But what happens when an actual person goes to **compute** the volume ... to measure the system?

A standard observation in fractal geometry is that to compute the length of the coastline of Britain, one has to use a **ruler** of finite length. And the length of one's ruler determines the **measured** length of the coastline. The actual coastline doesn't change, of course; it's a matter of how many inlets and mini-penninsulas the ruler takes account of. Where fractals are concerned, measurement must take place on a certain specific **scale**.

A related conclusion holds concerning the measurement of volume in phase space. In order to actually **measure** the volume of a region, one must use some "measuring instrument" with a finite scale. For instance, one can divide phase space up into little tiny boxes and count **how many boxes** contain part of the region.

Multiplying by the volume of a single box, this should give the total volume of the region concerned, right? Well ... not exactly. The problem is that, when a region spreads out and gets all stringy, it will have a **little tiny portion** in a great number of regions. So it will **appear** to take up more space than it actually does.

This, as it turns out, is the source of entropy. Though it is measurable in the chem lab, entropy is nonetheless at bottom a **subjective** quantity. It arises from the necessity for **finite-scale measurement accuracy**. When measuring an evolving dynamical system, if one wants to keep up a constant accuracy of measurement, one has to continually measure on a **finer and finer scale**. But then one is no longer computing entropy on the same partition of states ... the entropies measured at two different times cannot be fairly compared. Entropy is free to increase ... as the Second Law says it must.

Entropy and Metric Entropy

So a system can be represented as a collection of states, each of which is a point in "phase space." And the region of phase space representing a system will tend to spread out, octopuslike, blurring the process of measurement. But why does this happen?

The answer is nothing but **chaos**. The mathematics of chaos theory may be used to make the informal idea of "spreading out like an octopus" quite precise. Chaos causes nearby trajectories to diverge from each other -- and hence blurs phase space, producing entropy. The following two conclusions, both simple mathematical results, are truly remarkable in their implications. They provide a complete mathematical justification for the informal "spreading out and getting stringy" argument that was used above to explain the origin of entropy.

First of all, there is an inequality relating the metric entropy with the **real** entropy:

METRIC ENTROPY < ENTROPY/k (13)

This is simple to interpret. It means that, if the entropy of an evolving system is small, then the system doesn't spreadthings out much. On the other hand, if the system spreads things out a lot, then the entropy is **big**.

And then there is an equation, actually a "limiting equation," regarding what happens once a system has been evolving for a **long time**:

RATE OF CHANGE OF ENTROPY --> METRIC ENTROPY (14)

Suppose that, as the system evolves longer and longer, one computes the **rate of change** of the entropy at every time step. How fast does the entropy increase? Eventually, this equation says, the rate of entropy increase is just about equal to the **amount of spreading** ... to the metric entropy.

These are not "deep" results; they follow easily from the technical definitions involved. But they do a most admirable job of connecting **chaos theory** with **far-from-equilibrium thermodynamics**. Subtracting all the technical stuff, what they say is quite simply that **chaotic systems have a lot of entropy, and they produce entropy faster than nonchaotic systems**. Note that we have not ruled out the possibility of a nonchaotic system with high entropy. But this system, if it exists, will produce entropy slowly. If a system produces entropy fast, it's got to be chaotic.

Baranger's Law

The Second Law says that entropy goes up. But how **fast** does it go up? One possibility is that it always **increases as fast as it can**, given the constraints of the particular situation. This principle of "maximum entropy production" was put forth in 1989 by the system theorist Rod Swenson; its philosophical implications were extensively explored by Sally Goerner in her 1993 book *The Evolving Ecological Universe*.

When one considers the matter carefully, one finds that Swenson's law of maximum entropy production is a serious overstatement. It is simply **not true** that, in all circumstances, entropy increases at the fastest possible rate. There are situations in which a system has a choice between

two states -- and it choose the one which produces entropy **more slowly**. A simple example is water being forced through a pipe. As it rushes faster and faster, it makes a transition from smooth, "laminar" flow to chaotic turbulence -- and during this transition, it passes through a stage known as "intermittency," in which smooth and turbulent flow rapidly succeed one another. In this sort of transition, pathways of rapid entropy production are routinely passed up in favor of those which are slowpokes in the entropy game. Swenson's law does not hold.

But there are also interesting situations in which Swenson's idea works. One example is a peculiar apparatus called a **Benard cell**. A Benard cell is a flat box filled with some fluid, often water. The bottom of the box is a metal plate, and the critical parameter of the system is the **temperature**. With no heat, the cell is random; the only action is random molecular collisions. Brief molecular inhomogeneities arise, but these are mere ephemeral fluctuations from the equilibrium state. And if you take a Benard cell and heat it up just a little, the heat willdiffuse slowly through the cell, returning the system gradually to equilibrium. But suppose instead you heat the cell up a **lot** -- that's where the fun starts!

Hot collections of molecules are lighter than cold ones -- and if a collection is hot enough, its buoyancy will be sufficient to overcome the viscous drag of the surrounding fluid, and it will float around as a coherent unit. Once a certain **critical temperature** is reached, some of these overheated collections will shoot up all the way to the top of the cell, and will pull other **fairly** hot molecules up along with it.

But then, what happens when the hot molecules finish the voyage to the top? The air outside cools them down ... they sink back down, pulling other **fairly cool** molecules with them. Then, once they get to the bottom, they heat up again. Up and down, up and down the molecules cycle -- thus setting a **trend** for other molecules to follow. Before long a whole region of the cell is rolling over and over. This is a classic example of **self-organizing** behavior. No one tells the cell what to do: it just organizes itself into a coherent high-level pattern, based on a tiny change in the critical parameter of temperature.

Why doesn't the cell return to equilibrium? In dynamical language, the rolling motion is a kind of **attractor**. As the molecules roll up and down, they create **friction** and pull in more energy from hotter parts of the cell -- so there is no way for the situation to settle down, for the energy difference to disappear. This would not be possible in a closed system -- but the system is dramatically open; it is rapidly dissipating energy, and increasing entropy. The system **keeps itself** away from equilibrium. Instead of the "fixed point" attractor of equilibrium, it converges to the more complex attractor of continual rolling motion.

Eventually, uniform "rolls" emerge, and occupy the whole cell. Very elegant, totally spontaneous -- and, as shown by the computer simulation of David Hogg (1992), **entropy production maximizing**. As one increases the temperature of the Benard cell, the system systematically chooses the pathway that **produces entropy fastest**. When it finally opts for pulsating rolls instead of statistical homogeneity, this is because the former situation zooms us more quickly toward heat death!

Two Types of Bifurcation

Fluid forced through a pipe doesn't obey maximum entropy production. Emergence of **rolls** in the Benard cell does. Thus, an intellectual puzzle emerges. What is it that differentiates the two types of experiment? When is entropy maximized?

As it turns out, one difference between the two is the **type of bifurcation** involved. A "supercritical" bifurcation is a situation in which, when a certain parameter of a system passes a certain critical value, the system reaches a "choice" between two mathematically possible paths. One path is stable, the other unstable. The stable path, in practice, is the one selected -- and the transition is, in a way, smooth ... there is no discontinuous "jump" in system behavior. A "subcritical" bifurcation, on the other hand, involves a sudden **leap** from one state to another -- such as when water running through a pipehops from smooth to turbulent flow. In technical language, a supercritical bifurcation has a continuous bifurcation diagram, while a subcritical bifurcation does not.

Rolls in a Benard cell arise through a supercritical bifurcation: when the temperature passes a certain level, the future of the system splits up into two possible paths, one stable, one not. Turbulent flow through a pipe, on the other hand, arises **sub**critically: there is no smooth motion toward a "choice point" ... as the speed of the flow increases, there is a sudden **jump**.

Given these facts, analogical reasoning leads to the following conjecture: that Swenson's principle of maximum entropy production works only when a system is presented with a choice that is a **supercritical** bifurcation.

But, as it turns out, even this hypothesis is too strong -- another restriction is needed. What is necessary is first of all that, before the bifurcation, the system's state have a certain **property** which is a "dynamical invariant."

Losing Structure through Chaos

A "dynamical invariant" is any property of a system that **doesn't change** as the system evolves. For instance, the Benard cell, until the bifurcation, has the property of **statistical homogeneity** ... meaning it looks the same everywhere. This property doesn't change as the system evolves: it's a dynamical invariant. Often, in physics, dynamical invariants take the form of **symmetries** ... for instance, the uniform appearance of the Benard cell may be characterized as a "translational symmetry," meaning that when you shift your eyes from one part of the cell to another (i.e. translate your coordinate system), the appearance of the cell remains exactly the same.

What is needed to make Swenson's principle work is that a system have a **dynamically invariant property**, which is **lost in a supercritical bifurcation**. In other words, the bifurcation splits the system's evolution into two "possible" branches, one stable and one unstable ... and the invariant property follows the **unstable** branch, the branch that doesn't happen. In this case, as it turns out, the **stable** branch without the property has a **higher entropy**. This, finally, is **Baranger's Law** -- proved by Michel Barenger in 1993, in a brief paper modestly entitled "An Exact Law of Far-from-Equilibrium Thermodynamics."

In the Benard cell, when one reaches the bifurcation point, the homogeneous, symmetric condition is **no longer stable**. There is, mathematically, a possibility for the system to continue in a symmetric way -- for the cell to forever keep from rolling. But this possible path never materializes, because it is **unstable** -- the slightest random fluctuation is enough to destroy it. On the other hand, the **stable** state, the complex self-organizing rolling behavior of the cell, fails to display the symmetry property that the system had before the temperature reached its critical value.

The proof is somewhat involved, but the **reason** for this beautiful law is not so hard to see. Consider: one never has **perfect obedience** to any invariant property -- in reality one has situations that are "almost homogeneous", "almost circular", etc. Even **before** the bifurcation, what one is actually observing is not a trajectories possessing a dynamical invariant, but a trajectory **almost** possessing a dynamical invariant. Before the supercritical bifurcation, these near-invariant-possessing paths stay near the "phantom" paths that **really** have the dynamically invariant property. Afterwards, though, chaos sets in ... large Liapunov exponents develop in directions **away** from the phantom paths, and the real paths lose the property. But we have seen that chaos -- Liapunov exponents, metric entropy -- is proportional to **rate of entropy production**. Nature, by following chaos, maximizes entropy production!

A complex line of reasoning ... but yet, a very simple conclusion. What we have concluded is, quite simply, that **smoothly losing structure increases entropy**. The "structure" may be symmetry or any other dynamical invariant. When a supercritical bifurcation occurs and this structure is broken, the structure-shattering chaos produces entropy. In the process, the system may well develop new, more complex structures, like the rolls in the Benard cell ... but the old invariant structure is lost.

Chaos and Computation

Let us return to the Benard cell, one more time. Unfortunately for Swenson's Law, the emergence of rolls is not the whole story. It turns out that, if the temperature is increased at just the right rate, the rolls give way to a sort of "honeycomb" of equally sized hexagonal cells. If the entropy went up here, one would have another example of entropy increase driving form creation.

But in fact there is no reason to believe this is the case. This is a **sub**critical bifurcation, not a **super**critical one. The hexagons, which are after all a more interesting structure than the rolls, are **not** formed by entropy increase at all.

What Baranger's Law **doesn't** get at is the **production of new structure** that often follows the destruction of an old structure. It explains the demise of the homogeneous structure of the Benard cell, but not the emergence of the rolling structure ... let alone the emergence of the hexagonal structure. But it is these phenomena, rather than the initial symmetry, which make the Benard cell interesting.

To see the relation between **chaos** and **structure**, we have to return to **computation theory**. Given the relationship between metric entropy and entropy, Brudno's Equation is easily seen to imply that

ALGORITHMIC INFORMATION < 2*ENTROPY/k (15)

So if a physical system has almost no entropy, all its trajectories will be very **simple** to compute. Only a system with large entropy can have algorithmically complex trajectories. But then, complex trajectories lead to high metric entropy, which means **rapid entropy production**. Everything connects with everything else. Abstract computational patterns are not allowed to forget their origins in an actual, physical dynamical system. Thermodynamics, chaos and computational complexity are threedifferent views of the same thing: complexity.

CHAPTER SIX

EVOLUTION AND DYNAMICS

6.1 INTRODUCTION

In this chapter we will return to the genetic algorithm, which was introduced in Chapter One. The relevance of the genetic algorithm to the psynet model has already been established -- GA's, it seems, are an abstract, archetypal model of a certain type of psychological creativity. Here we will be concerned with genetic algorithms as dynamical systems, and with the use of genetic algorithms to evolve other dynamical systems. Rather than merely cranking out genetic-algorithm applications, the focus is on understanding what the genetic algorithm is, what it can do, and why it is relevant to human and computer creativity.

We will begin by studying the mathematical dynamics of the simple GA. New mathematical tools will be introduced for studying the nature of evolutionary dynamics. Using these new methods, two important conclusions will be reached. First, it will be shown that, while GA's cannot be expected to converge to a globally optimal population, nevertheless, as population size tends to infinity, this kind of convergence gets **closer and closer** to happening. Finite population size is a generator of diversity, but a distractor from ultimate convergence.

Next, the convergence rate of the GA will be studied. An equation will be derived which suggests a new way of looking at the GA -- as a kind of progressively focussing stochastic search. The GA with crossover, it seems, displays a pattern of learning similar to that observed in humans and animals: a phase of rapid initial learning followed by slower phase of knowledge refinement.

A GA with mutation only gives the "wrong" shaped learning curve; a GA with crossover gives the "right" shape." This mathematical result, which is borne out by GA practice, is extremely intriguing from a psychological point of view. It suggests very strongly that, if we are to model minds as evolutionary learning systems, we **must** include crossover operations in the mix.

Finally, it is noted that evolution itself does not provide a complete explanation of any real-world complex system. There is always another side to the story -- one which represents a force of order-maintenance, as opposed to evolution's force of change and adaptation. In evolutionary biology this other side is called **ecology**. Ecology expressed the interrelations of things: it both constrains evolution and gives it power. In a psychological context, I will usually call this "other side" **autopoiesis**. The idea is the same. Evolution provides innovation, creativity. Autopoiesis keeps things alive, and generates emergent pattern binding different parts of systemstogether. The two processes are two different views of a unified life-process.

In order to get at the full picture, we will describe the Simple Evolving Ecology (SEE) model, a genetic algorithm generalization which is flexible and robust enough to be considered as a serious model of mind. SEE is a spatially distributed GA with ecological effects and a dynamically evolving spatial structure. Experimentation with SEE has just begun. Initial experiments, however, indicate that the behavior of SEE has a great deal to teach us regarding the behavior of complex psychological systems, and complex systems in general.

The accomplishments of the GA as an optimization and machine learning algorithm have been described in many other places, and I feel no need to review them in detail here. Instead I will report several experiments using the GA to evolve complex structures -- in particular, fractal structures. The GA will be used to evolve strange attractors for plane quadratic maps; and to evolve invariant measures for iterated function systems, which are interpreted as pictures or music.

The details of these applications have various connections to ideas described in other chapters. For instance, to understand the dynamics of the GA as it evolves strange attrctors, we will find it convenient to introduce the notion of **autopoiesis**; and to invoke the idea of "second-order evolution," as discussed in the previous chapter. In order to use the GA to produce musical melodies, we will need to arrive at a method for computationally assessing the "fitness" or "quality" of different melodies. As it turns out, a strategy very similar to the Chaos Language Algorithm will come in handy here.

These experiments may be considered in purely computational terms; or, speaking quite loosely, they may be interpreted in terms of cognitive science. One psychological application of the GA is to learning theory -- one wants to hypothesize that when we solve optimization problems of various sorts, we are using crossover and mutation operations on a pool of possible answers. Another, perhaps more interesting, application is to **creativity**. Here one is not trying to use the GA to simulate the process of answering a question, but rather to simulate the process of coming up with new forms. This is where the present experiments come in. One is, quite simply, trying to get the GA to come up with interesting things.

In traditional applications, where one uses the GA to solve problems, one is not exploiting the full information available in the GA population at a given time. One is extracting the "best" answer from the population and ignoring the many other interesting structures contained therein. In these creativity-oriented applications, on the other hand, the full population is used. As in real biological and psychological systems, the full **efficiency** of the evolutionary process is exploited.

6.2 THE DYNAMICS OF THE GENETIC ALGORITHM

Dynamical systems models are often categorized based on thecriterion of discreteness versus continuity. There are continuous systems, such as differential equations. There are discrete-time, continuous-space systems, i.e. dynamical iterations such as the Baker map. And there are wholly discrete systems -- such as the evolution game described above, and the genetic algorithm.

Each of the two varieties of dynamical system has its advantages and disadvantages. Most continuous dynamical systems of practical interest are too complex to study analytically; and so, in the study of continuous dynamical systems, one quickly resorts to computer simulations. One replaces one's continuous system with a nicely space- and time- **discrete** dynamical system. But even where they are intractable, continuous models are still often useful as guides for the intuition. For instance, differential equations models are naturally appropriate to physics, because the basic conceptual models of physics involve continuously varying quantities like position, momentum and energy.

In areas like population biology and cognitive science, on the other hand, the basic conceptual models tend to involve discrete entities: organisms, species, thoughts, neural assemblies,.... So one is naturally working with discrete models, such as the genetic algorithm. The disadvantage of working with discrete dynamical systems, however, is that, if one **does** wish to do mathematical analysis, the tools of calculus are not directly available. Computation is simple and immediate, but the extra understanding that comes with mathematical theory is difficult to come by. There are three options: to give up on analytical understanding, to struggle with discrete mathematics despite its limitations, or to look for ways to approximate discrete models by **continuous** systems, which are more easily analyzable.

In this section I will take the third option. I will describe some recent research in which the genetic algorithm, a discrete dynamical system, is approximated by a **continuous** dynamical system called the IMGA, and studied in this way. First I will present a convergence theorem for the simple GA with very large population size. Then I will turn to the question of how fast the GA converges. I will describe mathematical results which suggest that the GA follows a learning curve similar to that demonstrated by humans and animals: a power law, with rapid improvement at first followed by slower learning afterwards. This is a suggestive (though far from conclusive) piece of evidence that the **crossover** operation has something to do with biological learning.

Convergence, Convergence Rate, and the IMGA

Davis and Principe (1993) have shown, using methods from the theory of Markov chains, that the simple GA cannot, for any finite population size, be guaranteed to converge to a globally optimal population. However, they also present numerical results suggesting that, as population

size tends to infinity, convergence to a globally optimal population becomes more andmore likely. I have shown, in a recent paper, that this is in factthe case: that convergence to a globally optimal population holds in the limit as population size tends to infinity. This proof was obtained in the manner described above: by approximating the GA with a continuous iteration.

This continuous approximation to the GA is called the IMGA or "iterated mean genetic algorithm." It is a real nonlinear iteration which approximates the behavior of the GA arbitrarily closely as population size increases infinity. Because it is a differentiable function, one may treat it by the methods of calculus: one may calculate the Jacobian derivative of the IMGA, which turns out to be an exceedingly interesting quantity.

The first study of this determinant, given in an appendix to *The Evolving Mind* (Goertzel and Ananda, 1993) and in the Proceedings of the *COMPLEX94* Australian Complex Systems Conference (Goertzel, Ananda and Ikle', 1995), was marred by some persistent calculational and typographical errors. A corrected and completed treatment was presented at the 1995 IEEE *International Conference on Evolutionary Computing*, and is given in detail in (Bachman, Goertzel and Ikle', 1996).

To study convergence, the eigenvalues of the Jacobian are evaluated for the case of a globally optimal population. In the case of no mutation and a fitness function with a unique global maximum, these eigenvalues are bounded by 1 in magnitude, thus showing that a globally optimal population is an attractor for the IMGA. Furthermore, the eigenvalue formulas yield a bound for the asymptotic rate of convergence of the GA for large population size.

To study rate of convergence, on the other hand, we will deal with the Jacobian determinant. By a remarkable combinatorial circumstance, it is possible to give a simple formula for this determinant that holds at **all** points on the trajectory of the IMGA, not just in the vicinity of the answer. This formula gives valuable hints as to the behavior of the GA. In words, what it tells us is that: The fitter the population gets, the slower the GA converges.

Mathematical Apparatus

Throughout this section, we will consider a simple genetic algorithm, with a population of size N, consisting of M-bit binary sequences, evolving with non-overlapping generations, and selection probability proportional to unscaled fitness. For the purpose of the convergence proof we will set the mutation rate equal to zero. Finally, where f is the nonnegative-valued fitness function, we will use the notation $A_i = f(i)$.

To facilitate the expression of formulas involving crossover, we will represent a crossover operator as a collection ofprobabilities: P_{kli} will denote the probability of crossing bit string k with bit string l to obtain bit string i. If the equations

$$P_{jij} + P_{ijj} <= 1$$

$$P_{jji} = 0 \quad (1)$$

hold whenever i and j are unequal, then we will say that we aredealing with a true crossover operator. It is easy to see that the case of crossover at a single randomly-selected cut-point falls into this category.

Suppose that p_i gives the proportion of bit strings of type i in generation t-1. Then it is not hard to see that the expected proportion of type i in generation t is given by

$$Phi = (p_k p_l A_k A_l P_{kli}) / (p_r A_r) 2 \quad (3)$$

Where $p = p_1,...,p_K$), where $K = 2M$. The iterated mean genetic algorithm (IMGA) is defined by the formula:

$$pt = Phi\ (pt\text{-}1) \quad (3)$$

where the initial vector $p0$ represents the vector of proportions derived from the fixed initial population.

In order to show the validity of this approximation, one may represent the GA as a mapping from vectors of probabilities into probability measures on the space of probability vectors, *sigma*: [0,1]*K* --> P([0,1]*K*). This representation is exploited extensively by Davis and Principe in their Markov chain analysis of the simple GA. Where *sigmar* denotes the mapping of this sort resulting from r generations of the GA, the following lemma is easily obtained.

Lemma. Fix an initial vector of proportions p, a number of iterations t, an *epsilon* > 0 and a *theta* < 1. Then there exists a number N′ so that, whenever the population size N exceeds N′,

$$Pr\{ |pt - sigmat(p)| < epsilon \} > theta \quad (4)$$

Though the function *Phi* is a perfectly accurate representation of the iterated mean path, it is sometimes useful to introduce a different representation, one which takes into account the implicit contraint that all the arguments of *Phi* must sum to one. By replacing *Phi* with the function *Psi* defined by

$$Psi(p_1,...,p_K) = Phi(1\text{-}p_1 - ... - p_{k-1},p_{k+1},...,p_K) \quad (5)$$

one obtains a function which acts on an entire hyperoctant of (2*M*-1)-dimensional space. (The zeroth coordinate has been singled out for definiteness; obviously any coordinate could have been chosen.)

The results to be described here regard the Jacobian derivative of the functions *Phi* and *Psi*, as given by the following formulas:

$$Phi'_{ij} = A_j (p_k p_l A_k A_l (P_{kji} + P_{jki} - P_{kli})) / (p_r A_r)3$$

$$Psi'_{ij} = Phi'_{ij} - Phi'_{i0}$$

(6)

Convergence Result for the IMGA

To establish the convergence of the IMGA from a sufficiently good initial population, it suffices to compute the eigenvalues of the matrix *Phi'*[e_m], where e_m is apopulation vector corresponding to an optimal population. These eigenvalues are all less than one in magnitude, thus yielding the desired result.

Theorem. Assume the IMGA with zero mutation rate. Suppose that m is the unique global maximum of f over {0,...,N}. Let e_m denote the i'th unit vector of length N+1 (i.e., the vector indexed from 0 to N with a 1 in the i'th position and 0's in all other positions).

Then the eigenvalues of *Phi'*[e_m] are given by the formula

$lambda_0 = 0$

$lambda_i = (A_j/A_m) (P_{mji} + P_{jmi})$, i=1,...,N (7)

As a consequence, the vector e_m is an attracting fixed point for the iteration *Phi*. The asymptotic convergence rate is bounded by the magnitude of the largest eigenvalue.

The magnitude of the largest eigenvalue gives a bound on asymptotic convergence rate; and the form of this bound has avery clear intuitive meaning. One has $(A_j/A_m)d(i,m)$ where $d(i,m) = P_{mii} + P_{imi}$ is a measure of the proximity between i and m (d(i,m) is largest when i=m, smallest when i and m are Boolean complements, and grades fairly smoothly from the one extreme to the other). Thus, convergence is slowest when there is a genotype with fitness close to the fitness of m, which is also close to m. A fit genotype which is not so close to m matters less, as does a nonfit genotype which is close to m.

This result substantiates the observation of Davis and Principe that, as population sizes get larger and larger, the GA gets closer to demonstrating typical convergence behavior. More philosophically, it illustrates a fundamental difference between discrete and continuous dynamical systems. The discrete system here, the GA, illustrates a fundamental **diversity** that is absent from its continuous approximant, the IMGA. By averaging out the GA, the finite-size variations of particular populations are lost, and only the overall trend toward the optimum remains. This is good from an optimization perspective, but it may not be good from all perspectives. In real-world systems, there may be some value in the obstinate **refusal** to converge displayed by finite-size evolutionary dynamics. Convergence is not necessarily always a more important value than diversity.

On the Rate of Learning

At present no method for computing the eigenvalues of *Phi'* or *Psi'* away from the unit vectors e_i has been discovered. However, if one assumes crossover at a single randomly chosen cut-

point, then it is possible to compute the **determinant** of *Psi'* for the general case (the determinant of *Phi'* is of course zero). This determinant is given by

Det *Psi'* = c_M (A_i) / (p_i A_i) *K* (8)

where

c_M = r*V* / (M+1)*K-M-1* , V=(r-1)2*M-r* , M>1

c_1 = 1

This is a remarkably elegant and intuitively meaningful formula; unfortunately, the only known proof is quite lengthy and complicated.

The assumption of zero mutation rate is convenient but not crucial to this formula. To incorporate a mutation rate it suffices to consider the GA with mutation as a two-stage process: one stage of true crossover and another stage of mutation. The mutation stage may be represented as a constant linear operator on

probability vectors, namely

$M(p)_j = P_{1j}p_1 + ... + P_{nj}p_n$ (9)

where P_{ij} denotes the probability of i mutating into j.

The dimension of this matrix may be reduced just as with *Phi*.

According to the chain rule, the Jacobian of the two-stage GA is then given by the product of the determinant of this mutation matrix with the determinant of the crossover Jacobian. Thus mutation merely serves to add an extra constant outside the formula of the theorem. For most mutation schemes, the mutation matrix is only singular for finitely many values of the mutation rate, so that the constant will rarely be zero (numerical calculations verify that this is true for independent point mutation with M<6).

As noted above, this determinant formula has interesting philosophical implications. The numerator simply scales the determinant by a power of the geometric mean of the fitness function. And the denominator decreases the determinant whenever the fitness increases; and vice versa. The quantity |Det *Phi'*| measures the proportion by which *Psi* stretches areas in probability vector space; thus it is a rough measure of the **scope of the search** carried out by the genetic algorithm. The theorem suggests that genetic algorithms accomplish their remarkably effective optimization by the simple strategy of **adaptively narrowing** the scope of their crossover-based statistical search, based mainly on the average fitness of the population, and not on the specific makeup of the population. This is a new and powerful idea.

The fact that crossover gives a fitness-dependent Jacobian rate, while mutation gives a **constant** Jacobian, is also highly significant. What it says is that mutation is an inherently

inferior form of adaptation. Mutation does not close in closer and closer on the answer, it just keeps moving toward the answer at the same rate. This less sophisticated methodology is reflected in practice in the poorer performance of mutation-based GA systems. Crossover is needed to give the GA its initial burst of creativity, which brings it into the neighborhood of efficient problem solutions.

Discussion

The concept of an "infinite population size GA" may seem somewhat vexatious. After all, an infinite population is guaranteed to contain infinitely many copies of the optimal genotype! However, as the numerical results of Davis and Principe show, the behavior of the simple GAapproaches the infinite population size behavior rather quickly as population size increases. Thus, while the IMGA is indeed an infinite-population-size approximation to the GA, it may just as correctly be thought of as a deterministic nonlinear iteration which **intuitively** reflects the behavior of the GA.

Essentially, in biological terms, the IMGA is a "no genetic drift" approximation. The convergence result given here shows that, in the absence of mutation and random genetic drift, evolution **would** indeed converge. At least in the case of the simple GA, it is not only mutation but also genetic drift that plays a central role in preserving the diversity of evolving populations.

On the other hand, the determinant theorem shows how convergence proceeds. Fast at first, when fitness is low. Slow later on, when fitness is high. This is what the IMGA formulas say; it is what practical GA experience shows. And it is also, tantalizingly enough, what is seen in **human and animal learning**. The famous "power law" of learning shows that, when approached with a new situation, we learn very quickly at first, and our learning then slows down once we reach a certain level. This is exactly the pattern shown by the GA with crossover -- but **not** the GA with mutation. The GA with mutation shows a very un-biological **linear** learning curve. This observation does not prove anything conclusively, but it is extremely suggestive. It says that, if we believe learning is evolutionary, then one should believe learning involves **crossover**, rather than just mutation.

6.3 THE SIMPLE EVOLVING ECOLOGY (SEE) MODEL

The simple genetic algorithm is a beautiful and, as the name suggests, simple model. It captures the essence of crossover-based natural selection in a way that is highly amenable to computer simulation and is, to a lesser extent, amenable to mathematical analysis as well. However, as pointed out in the chapter introduction, it is an incomplete model in that it leaves out ecology. This omission will become glaringly apparent in later chapters as we attempt to use the genetic algorithm as a metaphor for more complex psychological processes. In a psychological context, a metaphor for evolution which lacks ecology will come up short far too often.

It is therefore interesting to seek a minimal formal model of **evolution plus ecology** -- a model which does for evolving ecology what the simple GA does for evolution. The goal of this section is to describe such a model, called the Simple Evolving Ecology model. Computer

implementation of the model is currently underway, in collaboration with Dr. Matthew Ikle'. Here we willdescribe the basic structure of the model and mention the qualitative nature of some preliminary numerical results.

In essence, SEE is a genetic algorithm living on a graph. Each node of the graph contains a certain population, which evolves within the graph according to the GA. Unfit individuals at a node of the graph can move to neighboring nodes in search of more favorable conditions. The fitness of an individual is determined, not only by an "objective" fitness function native to each node, but also by how the individual "fits in" with its neighbors. This combination of spatial structure, GA evolution and ecology gives the model the possibility to give rise to complex emergent structures quite different in character from the emergent structures observable in a simple dynamical system like the GA.

I will argue that SEE, unlike the GA, has the potential to serve as a minimal computational model of mind. Evolution alone is not enough to give rise to the full array of emergent psychological structures. Evolution plus ecology is.

Specification of the Model

I will now give a formal specification of the SEE model.

Geometry . Consider a weighted graph G, called a "world-graph." As a first approximation G may be taken to be a two-dimensional lattice. Each node of the world-graph G is called a "cell." The set of all cells, the node set of G, is called W. Each cell C has a neighborhood N(C), which is a certain connected subgraph of W, containing C. It may be useful to distinguish different neighborhoods for different purposes, e.g. $N_1(C)$, $N_2(C)$....

Each cell C contains a certain population P[C] of entities called "agents." These agents are drawn from some space S, e.g. the space $\{0,1\}^n$ of binary sequences of length n. One may also speak about the population of a collection of cells, e.g. the population P[N(C)] of a neighborhood of C. Where time-dependence needs to be emphasized, we may speak of P[C;t] to indicate the population of cell C at time t. Also, $N = N_C = N_{C;t}$ will denote the number of elements in P[C;t]; i.e., the population size.

Fitness, Environmental and Ecological. This gives the basic geometry of the model. Next the evolutionary aspect of the model must be specified. For this purpose, each cell C is endowed with two real-valued functions: an **environmental** fitness function, and an **ecological** fitness function. These two fitness functions embody the opposing principles of evolution and ecology, or adaptation and autopoiesis.

The environmental fitness function $f_{env;C}: S^N \to R$ is defined separately for each cell, and represents the physical surroundings in that cell. If agent x lives in cell C, $f_C(x)$ is called its "environmental fitness."

The function $f_{ec;C}$: P[N_{ec}(C)] x S -> R, called the "ecological fitness," is defined so that the quantity $f_{ec;C}$(x) determines the degree of "fit" between x and the other entities in N_{ec}(C). The neighborhood N_{ec} may be taken equal to C, or it may be larger.

item The "total fitness" of an agent x living at cell C is then the weighted sum of its environmental and ecological fitnesses,i.e. f(x) = c $f_{env;C}$(x) + (1-c)$f_{ec;C}$(x). The weight c is a global system parameter.

Interaction. The fitness functions specify how agents in the model are to be judged. It remains to specify how they are to **interact**. This is intentionally left quite general. An "action operator" is defined as a function r: S*m* -> S for some integer m. The system as a whole is endowed with a collection of action operators, R = {r_i,i=1,...,J}. The r_i may be reproductive operators, as in mutation, with m=1, or crossover, with m=2. Or they may be operators which allow the agents to rewrite each other, as in classifier systems. Each operator has a certain real-number "importance" I[r_i], which determines how often it will be utilized.

The dynamics within a single cell, at a given time, are as follows. First, a stage of selection according to fitness. Then, a stage of mutual interaction, leading to the creation of new agents within the cell.

In the selection stage, the population of the cell at time t, P[C;t], is replaced by an interim population Q consisting of elements of P[C;t] with different frequencies. item In Q, the frequency of an agent x is approximately proportional to its total fitness.

In the interaction stage, the elements of Q are chosen at random to interact with each other. The process is as follows. First an action operator is chosen. The probability of selecting operator r_i is proportional to its importance I[r_i]. Then, where m is the arity of r_i, m agents are chosen at random from Q. The product of r_i, applied to these m agents, is placed in the new population P[C;t+1]. This process is repeated N times, at which point the new population P[C;t+1] is full.

In the most general case, each edge (C_i,C_j) of the world-graph G may be understood to have a weight with two components, one component (w_{ij}) determining the base migration rate from C_i to C_j, the other (w_{ji}) determining the base migration rate from C_j to C_i. The sume of the w_{ij}, for fixed i, cannot exceed 1.

Migration. Finally, each agent x has a certain "propensity to migrate," *mu*(x), which is a monotone function of the rate of change of its fitness over the last g generations. If the total fitness of x has increased over this time period, then *mu*(x) is nonnegative. If the total fitness of x has decreased over this period, then *mu*(x) is nonpositive. E.g., in the simplest instance, one might set *mu*(x)=1 if the fitness of x has gone up since the last generation, *mu*(x)=-1 if the fitness of x has just gone down, and *mu*(x)=0 otherwise. Or, *mu*(x)=0 uniformly is also permissible. Newly created agents are assumed to have *mu*(x)=0.

The sum of *mu*(x) over all elements in a cell C_i is called *Eta*$_i$. The total migration rate from C_i to C_j is then given by *Mu*$_{ij}$ = w_{ij} + *a Eta*$_i$, where *a* is a constant. The constant *a* must be chosen so that the sum of *Mu*$_{ij}$ for fixed i does not exceed 1.

The actual migration from C_i to C_j is carried out as follows. An agent is chosen from $P[C_i]$, where the probability of x being selected is increasing with respect to *mu*(x). This agent is then removed from $P[C_i]$ and placed in $P[C_j]$. This process is repeated until $N[C_i]$ Mu_{ij} agents have been moved.

The weights w_{ij} may change over time. This represents a kind of "ecological learning" -- a selection of the optimalenvironment. For instance, one might decrease or increase w_{ij} based on whether the agents that pass through it tend to decrease or increase their fitness.

Initially, the weights may be set at random; or they may be set hierarchically, in such a way as to decompose G into regions within regions within regions. For instance, in a 32 x 32 lattice, one might have four 16 x 16 lattices separated by weights of .1, each consisting of four 8 x 8 lattices separated by weights of .2, each consisting of four 4 x 4 lattices separated by weights of .3.

Emergent Structures

At the time this book is being written, computational experimentation with SEE is just beginning. A major problem has been the design of an appropriate user interface for observing the dynamics of such a complex system. Long tables of numbers are not particularly intuitive. To study the dynamics of the genetic algorithm, one generally looks at episodic "dumps" of the whole population; but when there are a number of cells each containing their own population, these dumps become prohibitively complex. The Tcl/Tk scripting language has been used to build an interactive interface, in which the world-graph G is represented graphically on the screen, and the user can track one or more selected cells at a given time by clicking on their icons.

Up to this point, experimentation with SEE has taken place primarily with the ecological fitness function set to zero. Also, we have assumed agents modeled by bit strings, reproducing by one-cut-point crossover with mutation as in the simple GA. In this case, what one has is a spatially distributed GA, with the possibility for migration. In this circumstance, SEE leads to fixed point attractors, with the population distribution determined by the fitness functions at the different cells. In many instances one notes extreme sensitivity to initial conditions, whereby a very slight change in the initial population can cause much of the population to migrate into one cell rather than another. In general, there seem to be a variety of different network-wide attractors, spaced fairly close together in the state space.

In order to get more interesting dynamics out of SEE, one has to introduce ecological fitness. One can still get fixed-point dynamics here -- but one can also get total, unbridled chaos. As usual in such cases, the interesting behaviors are found by setting the parameters **between** those values that lead to boring behavior (fixed point) and those that lead to excessively chaotic behavior. The key goal, initially, is to get interesting large-scale emergent structures to arise in the network. Having gotten these structures to arise, one can then ask how often they will arise -- how large their basins are in the overall state space of the system.

In the preliminary experiments done to date, with the simple GA-like bit string model, the most striking result is the emergence of **continuous** maps across the network, whereby the

agents in each cell tend to be similar to the agents in neighboring cells (in terms of the Hamming metric). The reasonfor the emergence of this associative structure is obvious. If two neighboring cells have very different populations, then whenever an individual migrates from one to the other, it will be at severe risk of dying. It will bring down the fitness of the population of the new cell. On the other hand, if two neighboring cells have similar populations, then migration will be more likely to yield individuals who are successful in their new homes. Co-evolution of groups of neighboring cells means that the fitness of **both** cells will be higher if the populations "cooperate" by breeding similar populations.

It was hoped to also find an hierarchical attractor structure, whereby distinctive, large-basined attractors emerge on an hierarchy of levels (e.g., in a lattice world-graph, 2 x 2 squares, 4 x 4 squares, etc.). This has not yet been seen to emerge on its own. If it is elicited by means of an hierarchical migration structure, however, such an attractor structure **does** emerge and **will** preserve itself. The regions delineated by the migration coefficients will become **dynamically** separate as well as **geographically** separate.

These results need to be validated by further experimentation. Different kinds of emergent structures will also doubtless be observed. Looking further toward the future, one can see a wide range of possibilities for the SEE model.

For example, it is quite possible to envision SEE as a learning system. Certain cells may be distinguished as "input" cells, others as "output" cells. The other cells are "processing" cells. In a machine learning context, the goal of the system is to map inputs into appropriate outputs. I.e., at time t the fitness of elements in the output cells are judged based on whether they are an appropriate response to the input cells at time t-s, for some fixed delay s. This may be viewed as a special choice of the function $f_{env;C}$ for the output cells C. The effects of high or low fitness in the output cells will propagate through the system causing reorganization.

Finally, in an artificial life context, the whole graph G might be considered as the brain of a single organism. The behavior of the organism is then evaluated in terms of the dynamics on the meta-graph in which it lives. The meta-graph might also be a Simple Evolving Ecology as outlined here.

Doubtless the SEE model, like the simple GA, has its shortcomings and will need to be modified or replaced. In its integration of evolution and ecology, however, it represents a large step forward. The dynamics of very complex systems may be viewed a kind of harmonious struggle between forces of change and expansion and forces of preservation. SEE incorporates both of these "forces" in a simple and intuitive way.

The SEE Model as a Psynet Implementation

One of the main motivations for the development of the SEE model is the hope that SEE will turn out to be a good way of implementing the psynet model of mind.

From an AI perspective, the SEE model is vaguely similar to the Darwin Machine model discussed in Chapter One, the difference being that, instead of neural modules, one has

evolvingpopulations of abstract entities. A single-cell evolving population in the SEE model may be taken as an analogue of a neuronal group. Instead of feeding charge to each other, the populations exchange individuals with each other.

Indeed, if one wishes to map the SEE model onto the brain, it is not difficult to do so. The genotypes in the populations at each cell of the SEE lattice world could be taken as small neural modules, or "first-order networks." A single-cell population then becomes a second-order network, defined as a **competitive pool** of first-order networks. The first-order networks are all competing, first to fit in well with one another, and second to solve some problem represented by the cell's objective function. The whole SEE network is then a network of competitive pools, interacting with each other.

In the end, this somewhat peculiar "SEE neural network architecture" is really no less realistic than feedforward networks, Hopfield networks, recurrent networks, or any of the other standard models. Even so, however, it seems clear that not much is **gained** by talking about the SEE model in the language of formal neurons. It is enough to observe that, like the standard formal neural network models, the SEE model is a complex cognitive system with intuitive similarities to the brain.

In preliminary experiments, primitive versions of the dual network have already been observed to come out of the SEE model. The associative memory network comes out of the above-noted propensity of neighboring cells to have similar populations. This makes the whole network a kind of "self-organizing feature map," similar in some ways to Kohonen-style neural networks.

Hierarchical structure is a little subtler. But it is very instructive to think about the ways in which a perceptual-motor hierarchy might be gotten to emerge from SEE. There are no explicit "governor processes," which will explicitly rule over regions of the SEE world. So if there is to be a hierarchy of first-level processes, second-order processes controlling first-order processes, third-order processes controlling second-order processes, and so on -- it will have to come out **implicitly**, in the dynamics of the world population. The higher order "controling processes" will in fact have to be **attractors** of regions of the network. For simplicity, consider the case of a lattice of side 2*n*. In this case, one might find the attractor of a 2x2 square to be the "controlling process" for that square -- not a particular process that lives anywhere, but rather an **emergent** process, part of the coe-evolution of the 4 individual cell population. Similarly, the attractor of a 4x4 square might be the "controlling process" for that square, regulating the 2x2 square controlling processes within it.

Experimentation with SEE has not proceeded to the point of sophisticated study of multilevel attractors. However, it has been observed that the most interesting dynamics occur when the migration rates are defined in an "hierarchical" way -- i.e., defined so that the network consists of clusters, within clusters, within clusters, etc. The precise nature of these "interesting" dynamics has not yet been studied -- but one suspects that it may consist precisely of hierarchical attractor structures, as suggested in the previous paragraph.

In sum, the SEE model presents a very concrete and workableframework for studying associative memory, hierarchical control, and the interactions between the two. It is a complex model, as compared to the simple GA or the toy iterations of dynamical systems theory. However, it is the minimal model that I have found which has any reasonable potential to demonstrate the main **structures of mind** (the CAM-Brain and Darwin Machine project being, in my view, a close second).

What is essential to both of these models is the combination of adaptive **spatial** structure with local, self-organizing learning. This is left out in current complex systems models, which have either localized learning (neural networks, GA's), or spatial structure regulating rigidly defined dynamics (CA's, lattice maps in dynamical systems). Mind requires intelligent adaptation in **both** space and time. Flexible complex systems models which provide this are, it seems, excellent candidates for intuitive, computational models of mind.

6.4 THE SEARCH FOR STRANGE ATTRACTORS

Now let us return to the ordinary genetic algorithm. In this section and those that follow, instead of generalizing the genetic algorithm into a more psychologically realistic model, we will look at the possibility of using the genetic algorithm in ways that are, very loosely speaking, more psychologically interesting. We will use the genetic algorithm, not as a tool for problem-solving, but as a method for generating a **diverse population of interesting forms**. First of all, we will look at the use of the GA for generating interesting populations of strange attractors.

In particular, the strange attractors in question are Julia sets of plane quadratic maps. A Julia set represents the region of parameter space for which a plane iteration converges to an attractor rather than diverging to infinity. In the original sense of the phrase, only analytic mappings can have Julia sets (Devaney, 1988), but in practice the concept may be applied to general vector-valued iterations. These "generalized Julia sets" need not have all the elegant mathematical properties of conventional Julia sets, but they exist nonetheless.

Here I will not be directly concerned with generating Julia sets, but rather with generating the **attractors** corresponding to the parameter vectors within the generalized Julia set of an iteration. While these attractors do not display the same fractal intricacy as Julia sets, they do present a striking variety of visual forms. Furthermore, they are fairly quick to compute -- unlike Julia sets, which are notorious for their gluttonous consumption of computer time.

In a recent article, Sprott (1993) has presented a method for the automatic generation of strange attractors. Essentially, his technique is a Monte Carlo search based on the Liapunov exponent as a practical criterion for chaos. But this technique, while novel, is computationally very crude. The question addressed here is the **acceleration** of Sprott's Monte Carlo technique: Is there any way to adaptively guide the search process, causing it to "zoom in" more rapidly on the chaotic domain?

My method for accelerating the attractor generation process is to replace Sprott's Monte Carlo method with a more sophisticated optimization scheme, the genetic algorithm (GA) (Goldberg, 1988). As it turns out, this exercise is not only successful with respect to its original goal --

depending on the specific implementation, the GA can locate strange attractors around 5 to 50 times more efficiently than the Monte Carlo method -- it is also instructive in regard to the dynamics of the evolutionary optimization. It gives rise to a new variation of the GA: the **eugenic genetic algorithm**, in which "unacceptable" elements are weeded out before they ever get a chance to enter the population. And it provides a simple, interesting illustration of two important evolutionary phenomena: 1) genetic drift (Kimura, 1984), and 2) second-order evolution, or the evolution of evolvability (as will be discussed in Chapter 6).

On some graphics systems, e.g. an X-Terminal, the time required for point plotting greatly exceeds the time required for search, so that little is to be gained from improving Sprott's search technique. But on other systems, e.g. on a slower PC clone, the search time dominates. This point is vividly illustrated by programming a screen saver which presents an endless succession of different strange attractors, not by summoning them from memory, but by conducting a continual search through the Julia set of an iteration. When one runs such a screen saver on a 25 Mhz 386-sx, the time required to locate a new attractor becomes quite apparent: one wishes there were some way to produce the new picture a little faster!

In fact, this trade-off between search time and plotting time is not only machine-dependent, but also equation-dependent. For the plane quadratic map considered here, plotting time dominates search time on an X-terminal, but for more complex iterations this would not necessarily be the case. In fact, it is not hard to see that, for every machine, there is **some** class of potentially chaotic iterations for which search time will greatly exceed plotting time.

Sprott's standard "test equation" is the plane quadratic, as described in Chapter 1. His Monte Carlo algorithm selects random values for the parameters a_i and, for each parameter value selected, it computes a certain number of iterates, continually updating its estimate of the Liapunov exponent. Most parameter values lead to unstable orbits (in practice, $|x_n| + |y_n| > 100000$), often within the first few iterates, and these can be rapidly discarded. Others lead to a very small Liapunov exponent (in practice, <.005), which indicates that a fixed point or a limit cycle is being approached, rather than a strange attractor. Roughly speaking, around 85% of parameter values result in instability, while of the remainder all but around 1-2% lead to limit cycles. The chaotic domain, while clearly of positive measure, is relatively scanty.

There is no mathematical or artistic reason for choosing the quadratic instead of some other iteration; however, for sake of comparison with Sprott's results, I will use this same iteration as my "test equation" here. In particular, I will intentionally use the same **parameter settings** as Sprott does, for the various internal parameters of the generation process. This is important because a minor adjustment in these internal parameters cansometimes lead to a drastic change in the effectiveness of the search.

For instance, Sprott chooses his guesses for each a_i from 25 values equally spaced over the interval (-1.2,1.2). If the interval (12,12) were used instead (which, admittedly, contradicts common sense), the advantage of the GA approach would be immensely increased, because the Monte Carlo method would lose 90% of its successes, whereas the GA would be hampered only initially, and would eventually hone on the correct region. Also, Sprott's estimation of the Liapunov exponent involves, at each step, doing one iteration from a point separated from (x_n,y_n)

by 10-*6*, and then computing the difference between the result of this iteration and (x_{n+1},y_{n+1}). In some cases the specific value (10-*6*) of this "standardized separation" makes a significant difference in the computed value of the Liapunov exponent; adjusting this value can therefore affect the proportion of attractors to be labeled "chaotic" rather than "limit cycle." Finally, Sprott uses x = y = .5 as an universal initial point; for obvious practical reasons he makes no attempt to distinguish a chaotic attractor whose basin of attraction is the whole plane from a similar chaotic pattern which is obtained only from certain initial values (in the worst case, perhaps only from x = y = .5).

The most natural way to evolve **parameter vectors** for quadratic iterations is to use the real vector GA . In the real vector representation which I will use here, crossover of two "parent" vectors v and w of length n is naturally interpreted to mean **swapping of entries**: one picks a random cut-point k between 1 and n-1 inclusive, and forms a "child" vector by prepending the first k entries of v to the last n-k entries of w. And mutation is naturally taken as the operation of replacing each **entry** v_i of a vector with some value chosen from a normal distribution about v_i (the variance of this distribution may be fixed or, most usefully, it may be taken to gradually decrease over the course of evolution).

Strange attractor generation is an atypical application of the genetic algorithm, because one is not seeking to find the single parameter vector that **maximizes or minimizes a certain function**, but rather to produce a large collection of parameter vectors that **satisfy a certain broad criterion**. In ordinary GA algorithm applications, one wants the population to converge to a certain "answer," or perhaps to a small collection of possible answers. Here, convergence to a single answer can be produced under certain conditions, but it is considered a failure. Thus, for instance, "fitness scaling," which is often used to force convergence to a single answer, would be counterproductive here. The objective is to produce a lot of different attractors, ranging widely across the chaotic domain, but without running down so many dead ends as the Monte Carlo method does.

Genetic Drift

How to use the GA for strange attractor generation? One approach would be to use the Liapunov exponent as a fitness function. Other numerical parameters, such as the correlationdimension, also suggest themselves. As one might expect, however, this approach is not effective; it leads to a population dominated by a single parameter vector with a slightly higher Liapunov exponent (correlation dimension, etc.) than the others.

A better approach is to define each twelve-dimensional parameter vector $a = (a_1,...,a_{12})$ as either **acceptable** or **unacceptable**, and contrive a fitness function as follows:

f(a) = 0, if a is unacceptable

(10)

f(a) = 1, if a is acceptable

Under this arrangement the algorithm has no reason to prefer one chaotic attractor to another; the population should, one would suspect, evolve to contain a fairly random assortment of acceptable parameter vectors.

Unfortunately, things are not quite so simple! The culprit is a peculiar twist on the biological phenomenon of **genetic drift** (Kimura, 1984). A mathematical understanding of genetic drift in GA's is difficult to come by, but on an intuitive level the phenomenon is not so difficult to appreciate. The crucial idea is that of a **schema** of vectors: a collection U of vectors with the property that, once a certain proportion of members of U have invaded the population, the continued presence of members of U in the population is, if not guaranteed, at least quite likely. In the bit string GA (Goldberg, 1988), schema take the form of **hyperplanes**; in the real vector GA with mutation there is no analogous elegant formulation, but the phenomena associated with schema in bit string GA's are still observable.

For, consider: if one begins with a random initial population, nearly the entire population will be unacceptable. If something acceptable emerges by chance, the odds are that it will form unacceptable children, and its lineage will die out quickly. But eventually at least one vector will begin to flourish, i.e., it and its minor variations will take over a significant percentage of the population. The source of "genetic drift" in attractor generation is (empirically speaking) the fact that, once a single vector begins to flourish, it becomes virtually unassailable.

Call this first successful vector Vector 1. Vector 1 and its mutants form a schema: when a vector crosses over with another very similar vector, the triangle inequality dictates that any "child" vector obtained will be very similar to its two parents. And in a population where almost everything else has fitness 0, any schema that arises will expand very fast: virtually ever pair selected for reproduction will be drawn from the system. Suppose another acceptable vector, Vector 2, appears; its fitness will be 1 just like the members of the Vector 1 system. But when Vector 2 is selected to cross over, its mate will most likely be a member of the Vector 1 system, so that its children are fairly likely to be unacceptable and at best its lineage will expand slowly.

To indicate the severity of this phenomenon, consider a few numerical results. In one run of tests, the population size was 2000 (fairly large), and the standard deviation of the mutation operator was .01. Over the course of 50 different experiments,it took an average of 320 generations for 90% of the population to converge to a single answer, within two decimal places of accuracy. Over all 50 experiments, this answer was never the same, making quite clear that the cause of the phenomenon lies in the general **dynamics** of the GA.

One way to get around this difficulty is to introduce **speciation** -- only let each vector mate with its neighbors. But empirically, it seems that this only results in genetic drift to several different vectors at once. The final population contains 2, or 4, or 10 different classes of approximately-equal vectors; there is no sustained generation of diversity. A more radical solution is required.

The Eugenic Genetic Algorithm

How can this pesky genetic drift be circumvented? The key, it turns out, is that one must never let the population get into a condition where genetic drift can dominate. So, first of all, instead of using a random initial population, it is far better to **stock the initial population with acceptable vectors**. And after this, whenever unacceptable vectors occur, one must **throw them out immediately** -- without even giving them a place in the new population. This latter step may seem like overkill, since an unacceptable vector would never be chosen to reproduce anyway. But in practice, if one permits unacceptable vectors to enter the population, they will soon come to dominate, and genetic drift will kick in.

These population-management policies are reminiscent of the draconic social policy called "eugenics," and therefore I have christened the GA thus modified the **eugenic genetic algorithm**. Although eugenics may be ethically unacceptable in a human context, in the context of strange attractor generation it would appear to be just the ticket! Having booted out unacceptables once and for all, one can cross and mutate to one's heart's content, and no vector will ever have the opportunity to dominate. Milder forms of genetic drift may still sometimes be observed: now and then there will arise a population consisting solely of, say, cigar-shaped attractors pointing in one direction or the other. But this is the exception, not the rule.

Practical experience with the ordinary GA teaches that mutation is of relatively little importance; and the same principle holds for the eugenic GA. Even without any mutation at all, the eugenic GA is a powerful tool for attractor generation. First one uses Sprott's Monte Carlo technique to generate, say, 10 - 50 strange attractors, saving the parameter vector underlying each one. Then one takes this collection of parameter vectors as an initial population, and repeatedly executes the following eugenic genetic algorithm:

1) select two different vectors at random (from a uniform distribution)

2) select a random cut-point (between 0 and 10, since the vectors have 12 entries), and produce a child vector

3) if the child vector is acceptable, choose a population

member at random and replace it with the child vector

The results are surprisingly good. For instance, over 10 different 300-generation experiments with population size 25, an average of about 30% acceptable children was obtained. 10 similar experiments with population size 10 yielded an average of 35%. The variances involved seem to be quite large; the highest proportion obtained in the course of these experiments was 60%, the lowest 11%. But these values are all much higher than the 1-2% obtained with the Monte Carlo method, so the intuitive moral would seem to be clear.

If one adds mutation, the success percentage unsurprisingly seems to go down somewhat, e.g. by 4-5% with a mutation standard deviation of .06. The advantage is that, in the limit anyway, one has the possibility to explore an infinite number of parameter vectors from a fixed initial population. But in practical terms, this advantage does not seem to amount to much.

These results do depend on the specific nature of the crossover operator. One might think to replace one-cut-point crossover with two-cut-point crossover, or uniform crossover in which each entry of the child vector is chosen at random from either of the two parent vectors. But this significantly degrades the results, in the latter case by approximately a factor of two. If one looks at the form of the quadratic iteration, however, this is not surprising; it probably stems from the fact that in some cases a_i is multiplied with a_{i+1} or a_{i+2} before it produces an effect (x_n,y_n). This interdependence makes it a benefit for pairs of nearby coefficients to stick together through the process of crossover.

A more curious feature of the results of these experiments is that they depend fairly sensitively on Step 3 of the eugenicized GA. This is a highly nonobvious result: why should the child vectors produce any better children than the parents? Nonetheless, if Step 3 is removed, one obtains much worse figures -- for instance, an average of around 7% acceptable offspring for population size 25, down from 15-16%. This suggests that some form of **second-order evolution** (as defined in the previous chapter, in the context of the evolution game) is occuring -- an evolution in the direction of forms with greater "evolvability." For some reason offspring obtained by crossover are better parents than typical acceptable vectors. The GA is arriving at a region of population space which is an "adaptive attractor": it is adaptive in the sense of being generally fit, and it is an attractor because the GA happens upon it as if by accident, without explicity coaxing.

As with genetic drift, the most sensible intuitive explanation involves subsets U of the generalized Julia set that are **schema** under crossover, in the sense that if one takes two vectors a and b within U, there is an relatively high probability that the child of a and b will also be within U. Over the course of its evolution, the GA would seem to be "converging" to these schema, in the same manner that repeated iteration of a function from an arbitrary initial value can lead to convergence to a fixed point of that function.

In the more general, system-theoretic language that will be introduced later, these "schema" subsets U are **autopoietic subsystem** of the GA dynamical system. They are self-producing systems of vectors or pictures, which produce each other over andover again, and in this way protect themselves against outside interference. What is happening with the GA here is a very simple example of a process that is crucial in psychological systems -- a process that, for example, underlies the maintenance of human belief systems.

The Question of Repetitiveness

The eugenic GA partially stifles genetic drift. But does it avoid it entirely? In other words, are the attractors generated by the eugenic GA somehow **more visually repetitive** than the ones generated by the Monte Carlo method? By re-using old coefficients, instead of generating entirely new ones, are we sacrificing some kind of originality?

It is clear from specific cases that the offspring of two attractors can look very different from its parents; thus the mechanism for generating diversity is present in the crossover operator. An example is given in Figure 4. Here, the two parent parameter vectors $(a_1,...,a_{12})$ are as follows:

Parent 1:

-1.10350, -0.00570, 0.23650, -0.82920, 0.72670, 0.35840, -0.75380, -0.45080, -0.37600, 0.53850, -0.99390, -0.80260

Parent 2:

1.00650, -0.03940, -1.17760, -0.49160, 0.24740, 0.10900, 0.30930, 0.16200, 0.73160, -0.00960, -0.73040, -0.48770

Both of these parameter vectors, as shown in the Figure, lead to fairly uninteresting attractors. The child is obtained from taking the first four entries of Parent 1, and prepending them to the last 8 entries of Parent 2:

Child:

-1.10350, -0.00570, 0.23650, -0.82920, 0.24740, 0.10900, 0.30930, 0.16200, 0.73160, -0.00960, -0.73040, -0.48770

Despite its derivation from the parameter vectors of its parents, the **shape** of the child bears little relation to the **shapes** of its parents. And this example is not unusual in this respect. What the prevalence of cases like this means is that the crossover operator does a fairly **broad** search of the Julia set of the quadratic operation. The Julia set is shaped in such a way that, when two points within it are crossed over, the result is reasonably likely to lie within the Julia set as well. And the "evolution of evolvability" is simply the tendency of a population to drift into regions of the Julia set which have a greater than average tendency to cross into the Julia set.

But the ability of crossover to produce dramatic new forms does not disprove the possibility that these new forms will tend to be repetitive. In an attempt to approach the issue of repetitiveness **quantitatively**, I have kept records of the distribution of the positive Liapunov exponents generated by theeugenic GA, as opposed to Sprott's Monte Carlo method. The results are not surprising. Sprott's method leads to positive Liapunov exponents which are distributed on a slowly decaying "hump" shaped curve, centered around .2 or .3. The eugenic GA, as tested on a population size of 20, leads to the same kind of distribution at first, but then as evolution progresses, the distribution grows a sharper and sharper peak around some particular point, or sometimes two or three particular points. During the initial period of more slowly-decaying positive Liapunov exponent distributions, the GA still generates chaotic attractors 5-20 times more often than the Monte Carlo method; but as the curve becomes peaked, the success rate can become even higher. What this represents is the indefatigability of **genetic drift**. Even the eugenic GA does not escape it completely.

Given the pervasiveness of genetic drift, if one wishes to search for strange attractors, the best course seems to be to alternate Monte Carlo search with low-mutation-rate eugenic GA search: e.g. to use Monte Carlo search to generate an initial population of 25 acceptable vectors, then use an eugenic GA to breed perhaps 100 attractors from these, then revert to Monte Carlo search again after a certain period to obtain a new initial population, etc. In order to determine when to

switch back to the Monte Carlo search, one can check the **entropy** of the distribution of the last, say, 50 positive Liapunov exponents generated. When this entropy becomes too low, one has reason to believe that genetic drift is predominating.

This algorithm is only a beginning; there is clearly much more "engineering" to be done. But even in its present form, the method provides a very rapid search of the interior of the Julia set. And, when applied to iterations less well-studied than the plane quadratic, it may provide a highly effective method of **discovering** strange attractors, where their existence is not previously known.

6.5 THE FRACTAL INVERSE PROBLEM

In this section, we will explore the possibility of evolving specified target fractals using the **iterated function system** algorithm for fractal generation. This exploration is similar in spirit to that of the previous setion, but quite different in detail. It provides further confirmation of the remarkable ability of the GA to generate diverse, intricate structures.

The theory of iterated function systems is a method for getting fractal sets from dynamical equations. It is both simple and elegant. Most of it is implicit in the half-century-old work of mathematicians like Hausdorff and Banach. Much less simple and much less classical, however, is the IFS **inverse problem**. Given, say, a picture or musical composition, which is in principle capable of being effectively produced by the IFS method, it is no easy task to actually **determine** the coefficients of the iterated function system which generates the picture or composition.

In this section I will describe some partially successful efforts to solve this difficult problem, using the genetic algorithm. As it turns out, a straightforward GA approach isworkable but inefficient. It is reasonably useful for simple one dimensional examples, but runs into difficulty with even the easiest two dimensional problems. The same computational obstacle was located by Mantica and Sloan (1988), using a completely different optimization method, so-called "chaotic optimization," leading one to the conclusion that the difficulty may lie not in the incapacity of the optimization methods but in the inherent complexity of the fractal inverse problem.

Barnsley's fractal image compression method involves the use of **local IFS's** instead of genuine iterated function systems. Local IFS's are mathematically less straightforward than IFS's; they are not contraction maps, so that convergence from an arbitrary starting point is not guaranteed. Also, unlike IFS's, they do not capture an image's global structure, but only the local regularities in its structure. Computational practicality is purchased, in this case, at the cost of philosophical and scientific interest.

In the following section, I will turn to a related problem which poses a different sort of difficulty. As in the application of the GA to strange attractor generation, instead of using a GA to solve an optimization problem, we will attempt to use a GA to evolve a **population** of entities satisfying a certain criterion. But this time, instead of evolving strange attractors, we will be trying to evolve one-dimensional IFS's whose attractors form **aesthetically satisfying melodies**. The eugenic genetic algorithm fits the bill nicely here; the trouble is in defining a fitness function

which quantifies the notion of "aesthetically satisfying." Drawing on some ideas from musical psychology, we will outline one way of approaching this difficult problem, and present some musical results.

Iterated Function Systems

An iterated function system, or IFS ,consists of N affine transformations w_i, each one associated with a certain probability p_i. In the one-dimensional case, the w_i take the particularly simple form

$$w_i(x) = a_i x + b_i \quad (11)$$

To specify a 1-D IFS with N transformations, then, requires 3N - 1 floating-point numbers (it is 3N-1 and not 3N because the N probabilities p_i are known to sum to one, which reduces by one the number of degrees of freedom).

In the two-dimensional case, the w_i take the form:

(12)

One may show that any such transformation can be expressed as the composition of one translation, three scalings, and three rotations. Two "classic" examples are given in Table 1.

The **deterministic IFS algorithm** consists of the following iteration:

$$A_{n+1} = w_1(A_n) \textit{ union } ... \textit{ union } w_N(A_n) \quad (13)$$

The "initial set" A_0 must be given. This is an iteration on **set space**, meaning that it takes sets into sets.

If one restricts attention to **compact**, **nonempty** subsets of some metric space (e.g. the line or the plane), then one can show that this algorithm necessarily converges, given only the assumption that **the determinants of the w_i are all less than one**. And even this can be strengthened; it suffices that the geometric mean of the determinants not exceed one. The limit set, the "fixed point attractor" of the iteration, is of course given by the equation

$$A = w_1(A) \textit{ union } ... \textit{ union } w_N(A) \quad (14)$$

At first this seems a remarkable result but it turns out to be quite "shallow," in the mathematical sense; it follows almost immediately from Banach's contraction mapping principle. The trickiest part, in fact, is the definition of an appropriate **metric** on set space (without some way of defining the distance between two sets, one cannot prove that the iterates converge to a limit). But this problem was solved half a century ago, by none other than Felix Hausdorff, of Hausdorff dimension fame (and also known, in point set topology, for his introduction of the "Hausdorff space"). The distance between a set A and a point y is given by d(A,y), the minimum over all points x in A of the pointwise distance d(x,y). The asymmetric distance between two sets

A and B, d(A,B), is given by the maximum over all points y in B of d(A,y). And, finally, the **Hausdorff distance** from A to B, h(A,B), is the maximum of the two asymmetric distances d(A,B) and d(B,A). It is in this sense that the deterministic IFS algorithm converges to its attractor: $h(A_n,A)$ tends to zero as n tends to infinity.

But the deterministic IFS algorithm is useful primarily for theoretical purposes. For practical computation one uses the **random iteration algorithm**. Given an IFS, one obtains an "attractor" by executing the following **decoding algorithm**:

1) Select an initial point x

2) Choose a transformation w_k (where the chance of choosing

w_i is proportional to p_i)

3) Replace x with $w_k(x)$, and plot the new x

4) Return to step 2, unless sufficiently many iterations

have been done already

In principle one should do infinitely many iterations; in practice one does perhaps 1000 - 50,000 iterations, and discards as "transients" the initial 100 - 2000 points.

The result of this 2-D algorithm is a density or **measure** m on a certain region A of the plane. Similarly, the 1-D algorithm gives rise to a measure on a certain subset of the line. The necessity to deal with measures means that, although the random iteration algorithm is very simple from the perspective of implementation, it cannot be fully understood without a formidable mathematical apparatus. The Hausdorff metric on a space of sets must be replaced by the Hutchinson metric on aspace of probability measures. However, the upshot is the same: Elton's "IFS ergodic theorem" says that the random iteration algorithm will eventually, with probability one, converge to a certain attractor or "invariant measure" determined by the w_i and the p_i.

The "support" of this invariant measure m is the same set A given in Equation (14) above. By construction, the measure assigns A an area of 1; m(A) = 1. The value of m(B), for a given subset B in A, is the "percentage" of points in A which are also in B, on a typical run of the decoding algorithm from an initial point within A (or, more precisely, it is the limit of this percentage as the number of iterations tends to infinity). So the w_i determine the region of support of the measure m, and the probabilities p_i determine the precise contour of the measure.

In graphics terms, this means that the p_i serve only to determine the "grey scale values"; whereas the w_i determine which points get some shade of grey (those points inside A), and which points get pure white (those points outside of A). In practice, however, if some of the p_i are set sufficiently low, then certain regions of A may not be likely to show up on the computer screen within a reasonable number of iterations; a fact which cannot be neglected when one is evolving IFS's by the genetic algorithm, as will be done in Chapter Five.

It is noteworthy that the IFS method deals only with **fixed point attractors**. One can define chaotic dynamics **on** the attractor set A -- by defining a map on A by $s(x) = w_{i(x)}\text{-}1(x)$, where i(x) is one of the integers j for which $w_j\text{-}1(x)$ lies in A -- but this is a different matter. There is no good reason why algorithms that demonstrate periodic and chaotic dynamics on **set space** and **measure space** should not be equally useful, or even more useful. But, for the present, such algorithms exist only in the realm of speculation; we are left with fixed points only.

Some Simple Experiments

Can one use the genetic algorithm to go from the fractal to the IFS that generates it? This section describes experiments aimed at answering this question, conducted on a 486 PC by Hiroo Miyamoto and Yoshimasa Awata at the University of Nevada, Las Vegas, under the direction of the author.

As a simple one-dimensional test case, we used a nonuniformly shaded Cantor set in [0,1] as a target picture. The IFS generating this target is given by the two affine transformations

$$w_1(x) = a_1x + b_1 = .333x$$

$$w_2(x) = a_2x + b_2 = .333x + .666 \quad (15)$$

with probabilities

$$p_1 = 0.3$$

$$p_2 = 0.7$$

To represent this as a parameter vector, we used the standardordering (a_1,b_1,p_1,a_2,b_2), where p_2 is omitted due to the relation $p_2 = 1 - p_1$, obtaining the vector (.333,0,0.3,.333,.666).

Note that this strategy for encoding probabilities only works for the case N = 2. In the more general case where there are N probabilities, one must encode at least N-1 numbers representing probabilities, but these numbers will not in general sum to one, and they must be normalized to sum to one before the vector is used to generate an image.

In our Cantor set example, preserving three digits of accuracy yields an 80-bit binary string. The task given to the genetic algorithm, then, is to search the space of 80-bit binary strings for the one which generates the shaded Cantor set.

The Fitness Function

Ideally, one would like to use a fitness function involving the Hutchinson metric on measure space (Barnsley, 1988). However, each evaluation of this metric involves a minimization on function space, and so this is not an effective computational strategy. The longer each function evaluation takes, the longer the algorithm will take to run. In practical applications, one is not directly concerned with measure-theoretic convergence, but rather with convergence to the

degree of approximation represented by a computer screen. Therefore, it seems sensible to use a lattice-based "distance measure," in which the distance between two finite sets A and B is determined by dividing the interval or square involved into n rectangular subdivisions. For each subdivision, one measures the magnitude of the difference between the number of points of A which lie in the subdivision, and the number of points of B which lie in the subdivision. Then one sums these values, to obtain the distance between A and B.

In the Cantor set example, the n subdivisions become n subintervals of [0,1], and one obtains a fitness function f defined so that the fitness of the i'th population element is given by

$$f(i) = 1 - (ob + dp) / (2 - T) \quad (16)$$

where:

T is the total number of iteration points for each IFS, which is taken equal to the number of pixels in the target image;

ob is the number of points generated by the i'th IFS which lie outside the interval [0,1];

$dp = dp(1) + ... dp(n)$,

$dp(i) = |ap(j) - tp(i,j)|$,

ap(j) is the number of points in the target image which lie in the j'th subinterval

tp(i,j) is the number of points in the attractor of the i'th IFS which lie in the j'th subinterval.

According to this scheme, the maximum fitness is 1, and the minimum fitness is 0. The probability of a given IFS in the population being selected is then given by the standard formula $f(i)/[f(1) + ... + f(n)]$.

Experiments

For our first run of experiments, we set the population size at 300, and permitted a maximum of 1500 generations. In the first run, the maximum fitness value of 0.802 was reached in generation 926. In the second run, 0.766 was reached after 850 generations, and there was no improvement after that. In the third run, a near perfect fitness of .972 was achieved after only 400 generations.

The successes were due, not to particularly fit initial populations, but solely to crossover effects. Mutation rate was kept very low (0.00015 per bit), so it is unlikely that mutations played a major role in the convergence. Over twenty runs, the maximum fitness achieved after 1500 generations averaged 0.858; and in all but three cases, this value was achieved before 1000 generations.

One might wonder what would happen if the number of transformations required to generate a given picture were overestimated. As an initial approach to this question, we set the genetic algorithm the problem of finding three affine transformations to generatethe Cantor set. The results were quite encouraging: over twenty runs, the maximum fitness achieved after 1500 generations averaged.721. In all but two cases, one of the three transformations ended up with coefficients very close to zero.

The behavior observed in these experiments is typical of the genetic algorithm: the desired answer is approached fairly quickly, and then never precisely obtained. However, the attractors with fitness over 0.85 are visually almost identical to the Cantor set, suggesting that the genetic algorithm may in some cases be able to provide IFS coefficients that are adequate for "lossy" solutions of the fractal inverse problem, i.e. for applications in which a decent approximation of the target image is enough.

The number of generations seemed to matter less than the populaton size. Continuing with the Cantor set example, for sake of variety we shrunk the population size to 50, and expanded the number of generations to 4000. In a typical run, after generation 84, the fitness value became 0.406 and the optimal parameter vector was (0.599, 0.169, 0.598, 0.377, 0.001). After this stage, the fitness value simply remained in the range between 0.335 and 0.415. The smaller population size kept the algorithm stuck in a far-from-optimal region of IFS space.

Though poor for the Cantor set, this result was typical of our experience with two-dimensional target pictures. For instance, a very simple two-dimensional fractal, the Sierpinski triangle, proved completely unapproximable. Only three affine transformations are needed, for a total of twenty real variables. But time after time, a population of 500 or 1000 ran for several thousand generations (96 hours of CPU time on our 486) without achieving fitness above 0.4. Clearly the longer bit strings required a much larger population size; but populations greater than 1000 are not feasible for PC implementation due to the extremely long running times involved.

Frustrated with these results, I decided to port the algorithm from the PC to a more powerful Unix system. Theresults, however, were unimpressive. A population of 10000 vectors, running for one hundred trials of a thousand iterations each, managed to break the .5 fitness barrier for the Sierpinski triangle only eight times. Only twice did the fitness exceed .8, producing a reasonable Sierpinski triangle. Further experiments were performed on fractals produced by 4 or 5 affine transformations, on random images, and very simple picture files not derived from IFS transformations; but these proved entirely unsuccessful.

Conclusion

Our one-dimensional experiments demonstrate the potential effectiveness of genetic algorithms for solving the IFS inverse problem. But our results for the two-dimensional case indicate that genetic optimization may not be an adequate tool. Further experimentation is required in order to determine the rapidity with which the population size must increase, in order to accomodate the longer bit strings required by higher dimensions and more transformations. However, it seems most likely that, in order to be useful for the fractal inverse problem, the

genetic algorithm must be hybridized with some other optimization algorithm. This is a natural avenue for future research.

6.6 EVOLVING FRACTAL MUSIC

Now let us turn from pictures to music. In order to use the GA to generate fractal music, we will stick with IFS's, but will return to the "target-less" form generation of Section 6.4. However, instead of merely trying to generate chaotic melodies, we will describe a programme aimed at the evolution of psychologically and aesthetically interesting melodies.

The key problem of algorithmic music composition is **structure**. Human-composed music combines mathematical and emotional structure in a complex way. But a computer composition algorithm, lacking access to the human unconscious, must draw structure from elsewhere. IFS music seeks to solve this problem by deriving musical structure from geometrical structure; more specifically, from the structure of the geometrical entities called **measures**. Goguen (1990) has used two-dimensional IFS's to generate fractal music (Goguen, 1990). However, it seems more natural to use one dimensional IFS's to represent melodies; so this is the approach I have taken here. Music is indeed multi-dimensional, but its core organization is one-dimensional, defined by the axis of time.

The fractal melodies generated by 1-D IFS's are not in any sense random; they possess structure on all different levels, from the most local to the most global. The worst of these fractal melodies are completely chaotic or excessively repetitive. The best, however, are at least moderately "catchy." Because they are generated without attention to the rules of Western tonal music, they tend to contain strange delays in timing and occasional "odd notes." Most of these aberrations could be filtered out by a "melody postprocessor" programmed with the elements of tonal music theory. Here, however, I have takenan alternate approach: I have chosen to play the melodies in a genre of Western music which is relatively insensitive to the niceties of music theory -- what is known as **"industrial music"**. As it turns out, fractal melodies sound quite natural when played as feedback-infused guitar samples against the background of a loud drum machine. This is not because industrial music is indifferent to melodic quality, but simply because industrial music is forgiving of strange delays in timing and occasional odd notes.

Fractal melodies are structured on many different scales. Admittedly, however, there are differences between this **fractal** structure and what we ordinarily think of as **musical** structure. The most important difference is the sense of **progressive development over time** -- most human music has this, but very little fractal music does. The most interesting fractal melodies are those which, coincidentally, **do** happen to display progressive development. In order to isolate these particular fractal melodies from the larger mass of uninteresting ones, I have implemented a **genetic algorithm** (GA), acting on the space of "code strings" for 1-D IFS's.

The genetic algorithm requires a "fitness function" to tell it which qualities make one melody superior to another. To supply such a fitness function is a very difficult task -- obviously, no one knows exactly what it is that makes a melody sound good. However, if one's goal is merely to

separate the passable from the terrible, and to do so statistically rather than infallibly, then some headway can be made. I have implemented a fitness function based on Meyer's (1956) theory of musical semantics and the mathematical psychology of *The Structure of Intelligence*, the basic idea of which is that good melodies demonstrate the **surprising fulfillment of expectations** (SFE). The SFE-based fitness function is certainly not a litmus test for melody quality, but it does serve as a rough "filter." In practical terms, it decreases the number of poor fractal melodies one must listen to in order to find one good one.

At the present stage of development, 1-D IFS's are primarily useful as a compositional tool rather than as a stand-alone composition method. Particularly when augmented by the GA, they are a reliable source of novel melodic ideas. It seems possible, however, that in the future the role for IFS's in music composition could become even greater. Combined with a tonality-based postprocessor and an improved SFE-based fitness function, IFS's might well be able to produce "finished compositions." The initial experiments described here provide strong encouragement for research in this direction.

From Mathematics to Melodies

How does one translate **measures** into **melodies**? The simplest strategy is to begin with a fixed interval I containing all or most of the attractor region A. One then divides the interval into M equal subintervals I_r, where M is the number of notes in the melody being constructed, and runs the IFS decoding algorithm keeping track of the number N_r of points which fall into each interval I_r. If I_r does not intersect the attractor A, then after perhaps a few stray transients, no points should fall inI_r at all; N_r should be zero or very close to zero. On the other hand, in theory N_r may become very high for certain intervals; it is even possible that every single point plotted will occur within the **same** interval. What is needed to produce music is a formula for converting each number N_r into a **note** or **rest**.

There are many ways to carry out this conversion. For an initial experiment, I have used the following scheme:

1) First, compute the minimum and maximum N_r over all the

subintervals I_r; call these valus N*(m)* and N*(M)*.

2) Divide the interval (N*(m)*,N*(M)*) into 12 equal-sized subintervals R_l, l = 1,...,12.

3) Assign the interval I_r the l'th note in the octave if N_r

lies inside R_l.

This is very simplistic and may be varied in several obvious ways. However, although variation of the conversion scheme does affect the character of the melodies obtained, it does **not** sensitively affect the presence of complex multilevel structure in these melodies. Under any reasonable conversion scheme, one finds melodies with repetitions, repetitions within repetitions, and so forth, and continual minor variations on these repetitions. The complexity of the melody

increases as one increases the length of the "code vector" which gives the coefficients of the IFS. Too few transformations, and generally repetitive melodies result. As a rule of thumb I have chosen the number of transformations to be half the number of intervals (notes); but this is largely an arbitrary decision.

1-D IFS's are a surprisingly interesting compositional tool. Fractal melodies obtained from 1-D IFS's in the manner described above sound far different from, and better than, random melodies. Some are overly repetitive, but many are complexly structured on multiple scales. Not all of this structure is humanly appealing -- much IFS music sounds markedly alien, neither unpleasant nore particularly "listenable." Occasionally, however, one happens upon a fractal melody which displays the sense of **patterns unfolding over time** that we expect from human-composed music.

Surprising Fulfillment of Expectations

My approach to assessing the fitness of a melody involves Meyer's (1956) theory of musical emotion, which states that the key to melodic structure is the artful combination of predictability and surprise. A successful melody, Meyer proposes, works by **surprising fulfillment of expectation** (hereafter abbreviated **SFE**). SFE means, quite simply, first **arousing an expectation**, then **fulfilling the expectation in a surprising way**. This approach to musical aesthetics has the merit of being easily programmable: one may recognize patterns in a melody, and compute the degree to which their fulfillment is "surprising" in terms of the other expectations implicit in the melody.

I will now describe in detail how I have implemented SFE to evaluate the "quality" of a melody. The basic method here isvery reminiscent of the Chaos Language Algorithm described above, although the details of the implementation are somewhat different. The idea, here as in the CLA, is to recognize repeated subsequences in a time series, and use these repeated subsequences to get at the **structure** of the series. One is getting at patterns in dynamics, and, now, using the genetic algorithm to evolve dynamics manifesting appropriate patterns.

Following the approach taken in *The Structure of Intelligence*, the **intensity** of a pattern P is defined as one minus the compression ratio which the recognition of P provides. So, suppose one is dealing with a pattern of the form "This note sequence of length k is repeated r times in a melody of length t." If one substitutes a single "marker" symbol for each occurrence of the sequence, then one has reduced the string of t notes to a string of t - r(k-1) symbols. On the other hand, it takes k symbols to store the string itself, thus giving a total of t - r(k-1) + k symbols for the compressed melody, as opposed to t symbols for the uncompressed melody. The intensity of the pattern is then given by

$$1 - [t - r(k-1) + k]/t = [r(k-1) - k]/t \quad (17)$$

For very long melodies this formula is not correct, since one will eventually run out of symbols (assuming as usual a finite alphabet), and will have to start inserting two or three symbols to denote a single repeated string. One must then become involved with the tedious details of Huffman coding or arithmetic coding (Bell, Cleary and Witten, 1990). For the short melodies

involved in the current experiment, however, equation (3) is a perfectly adequate approximation. In fact, as a practical matter, for short melodies it seems to work better to replace the "-k" term with a "-1". This means that once a pair of notes occurs **once** it is a pattern; it is a falsehood in terms of information theory but may possibly be more psychologically accurate, and is in any event useful for the evolution of short melodies. Thus one obtains the simpler formula

$$[r(k-1) - 1]/t \quad (17')$$

If one is dealing with a **contour** pattern or an **interval** pattern instead of a note pattern, formula (8) must be modified in a somewhat ad hoc fashion. For clearly, a contour sequence of length k, occuring r times, should be counted as less intense than an interval sequence of the same length which has occurred the same number of times; and the same should be true of interval sequences compared with note sequences. The question is really one of musical psychology: from the point of view of a human listener, how prominent is a repeated contour sequence as opposed to a repeated interval sequence or note sequence? Studies show that contour patterns are very important to human melody identification, but that interval patterns also play a significant role. The easiest way to model this balance is to introduce **weights** $0 < w_C < w_I < 1$. The intensity of repeated contour and interval sequences are then defined by the formulas

$$w_C\,[r(k-1) - k]/t \quad (18)$$

$$w_I\,[r(k-1) - k]/t \quad (19)$$

respectively, where for short melodies one may in practice wish to replace "-k" with "-1" as in (8).

The values of the weights are actually quite important. In most fractal melodies, contour patterns are the most common patterns, so if w_C is set below .5 most fractal melodies will have very few significant patterns. As a rule I have set $w_I = .8$ and $w_C = .6$. In fractal melodies there are very few interval patterns which are not also pitch patterns; this is clearly one significant difference between fractal music and Western tonal music. In this respect fractal music is more similar to, e.g., Balinese gamelan music, which is almost entirely based on contour with no concept of absolute pitch at all.

This way of measuring intensity is time-independent, in the sense that it counts patterns from the distant past (the beginning of the melody) just as much as more recent ones. Even for very short melodies, this is not psychologically realistic; there is always some degree of **damping**. This can be modeled by

computing the intensity over some interval **shorter** than the one from the beginning of the melody (time 0) to the current point in the melody (time t). Having measured the intensity of a

pattern P over several intervals [t-s_i,t], one then averages the intensities obtained in this way to obtain the total intensity of P.

Ideally one would wish to let s_i run from 1 through t. Writing the intensity of pattern P over [t-s_i,t] as IN[P;s_i], one could then define the damped intensity as the sum

v(1)IN[P;1] + ... + v(t)IN[P;t] (20)

where v(x) is a "decay function" monotone decreasing in x, satisfying v(1) + ... + v(t) = 1/t. In the current experiment I have implemented a simplified version of this scheme in which only two values of s_i are used, s_1 = t giving the whole melody and s_2 = 7 +/- 2, giving only a brief window of past history. Specifically, I have chosen s_2 = 5 note patterns, s_2 = 6 for interval patterns, and s_2 = 8 for contour patterns.

This idea of a "short scale" versus a "long scale" is obviously a simplification, motivated primarily by practical rather than conceptual considerations. But it does have some psychological relevance: recall the "magic number 7 +/- 2" of short term memory capacity. The short scale intensity of a pattern may be reasonably interpreted as the intensity of the pattern in the part of the melody "currently psychologically present," while the long scale intensity is the intensity of the pattern in the melody as a whole, as stored in long-term memory. Perhaps not coincidentally, 7 is also an approximate cut-off for the **size** of frequently observed repeated patterns in melodies. For note or interval patterns k>5 occurs only rarely; and for contour patterns k>7 is infrequent (it is well nigh impossible to find two human-composed melodies with the same contour in the first 15 notes).

So, given these two estimates of intensity -- short scale and long scale -- how does one estimate the degree of SFEdelivered by a given note in a melody? The key, I suggest, is to **compare** the amount of short scale intensity delivered to the amount of long scale intensity delivered. If patterns with long scale intensity are selected **at the expense** of patterns with short scale intensity, then this means that immediate goals are being frustrated in the service of long-term satisfaction. This is certainly not the **only** kind of SFE, but it is certainly one important species of SFE, which has the advantage of being easily quantified.

At each note t, each pattern is assigned a **long scale percentage**, obtained by dividing its intensity over [0,t] by the intensities of all other patterns over [0,t]; and a **short scale percentage**, obtained similarly from intensities computed over the interval [t-7,t]. The **SFE value** of the note is defined as the positive part of bA-B, where

A = the sum of the long scale percentages of all patterns

completing themselves at note t

B = the sum of the short scale percentages of all patterns

completing themselves at note t

and b>1 is a **balance factor** designed to compensate for the natural tendency of B to exceed A in all music. In the experiments I have done, a balance factor $1.5 < b < 2.5$ has yielded the most sensible results, but this may be different for non-fractal melodies.

The basic idea is that, if B exceeds A by a sufficient ratio, the SFE value is zero; if A exceeds B the SFE value is the amount of the excess. Thus the SFE value of a note is the extent to which the note affirms long term structure at the expense of short term structure.

Now, a good melody need not -- and in general **will** not -- have universally high SFE values. Sometimes short scale patterns need to be fulfilled; this the best way to build up the long scale intensity that gives **future** notes high SFE values. Therefore, in this context, the quality of a melody is most sensibly estimated as **the average of the SFE values of its notes**. This is a long way from being a "litmus test" for distinguishing good from bad melodies -- aside from the question of the general validity of the SFE theory, many serious simplifications have been made along the way. But it is a start.

Fractal Industrial Music

IFS's, as used here, generate only **note strings**. To get a real melody from a note string, one must deal with other features such as duration, velocity and volume. I have found that, while

velocity and volume can be set constant without great loss, the assumption of a constant duration results in unnecessarily "wooden-sounding" melodies. Therefore, in order to produce "listenable" melodies, I have introduced an **ad hoc** rule for producing durations. Namely: durations are chosen at random from the set {WHOLE, HALF, QUARTER, EIGHTH, SIXTEENTH}, but if a string of notes is repeated, they must all get the same duration. This rule ignores the manifold difficulties of identifying phrasestructure, and is obviously not intended a general solution to the problem of matching note strings with durations. However, the rule generates much better-sounding melodies than a constant- durations, IFS-generated-durations, or random-durations approach.

In addition to duration, **timbre** cannot reasonably be ignored. Melodies can sound quite different when played on different instruments -- for instance, few melodies sound their best played on the "beep" of an IBM PC speaker. In order to play the fractal melodies I have chosen sounds from the genre of **industrial** music. Generation of industrial music is a particularly appropriate application of fractal music because industrial music tends to be forgiving of strange pauses or "off" notes.

The genetic algorithm is of limited but definite use in the generation of industrial music with IFS's. If one runs a GA using the SFE-based fitness function, and plays each melody which the GA evolves, one will hear a better sequence of melodies than one would hear by trying IFS's at random. The improvement is not so dramatic as had been hoped, but is significant nonetheless.

A substantial part of this improvement is due to the elimination of extremely repetitive melodies, which consist primarily of long strings of one note repeated over and over. This might

seem to be a trivial kind of improvement, but it is not, because there is no obvious way to predict which IFS "code sequences" will give rise to repetitive attractors. The GA shifts the melody population until it resides primarily in that part of code sequence space containing few repetitive attractors. It tends not to converge quickly to a single melody, which is just as well, because the point is to generate a diversity of melodies. Eventually, however, it does converge.

One way to avoid this outcome is to introduce the Eugenic Genetic Algorithms developed above -- to replace the SFE fitness function f(x) with a fitness function g(x) defined by

$$g(x) = \begin{cases} 0, & f(x) < c \\ 1, & \text{otherwise} \end{cases} \quad (21)$$

Implementing this modified fitness function leads to an endless stream of novel, more-interesting-than-average melodies. Numerically, for long melodies (100-1000 notes), with b = 1.5 and c = .3, one obtains melodies with SFE value averaging around .35 or .4 (out of a possible 1.5); whereas the average melody has SFE value in the range .26 - .27. For short melodies the results are erratic and the variances are so high that I was not able to obtain meaningful averages.

Even without the GA, however, interesting-sounding fractal melodies are not that hard to come by. Long fractal melodies, the ones which the GA does the most consistent job of locating, tend to grow dull. IFS's seem to be best at generating short "riffs" or melodic ideas. My personal favorite is:

Example 1:

A WHOLE A# QUARTER A WHOLE B SIXTEENTH A# WHOLE

G# QUARTER F QUARTER E WHOLE C# EIGHTH D# EIGHTH

C# EIGHTH F QUARTER C# WHOLE F QUARTER C# EIGHTH

C EIGHTH B HALF B# WHOLE C SIXTEENTH A# EIGHTH

G QUARTER A EIGHTH A EIGHTH A EIGHTH G HALF

G HALF G HALF G HALF G# EIGHTH F # WHOLE

Played on a sampling keyboard through a feedback-infused guitar sample, against a background of crashing drums and cymbals, this is quite a natural-sounding riff.

Example 1 was obtained without the GA, by random search. The following melodies of length 30 were found by the GA, and are fairly typical of those produced by the genetic algorithm with a population size of twenty, after a dozen generations of search:

Example 2:

C# EIGHTH C# EIGHTH C# EIGHTH B EIGHTH G# HALF

D QUARTER E EIGHTH D SIXTEENTH A HALF E QUARTER

E QUARTER E QUARTER E QUARTER E QUARTER G HALF

E SIXTEENTH E SIXTEENTH F WHOLE F WHOLE D# QUARTER

E WHOLE D# QUARTER E EIGHTH D SIXTEENTH D SIXTEENTH

D SIXTEENTH D SIXTEENTH C QUARTER C QUARTER D WHOLE

Example 3:

E HALF D# QUARTER G SIXTEENTH F# EIGHTH C HALF

G EIGHTH F# EIGHTH F EIGHTH F EIGHTH F# EIGHTH

G# HALF G EIGHTH G EIGHTH G EIGHTH G EIGHTH

A# HALF A# QUARTER A# QUARTER D# SIXTEENTH

C# WHOLE B# WHOLE G# EIGHTH F# QUARTER G WHOLE

G WHOLE G WHOLE G WHOLE E QUARTER F# WHOLE

Example 3 is a fine argument in favor of the SFE approach to musical aesthetics. The existence of two separate strings of four repeated G notes, in itself, leads to an above average SFE value for the melody. For, the second time this string is reached, it does not fulfill any short-scale patterns, but it does fulfill the long-scale pattern set up by the repeated G's the last time they came around. The musical effect of this SFE would be ruined by random selection of durations, but the duration algorithm used works quite nicely; the four eighth-note G's appear as an "intentional foreshadowing" of the four whole-note G's, giving the essential impression of development over time. In fact, the repetition of the same note sequence with different durations is itself an example of "surprising fulfillment of expectations," though this fact was not picked up by the currently implemented SFE fitness function due to its exclusive focus on pitch values.

Finally, for sake of completeness, the following is a fairly typical example of a melody rejected by the GA:

Example 4:

C WHOLE C WHOLE C WHOLE C WHOLE C WHOLE

C WHOLE C WHOLE C WHOLE D QUARTER C# HALF

C# HALF C# HALF C# HALF C# HALF D WHOLE

D WHOLE D WHOLE D WHOLE D WHOLE C HALF

C HALF C HALF C HALF C HALF C HALF

C HALF C HALF C HALF C HALF C HALF

The 60 IFS coefficients which generated this melody look no different to the human eye than those which generated Examples 2 and 3; the genetic algorithm, however, is able to learn to predict, with a certain degree of accuracy, which code vectors will lead to such monotonous and obviously terrible melodies. In the SFE framework, the long repeated strings are strong short-scale patterns, which are not outweighed by any long-scale patterns.

On the Aesthetics of Computer Music

The use of the computer to generate pictures and music naturally raises the question of the **aesthetics** of these digitally-conceived art forms. In this section I will sketch out a few ideas on this subject, specifically focused on the fractal music generated in the previous section.

The question of musical aesthetics is a broad one. All sorts of factors affect the subjective judgement of musical quality: the timbre of the sounds involved, the comprehensibility of the melody, the nature of the harmonies, the kinesthetic "feel" of the rhythm, the listener's familiarity with the musical style.... Beneath all these different factors, however, there would seem to be some kind of core. The essence of musical quality, one feels, has to do with the patterns of the rising and falling of notes. Without attractive note patterns, none of the other factors will come into play in the first place. The essence of musical aesthetics resides in the single melody line.

A single melody line -- just a series of notes with varying pitch and duration, perhaps twenty to fifty notes long. This is a very discrete thing, just a selection from a finite number of permutations of pitch/duration combinations. The million dollar question is, why are some permutations so much more attractive than others? This question becomes particularly acute when one uses computer programs to generating melody lines. I have used perhaps a dozen different mathematical structures for generating melody lines, with widely varying results. Most, such as the chaotic attractors of quadratic maps), produce melodies that sound essentially random, despite the obvious presence of mathematical structure. Others, such as the invariant measures of large random one-dimensional iterated function systems, produce predominantly over-repetitive melodies. On the other hand, there is the odd mathematical structure which produces truly interesting and "catchy" tunes; an example, as described in the previous section, is the invariant measures of a small one-dimensional IFS (Goertzel, 1995). But the frustrating thing is that, even when one happens upon a structure that produces good melodies, one has no idea of why. What is it about small IFS's that "rings a bell" with the human mind?

The first step toward a resolution of the problem of melody lines, I propose, is the recognition that musical appreciation is a kind of emotion. In particular, it is a kind of pleasure. Even a

favorite sad piece of music brings intense pleasure as it moves one to tears. In order to understand why certain melodies are enjoyable, we must understand the structure of the emotions to which they give rise.

Mandler (1985) proposes that each emotion may be resolved into a "hot" part and a "cold" part. The hot part of the emotion is the "raw feel" of it, the conscious experience of the emotion. The cold part, on the other hand, is the underlying logic of the emotion, the interplay of other factors which gives rise to the emotion. When I speak of the structure of emotion, I am speaking of this "cold" part. The topic in question is the structure of the emotion of musical appreciation.

There are, basically, two approaches to the psychological analysis of the structure of emotion. The first is the combinatory approach, which begins with a few basic emotions and tries to build other emotions out of these. The second is what might be called the "frustrated expectations" approach, according to which an emotion occurs only when some mental process doesn't get what it expected. The two approaches are not necessarily incompatible, but they represent different perspectives.

The frustrated expectations approach originated with the French psychologist Paulhan in 1887. Paulhan was long on philosophy and short on specifics, but he did give one very precise definition. Happiness, he said, is the feeling of increasing order. And unhappiness is the feeling of decreasing order. Happiness occurs when processes receive an overabundance of order, more order than they expected; when chaos is surprisingly replaced with structure. On the other hand, unhappiness occurs when processes are let down by the absence of of order, when they get chaos but expected structure.

The combinatory approach, on the other hand, would seem to have a firmer biological foundation. There are certain emotions that seem to be grounded in human physiology. Rage and lust are the two most obvious ones: even reptiles would seem to experience these. Also, there is a feeling of warmth and tenderness that develops in mammals between the infant and its mother. There is a feeling of curiosity and exploration, which is easily detected on EEG recordings in humans, rats and monkeys.

It seems clear, however, that these basic emotions may be interpreted in a Paulhanian fashion. Lust, tenderness and exploration are all feelings of increasing order -- they are special kinds of happiness. On the other hand, rage is a feeling of decreasing order; it occurs when something threatens the order of one's world. So, on a very abstract level, we may arrive at a unified view of emotion. The question is, what does this general picture tell us about the specific emotion of musical appreciation?

As mentioned above, this path has been trodden before, by the excellent musical psychologist Meyer, in his book "Emotion and Meaning in Music." Meyer, drawing on the frustrated expectations theory of emotion, argues that high quality melodies work by frustrating expectations, and then fulfilling them. In this way they cause a temporary unhappiness, followed by a much greater happiness -- a feeling of a lack of order, followed by a more intense feeling of increasing order. It is worth noting that the temporary unhappiness is necessary for the ensuing happiness: there is a practical "ceiling" to the amount of order a mind can experience, so that,

after a certain point, in order to have the feeling of increasing order, one must first have the feeling of decreasing order. The key is to emphasize theincrease, the happiness, over its counterpart. I like to summarize Meyer's theory with the phrase, "surprising fulfillment of expectations," or SFE. A good melodic line is one that makes the mind feel: "Oh, wow! Yes -- that's what those notes were doing back there. I see now: it all fits. Beautiful!"

The trouble with Meyer's theory, however, is the difficulty of assessing what kinds of expectations are involved. Meyer, in his book, makes use of standard ideas from the theory of Western classical music. But these ideas are of little use in analyzing, say, Balinese gamelan or be-bop or punk rock, let alone experimental computer-composed music.

In the previous section I sought to use the theory of SFE as a guide for the computer composition of melodies. I wrote an algorithm which assessed the degree to which a given melody manifested the surprising fulfillment of expectations, and used a technique called the genetic algorithm to cause the computer to "evolve" melodies with a high SFE value. But the results, though exciting and in many ways successful, were not entirely positive. My implementation of SFE, in terms of repeated note, interval and contour patterns, did serve the role of filtering out pseudo-random melodies. But it did not filter out all of these and, what is worse, it did not reliably distinguish interesting melodies from boring, repetitive ones. Simple repeated patterns, balanced by a fixed algorithm, do not provide enough insight.

And this brings me to my key contention. I believe that the SFE theory is essentially correct. But this does not imply that the quality of a melody can be assessed by mathematical (as in my program) or music-theoretic (as in Meyer's book) criteria alone. it may be that certain note patterns have an intrinsic emotional value, due to their associations with other experiences. These note patterns, when used appropriately in a musical composition, will cause a much greater experience of pleasure than will other note patterns. A surprising fulfillment of expectations is all right, but not as good as a surprising fulfillment of expectations by the right kind of pattern!

But what is, precisely, the "right" kind of pattern? Here I will turn back to the 1800's again, to another philosophy of music: that of Arthur Schopenhauer. Schopenhauer, as is well known, argued that music had a special place among the arts. Music, he said, was closer than anything else to tracking the movements of the Will. The patterns by which notes move are very similar to the patterns by which our own free will moves. In Schopenhauer's view, the act of willing is only means by which we regularly come into contact with true reality; thus his philosophy of music gives music a very important role indeed!

Despite its excesses and peculiarities, I believe Schopenhauer's idea is a powerful one. Recall the two approaches to the study of emotion, discussed above: the frustrated expectations approach, and the combinatory approach. Meyer's theory of musical aesthetics corresponds to the frustrated expectations approach; and Schopenhauer's theory corresponds, a bit more loosely, to the combinatory approach. Schophenhauer argued that, in Nietzsche's phrase, "in music the passions enjoy themselves." The patterns of music, according to Schopenhauer, are precisely the same patterns according to which we act andfeel. The combination of musical elements into

musical compositions is like the combination of emotional and behavioral elements into complex actions and feelings.

So what does this all add up to? I suggest that, just as both theories of emotion are partly correct, so both theories of musical psychology are partly correct. One needs the structure of SFE, but one needs this structure to be supported by musical patterns that reflect our basic emotional and behavioral patterns, our psychic/bodily rhythms.

And, finally, what does this have to do with computer-composed music? At first glance, it might seem that the prognosis is not so good. Unless we can program a computer to determine those patterns which reflect our human actions and emotions, one might conclude, then computer music will continue to sound "sterile," as so much of it does today. But yet there are counterexamples; there are programs such as my own IFS algorithm, which generate interesting, listenable melodies that are not at all "sterile." Somehow, it would seem, this algorithm is hitting on some humanly meaningful structures. The IFS method itself is better at producing catchy melodies than my jury-rigged SFE algorithm is at distinguishing them from poor ones.

The IFS algorithm, as I have implemented it, produces structures that deviate from the "tonic" note, and then return to it, and then deviate again, and return, et cetera. And while it deviates and returns, it tends to repeat the same contour patterns. For short melodies, these repetitions do not occur quite often enough to become dull. This sort of overall structure has an obvious psychological meaning to it: in many different contexts, we journey to and from the same mental state, repeating on our different journeys the same patterns with minor variations. This is a crude but plain example of how mathematical structures may reflect psychological structures.

CHAPTER SEVEN

MAGICIAN SYSTEMS AND ABSTRACT ALGEBRAS

7.1 INTRODUCTION

Many systems theorists have stressed the "reflexive," "autopoietic" or "self-producing" nature of complex systems (Varela, 1978; Kampis, 1991; Goertzel, 1994; Palmer, 1994). This literature is intriguing and inspiring, and contains many detailed mathematical investigations. On the whole, though, a really useful mathematical model of the concept of reflexive self-production has never emerged. Varela's (1978) "Brownian Logic" may be considered a candidate for this role, as may be Kampis's "component-systems theory." But these theories are more conceptual than operational; they are not easily applied to obtain useful answers to specific questions about specific complex systems.

The goal of this chapter is to propose one possible path along which an **operational** mathematics of autopoiesis might be developed. I will take the simple notion of a **magician system**, introduced in Chapter One, and develop it into a more complete mathematical theory.

The magician system model is very general: it contains variants on the genetic algorithm, and also the psynet model of mind. Despite its generality, however, it is not without substance. It enforces a particular way of thinking about, and mathematically analyzing, complex systems.

I will explore both combinatorial and geometric perspectives on magician systems. First I will review how magician systems can be given a **geometric** formulation, in terms of directed hypergraphs. Then, for the bulk of the chapter, I will turn to algebra, and discuss how special cases of the magician system model can be formulated in terms of iterations on abstract algebraic structures. I will show by a detailed construction how the dynamics of an autopoietic system can be interpreted as a **quadratic** iteration on an abstract algebra. In the simplest case this algebra is a space of hypercomplex numbers. This construction implies a fundamental system-theoretic role for Julia sets and Mandelbrot sets over algebras -- for **hypercomplex fractals** and their yet more abstract cousins.

The conclusion that complex system dynamics can be expressed in the language of fractals over algebras has profound evolutionary consequences. For, evolution is often phrased in terms of the maximization of some "fitness criterion." But if the viability of a complex system depends on its position in certain hypercomplex Julia and Mandelbrot sets, then it follows that, in the evolution of complex systems, the **fitness criterionis a fractal**. If this is true, then the whole business of evolutionary fitness maximization is nowhere near so straightforward as has generally been assumed. Gregory Sorkin (1991) has provided the beginnings of a theory of simulated annealing on fractal objective functions; but we do not as yet have any theory at all regarding crossover-driven genetic optimization on fractal objective functions.

7.2 MAGICIAN SYSTEMS AND GENETIC ALGORITHMS

It is worth pausing for a moment to reiterate the mathematically obvious fact that the genetic algorithm is just a special kind of magician system model. Thus, the ideas of this chapter are a kind of generalization of the ideas of Chapter Six. For example, the "infinite population" theory given there could, in principle, be applied to other types of magician systems as well; however, the calculations would become even more difficult.

In the GA the magicians are the genotypes -- bit strings, real vectors, or whatever. The magician interactions are crossover: A acts on B to produce the offspring of A and B. The random nature of the crossover operator means that we have a **stochastic** magician system. The selection according to fitness may be modelled in several ways, the simplest of which is to include the **environment** itself as a magician.

To implement the standard GA, with survival proportional to fitness, the environment magician E acts on other magicians in the following way. First it acts on all the other magicians, in such a way that E*A = E', where E' is a modified E which contains information about the fitness of A. Then, having accumulated this information, it acts on the magicians in a different

way: E*A = A if E chooses A to survive, whereas E*A = -A, where -A is the magician that annihilates A, if E chooses A not to survive.

Of course, an environment magician E acting in this way is a somewhat contrived device, but this is because the selection mechanism of the simple GA is itself rather artificial. A more natural artificial life model would make survival dependent upon such operations as food consumption, location of shelter, etc. And these operations are very naturally representable in terms of magician action, much more so than the artificial mechanisms employed in the simple GA. The magician system model tends to push the GA in the direction of biological realism rather than (serial-computer) computational efficiency.

In the language of genetic algorithms, an "structural conspiracy" is nothing but a **schema**. The evolution of schemas was discussed in the previous chapter, in the context of the evolution of strange attractors for plane quadratic maps. A schema is a collection of similar individuals which tend to produce each other, under the crossover operator (i.e., under magician dynamics). The evolution of a GA population is commonly viewed as a path from broad schemas to narrow schemas, to yet narrower schemas. In dynamical-systems terms, this is the movement from one autopoietic subsystem to another attractorwhich is a **subset** of the first, to another attractor which is a subset of the second. From this point of view, the dynamics of the genetic algorithm are not peculiar at all, but are rather indicative of forces at play in a much broader class of complex, self-organizing systems. The genetic algorithm is a paradigm case of the interplay between **structure-maintenance** (autopoiesis; schema) and and **adaptive learning** (fitness-driven evolution).

Finally, it is worth noting that true autopoiesis, in the sense of spatial structure maintained by self-preserving dynamics, cannot appear in the GA which is non-spatial, but it can and does appear in the SEE model. Ecology transforms self- producing evolutionary subsystems into autopoietic subsystems.

7.3 RANDOM MAGICIAN SYSTEMS

Magician systems are a new mathematical entity; however, they have many interesting connections with familiar mathematical concepts. The remainder of this chapter will explore these connections, beginning with graphs, then moving to hypercomplex numbers and more complex abstract algebras, and nonlinear iterations on algebras.

An ordinary graph in the sense of discrete mathematics -- a collection of dots and lines -- is not sufficiently general to model a magician system. Instead, if one wishes to represent magician systems graphically, one needs to introduce the **dihypergraph**: a digraph in which each edge, instead of joining two vertices, joins three, four or more vertices. Formally, a dihypergraph consists of a collection V of "vertices," and a collection E of "hyperedges," each of which is an ordered tuple of elements of V. The length of the largest tuple in E is the "maximum edge size" of the dihypergraph.

It is easy to see how a magician system is a dihypergraph. If A and B combine to produce C, then one draws a directed edge from A and B to C. If A, B and C combine to produce D, then

one draws an edge (a directed hyperedge) from A, B and C to D. The ordered nature of the edges in the dihypergraph allows the noncommutativity of magician action: A can act on B, or B can act on A, and these need not yield identical results. Stochastic magician actions, whereby A can act on B yielding a variety of different possible results, can be modelled by **labelled** dihypergraphs, in which the label of an edge indicates the probability.

There are two ways to study ihypergraphs: on a case-by-case basis, or statistically. Here I will take the statistical approach; using some standard ideas from the theory of random graphs, I will look at **random dihypergraphs**. Technically, random graph theory applies only to graphs that are selected from sample spaces according to simple probabilistic models. However, several recent results (Chung and Graham, 1990) indicate that theorems obtained for **random** graphs can often be extended todeterministic graphs called "quasirandom" graphs without too much difficulty.

Using random dihypergraph theory, I will show that "phase transition"-like connectivity thresholds must exist for random magician systems. Inspired by this result, I will hypothesize that, in many circumstances, magician systems can only maintain autopoiesis in the neighborhood of the threshold. This implies that, in order to survive in the world, real magician systems must **evolve** the appropriate value for one of two parameters: either **size** or **connection probability**.

This idea, albeit speculative, is closely related to another concept which has attracted a great deal of attention in recent years: the "edge of chaos." A number of researchers (Packard, 1988; Langton, 1986) have come to the conclusion that complex systems are in a sense poised between order and chaos. The ideas given here suggest that this "edge of chaos" may be partially understood in terms of the threshold functions found in random graph theory.

Random Graphs

The basic fact about random graphs is the existence of **thresholds**. For example, suppose one has a collection of n vertices, and one connects these vertices with edges by choosing each edge independently with probability p = c/n. Then Erdos and Renyi (1960) showed that, for large n, the structure of the ensuing graph depends very sensitively on the value c. If c<1, then the graph almost surely has no components larger than O(logn). If c>1, on the other hand, then the largest connected subgraph almost surely contains all but O(logn) vertices. For the borderline case c=1, the largest connected subtraphs almost surely contains $O(n^{2/3})$ vertices.

This result is only the beginning. Choosing p = c logn/n, one finds disconnected graphs for c<1 and connected, Hamiltonian graphs for c>1. Choosing p just a little bigger, one finds graphs with arbitrarily large connectivity and minimum degree. Specifically, to get connectivity and minimum degree d, one must take

$$p = (\log n/n)[1 + (d-1)\log\log n/\log n + w_n/\log n] \quad (4)$$

where w_n tends to 0 arbitrarily slowly (Erdos and Renyi, 1960). And if one chooses p this way, then as a bonus one gets graphs which almost surely have **every specified degree sequence** $d_1 < d_2 < ... < d_k < d+1$ (Bollobas, 1985).

And connectivity is not the only graph property for which such a threshold exists. Bollobas (1985) defines a **monotone** graph property as any property which becomes more likely when one adds more edges to a graph, but keeps the number of vertices constant. He shows that **every monotone property** undergoes a "phase transition" at some point, just like connectivity does. Thresholds are not a fluky property of some specific mathematical function, but a basic conceptual, system-theoretic fact.

Random Dihypergraphs

Compared to the vast literature on random graphs, there has been surprisingly little work on random digraphs or hypergraphs. Lukasz (1990) has shown that the p=c/n threshold holds for digraphs; and Schmidt-Pruzan and Shamir (1983) have obtained a similar result for hypergraphs. Two of the authors (Goertzel and Bowman, 1993) have shown that the threshold for digraph connectivity also governs the behavior of fixed-length walks on random digraphs. Finally, the threshold existence results of Bollobas (1985) are basically set-theoretic in nature and can easily be seen to apply even to general random **dihypergraphs**.

Given the relative absence of work pertaining to random dihypergraphs, it is fortunate that many properties of random digraphs can be carried over to random dihypergraphs in a fairly straightforward manner. Connectivity is a prime example. To see how this works, one need only observe that a dihypergraph H induces a digraph G in a natural way. Namely, draw an edge from v to y in G iff there is an edge in H that **involves v and points to y**. This sort of "induction" is obviously a many-to-one relationship, in that the same digraph is induced by many different dihypergraphs; but this multiplicity need not be problematic. I may consider a dihypergraph to be "connected" if its induced digraph is connected, i.e. if for any v and y there is some edge leading from v **and some set of vertices** to to y

Suppose that I construct a random dihypergraph H by selecting each t-edge with probability

$$p^* = c/n^{t-1} \quad (5)$$

Then a simple calculation shows that each edge in the induced digraph G is chosen with probability c/n. So as c passes through the value 1, there is a bifurcation in the structure of the induced digraph. And this bifurcation, in itself, tells us something about the structure of the dihypergraph H.

First of all, if c<1, I know that the largest component of G is very small. This implies that the largest component of H is just as small. On the other hand, if c>1, I know that the chance of a path in G from v to y is very high; and it follows automatically that the chance of a path in H from v to y is the same. If almost all of G is connected, then almost all of H is connected. So, in short, if I assume equation (5), then the threshold at c=1 is there for dihypergraphs as well. Similarly, corresponding to the threshold function (4), I obtain for dihypergraphs

$$p = (\log n / n^{t-1})[1 + (d-1)\log\log n/\log n + w_n/\log n] \quad (6)$$

Not all questions about dihypergraphs can be resolved by reference to induced digraphs. However, the specific questions with which I am concerned here may be treated quite nicely in this manner. Below c=1, autopoietic magician systems of a reasonable size are quite unlikely. As c grows past 1, they get more and more likely. There is a critical point, past whichsignificant autocatalysis "suddenly" become almost inevitable. Then, as p passes c/n^{t-1} and $\log n/n^{t-1}$, connectivity gets thicker and thicker, until every vertex connects to arbitrarily many others.

Practical Implications of the Threshold

There is no reason to assume that the formula $p = c/n^{t-1}$ is adhered to by real magician systems. Consider, for instance, a system of proteins, some of which are enzymes, able to catalyze the relations between other enzymes (see e.g. Bagley et al, 1992 and references therein). This sort of enzyme system is naturally modeled as a magician system. What determines p in an enzyme system is the **recognitive ability** of the various enzymes. The parameter p represents the percentage of other enzymes in the system that a given enzyme can "latch onto."

What happens in reality is that p increases with n. For instance, if the average recognitive ability of enzymes remains fixed, and one adds more enzymes, then the chance of a reaction involving a particular enzyme will increase. But as one adds more and more enzymes, and p progressively increases, **eventually** p will reach a point where equation (5) is approximately valid. At this point, the network will just barely be connected. A few enzymes fewer, and virtually no pairs of enzymes will interact. A few enzymes more, and the system will be a chaotic hotbed of activity.

Or, on the other hand, suppose one holds n constant. Then one has a similar "optimal range" for **p**. In a system of fixed size, if the connection probability p is too small, then nothing will be connected to anything else. But if p is too large, then everything will connect to many other things.

Connectivity and Dynamics

What is the effect of the connectivity threshold on random magician dynamics? It is clear that sub-threshold conditions are not suitable for the origin of large autopoietic magician systems. A disconnected graph does not support complex dynamics. What I suggest is that, for many real-world dynamics, too much connectivity is **also** undesirable. This implies that useful dynamics tend to require a near-threshold balance of p and n.

Why would too much connectivity be bad? Well, consider: if the average degree is 1, then a change in one component will on average directly affect only one other component. The change will spread slowly. If the average degree is less than one, a change in one component will spread so slowly as to peter out before reaching a significant portion of the network. But if the average degree exceeds one, then a change in one component will spread exponentially throughout the

whole network -- and will also spread back to the component which originally changed, changing it some more.

Of course, this is only an heuristic argument. In some instances this sort of circular change-propagation is survivable, even useful. But even in these situations, I suggest, there issome **maximal useful average degree**, beyond which the "filtering" properties of the dynamics will cease to function, and nonproductive chaos will set in. This, then, is our key biological prediction:

Hypothesis. A typical magician system has a maximal useful average degree which is much less than n.

Given this, and assuming that a "typical" magician system is constructed in a roughly "quasirandom" way (Chung and Graham, 1990), then formula (4) above implies that the threshold for magician systems is not too far off from the simple threshold at $p = c/n^{t-1}$.

The Origin of Life, and the Edge of Chaos

To make these ideas more concrete, it is perhaps worth noting that they have an intriguing, if speculative, connection to the question of the origin of life. Oparin's classic theory (1965; see also Dyson, 1982) suggests that life initially formed by the isolation of biomolecules inside pre-formed inorganic solid barriers such as water droplets. These inorganic "cell walls" provided the opportunity for the development of metabolism, without which the construction of **organic** cell walls would have been impossible. They provided part of the "scaffolding" on which life was built.

The details of this process are well worth reflecting on. All sorts of different-sized collections of biomolecules could have become isolated inside water droplets. But only when the **right number** were isolated together did interesting dynamics occur, with the possibility of leading to metabolic structure. This lends a whole new flavor to the Oparin theory. At very least, it should serve as a warning to anyone wishing to calculate the **probability** of metabolism forming by the Oparin scheme. One cannot talk about metabolic networks without taking the possibility of threshold phenomena into account.

This way of thinking is very closely related to the idea of the "edge of chaos," reported independently by Norman Packard and Chris Langton and discussed extensively in Lewin's book *Complexity*, among other places. The idea is that most complex systems operate in a state where their crucial control parameters are poised between values that would lead to dull, repetitive order, and values that would lead to chaos. Packard and Langton found this to be true in simulations of cellular automata. Stuart Kauffmann (1993) discovered the same phenomenon in his experiments with random Boolean networks: he found, again and again, a strict connectivity threshold between dull static or periodic dynamics and formless chaotic dynamics.

One cannot rigorously derive the results of Packard, Langton and Kauffmann from random graph theory, at least not in any way that is presently apparent. But the relation between their results and the graph-theoretic ideas of this section are too obvious to be ignored. If there is ever

to be a precise theoryof the edge of chaos, one feels that it will have to have something to do with random graphs.

7.4 HYPERCOMPLEX NUMBERS AND MAGICIAN SYSTEMS

Having connected magician systems with graph theory, we will now turn to a different branch of mathematics -- abstract algebra. As it turns out, certain types of magician systems have a very natural representation in terms of the algebra of **hypercomplex numbers**.

Hypercomplex numbers are obtained by defining a **vector multiplication table** on ordinary d-dimensional space (R*d*). The vectors form an algebra G under multiplication and a vector space under addition; the multiplication and addition interact in such a way as to form a ring. Each coordinate represents a basic system component, and thus each vector represents a "population" of components. A vector coupled with a **graph of interconnections** constitutes a complete model of a system. The vector multiplication signifies the **intercreation** of components; i.e., A * B = C is interpreted to mean that component A, when it acts upon component B, produces component C.

In order to express magician systems in terms of hypercomplex numbers, we will at first consider the simplest case of magician systems: deterministic magicians, who are living on a complete graph, so that every magician is free to act on every other magician. In this case, when formulated in the language of hypercomplex numbers, magician dynamics can often be reduced to a simple **quadratic iteration**, $z_{k+1} = z_k 2$, where the z_i are vectors in R*d*, and the power two is interpreted in terms of the G multiplication table. If one adds a constant external environment, one obtains the equation $z_{k+1} = z_k 2 + c$, familiar in dynamical systems theory for the case where n=2 and the multiplication table is defined so that instead of hypercomplex numbers, one has the complex number system.

One of the most striking implications of this approach is that the the hypercomplex equivalents of **Julia sets** and **Mandelbrot sets** (Devaney, 1988) play a fundamental role in autopoietic system dynamics. This follows directly from the appearance of the equation $z_{k+1} = z_k 2 + c$, noted above. Specifically, the Julia set of a system, in a certain environment, contains those initial system states which lead to relatively stable (i.e. bounded) behavior. The Mandelbrot set contains those environments which lead to bounded behavior for a large contiguous region of initial system states.

The theory of Julia and Mandelbrot sets may need to be substantially generalized in order to apply to hypercomplex number systems (let alone to more general algebras), but the basic point remains: the dynamical properties which Julia sets and Mandelbrot sets identify in the complex plane, when translated into an hypercomplex setting, are precisely the **simplest** of the many properties of interest to autopoietic system theorists. For the first question that one asks of anautopoietic system is: Is it viable? Does it successfully sustain itself? Does it work? Julia sets and Mandelbrot sets address this question; they speak of **system viability**. Other system properties may be studied by looking at appropriately defined **subsets** of Julia sets.

So far as I have been able to determine, no serious mathematical work has been done on regarding hypercomplex Julia sets. However, several researchers have explored these sets for artistic purposes. Alan Norton produced videos of quaternionic quadratic maps, entitled *Dynamics of ei theta x(1-x)* and *A Close Encounter in the Fourth Dimension*; the background music in these videos, composed by Jeff Pressing, was constructed from the same equations and parameter values (Pressing, 1988). More recently, Stuart Ramsden has produced a striking video of the Julia sets ensuing from the iteration z*2* + c over the real quaternion algebra, using a new graphical method called "tendril tracing" (Ramsden, 1994). This kind of graphical exploration becomes more and more difficult as the dimensionality of the space increases; for after all, one is projecting an arbitrarily large number of dimensions onto the plane. But nevertheless, this sort of work is worthy of a great deal of attention, because what it is charting is nothing less than the dynamics of autopoietic systems.

Finally, I must be absolutely clear regarding what is **not** done in this chapter. I will not solve any of the difficult and interesting mathematical problems opened up by the concept of "hypercomplex fractals"; nor will I give serious practical applications of this approach to complex system modeling. The purpose of this chapter is quite different: it is simply to call attention to a new approach which appears to hold a great deal of promise. The hope is to inspire mathematical work on quadratics on hypercomplex number systems and related algebras, and practical work formulating system structure in terms of hypercomplex numbers and related algebras. In the final section I will propose a series of conjectures regarding complex systems and their algebraic formulations. The resolution of these conjectures, I believe, would be a very large step toward the construction of a genuine complexity science.

Magician Systems as Abstract Algebras

Having made appropriate restrictive assumptions, it is perfectly easy to express magician systems as abstract algebras. Let us begin with the case in which there is only one algebraic operation, call it "+". In this case, it is very natural to declare that the algebraic inverse of element A is the antimagician corresponding to A. And, once this is done, one would like the annihilation of magicians and antimagicians to be expressed by the identity A + -A = 0. The element "0" must be thus be understood as an "impotent magician," which is unable to annihilate anything because it is already null.

But if we are using the operation + to denote annihilation, then we need another operation to indicate the **action** ofmagicians upon one another. Let us call this operation "*," and let A * B refer to the **product** resulting when magician A casts its spell upon magician B. Thus a magician can act upon its own opposite without annihilating it, since -A * A need not equal 0.

This notion of + and * naturally leads one to view magician systems as **linear spaces** constructed over algebras. One need only assume that, where M denotes the magician system in question, {M,+,*} forms a ring. For instance, suppose one has a magician system consisting of 3 copies of magician A, 2 copies of magician B, and 4 copies of magician C. Then this system is nicely represented by the expression

3A + 2B + 4C

The result of the system acting on itself is then given by the expression

(3A + 2B + 4C) * (3A + 2B + 4C).

For what the distributive law says is that each element of the first multiplicand will be paired exactly once with each element of the second multiplicand. The production of, for instance, 12 A * C's from the pairing of the first 3A term with the last 4C term makes perfect sense, because each of the three A magicians gets to act on each of the four C magicians.

The annihilation of magicians and antimagicians is taken care of automatically, as a part of the purely mechanical step by which multiple occurences of the same magician are lumped together into a single term. For instance, consider

(A + B) * (A + B) = A*A + A*B + B*A + B*B

and suppose the magician interactions are such that

A*A = -B

A*B = B*A = B*B = B

Then the result of the iteration will be

(A + B) * (A + B) = -B + B + B + B = 2B

The single antimagician -B annihilates a single magician B just as the magician dynamic indicates, while the other two B's are combined into a single term 2B. Whereas the annihilation serves a fundamental conceptual function relative to the magician system model, the combination of like terms does not; yet both are accomplished at the same time.

As observed above, it is most natural to assume that -A * B = -(A * B). This means one has a genuine hypercomplex number ring; and furthermore it makes perfect intuitive sense. It means that the antimagician for A not only annihilates A, but in every situation acts in such a way as to produce precisely the antimagician of the magician which A would have produced in that situation. In this way it annihilates the **effects** of A as well as A itself, making a clean sweep of its annihilation by makingthe system just as it would have been had A never existed in the first place.

More generally, if the initial state of a magician system is represented by the linear combination $z_0 = c_1M_1 + ... + c_NM_N$, where the c_i are any real numbers representing "concentrations" of the various magicians, then the subsequent states may be obtained from the simple iteration

$z_n = z_{n-1}2$ (7)

And if one has a magician system with a constant external input, one gets the slightly more complex equation

$$z_n = z_{n-1}2 + c \quad (8)$$

where c is a magician system which does not change over time.

Julia Sets

As observed in the Introduction, the latter equation suggests a close relationship between magician systems and **Julia sets**. For, suppose one has a system of five magicians called 0, 1, i, -1 and -i, with a commutative multiplication obeying i * i = -1. This particular magician system, when used as the algebra G for an hypercomplex number space, yields the complex numbers; and the equation for an externally-driven magician system is then just the standard quadratic iteration in the complex plane. The Julia set corresponding to a given "environment" value c contains those initial magician system states which do **not** lead to a situation in which some magician is copied infinitely many times; and it also contains all values c which are **limit points** of stable environment values of this type. The boundary of the Julia set thus demarcates the viable realm of initial system states from the non-viable realm (note that what I call the **Julia set**, some authors call the "filled Julia set," reserving the term "Julia set" for the boundary of what I call a Julia set). The Mandelbrot set, on the other hand, is the collection of environments c for which the Julia set is connected, rather than infinitely disconnected (it is known that these are the only two possibilities; see Devaney, 1988).

How do these concepts generalize to other instances of hypercomplex numbers? So far as I know, the answer to this question is at present unknown. The most basic results, such as the non-emptiness of the Julia set of a polynomial mapping, all depend on the fact that a polynomial in the complex number algebra has only **finitely many roots**. This is not true in an arbitrary hypercomplex number system, and thus the mathematical theory for the more general case can be expected to be very different. So far mathematicians have not focused their attention on these questions; but this is bound to change.

Incorporating Space or Stochasticity

The above equations represent only the simplest kind of magician system: every magician is allowed to cast its spell on every other magician at every time step. The possibility of a spatial graph of interconnection is ignored. If one introduces space then things become much more awkward Where $z_{n-1} = c_1M_1 + ... + c_dM_d$, the iteration becomes

---

Rough Equation

z_n = SUM FROM {i,j} p_{ij} M_i M_j + c3

---

(9)

where the p_{ij} are appropriately defined integer constants. The simple case given above is then retrieved from the case where the p_{ij} are equal to appropriate binomial coefficients. While not as elegant as the equations for the non-normalized case, these more general iterations are still relatively simple, and are plausible objects for mathematical study.

The same approach works for studying **stochastic** dynamics, provided that one is willing to ignore genetic drift and use an "iterated mean path" approximation. Only in this case the interpretation is that each constant c_i is the **probability** that an arbitrarily selected element of the population is magician M_i. Thus the p_{ij} are not integers, they are determined by the expected value equation, and they work out so that the sum of the c_i remains always equal to 1.

Single and Multiple Action

We have not yet introduced **single** action, whereby a magician A, acting all by itself, creates some magician B. But this can be dealt with on a formal level, by introducing a dummy magician called, say, R, with the property that R*R = R, so that R will always perpetuate itself. To say that A creates B one may then say that R*A = B.

Is there a similar recourse for the case of triple interactions? As a matter of fact there is; but it involves two steps. Say one wants to express the fact that A, B and E combine to form C -- something that might, in a more explicit notation, be called

$$*_3(A,B,E) = C \quad (10)$$

The way to do this is to introduce a dummy magician called R, so that A * B = R and E * R = C. This reduction goes back to Charles S. Peirce (1935) and his proof that Thirds, or triadic relations, are sufficient to generate all higher-order relations.

The trouble with this reduction is that it takes two time steps instead of one. Somehow one must guarantee that R survives long enough to combine with E to produce C. The easiest thing is to have R produce itself. To guarantee that R does notpersist to influence future reactions, however, one should also have C produce -R. Thus R will survive as long as it is needed, and no longer.

By this sort of mechanism all triple, quadruple and higher interactions can be ultimately reduced to sequences of paired interactions. In practice, it may sometimes be more convenient to explicitly allow operations $*_i$ which combine i magicians to form an output. But these operations may be understood as a "shorthand code" for certain sequences of pairwise operations; they do not need to enter into the fundamental theory.

7.5 EMERGENT PATTERN

So far we have two algebraic operations: + meaning cancellation and formal summation, and * meaning action. Two is a convenient number of algebraic operations to have: nearly all of abstract algebra has to do with systems containing one or two operations. Unfortunately, however, in order to give a complete algebraic treatment of autopoietic systems, it is necessary to

introduce a **third** operation. The reason is the phenomenon of **gestalt**. In the context of pattern recognition systems, for example, gestalt expresses itself as **emergent pattern**. In many cases, a pattern A will emerge only when the two entities B and C are considered **together**, and not when any one of the two is considered on its own.

Our algebra (M,+,*) gives us no direct way of expressing the statement "A is a pattern which emerges between B and C." There is no way to express the word **and** as it occurs in this statement, using only the concepts of cancellation/summation and action. Thus it becomes necessary to introduce a third operation #, with the interpretation that A#B is the entity formed by "joining" A and B together into a single larger entity. The operation # can unproblematically be assumed to have an identity -- the zero process, which combines with A to form simply A.

In this view, to say that A is an emergent pattern between B and C, is to say that A is a pattern in B#C. If D is the process which recognizes this pattern, then we have

D * B#C = A (11)

where it is assumed that # takes precedence over * in the order of operations.

To get a handle on the operation, it helps to think about a simple example. Suppose the magicians in question are represented as binary sequences. The most natural interpretation of # is then as a **juxtaposition** operator, so that, e.g.,

001111 # 0101 = 0011110101

This example makes the meaning of inversion rather obvious. The expression B-*1*A refers to the result of the following operation: 1) if A has an initial segment which is identical to B, removing this segment; 2) if A does not have an initialsegment which is identical to B, leaving A as is. Thus inverses under # are unproblematic if properly defined.

It is also clear that # behaves distributively with respect to addition:

C#(A+B) = C#A + C#B (12)

(A+B)#C = A#C + B#C

Thus M(+,#) would be a ring, if one mades the somewhat unnatural assumption that 0 # A = 0 (here 0 is the zero of the additive group; it is the "empty magician system."). This equation, however, says that the result of juxtaposing with the empty magician system is the empty magician system. Unfortunately, it would seem much more logical to set

0#A = A (13)

thus equating the multiplicative and additive identities and sacrificing the ring structure; but on the other hand it is hard to see how the definition 0#A=0 could do any real harm.

The relation between # and * is even more difficult. Juxtaposition is clearly a **noncommutative** operation, since

0101 # 001111 = 0101001111 ;

so it would seems that in general # must be considered noncommutative. But with a little ingenuity, a kind of weak "one-sided commutativity" may be salvaged, at least in the important special case (discussed above) in which the operation * is an operation of **pattern recognition**. For, consider: the operation which takes A#B into B#A has a constant and in fact very small algorithmic information; and so, in the limit of very long sequences, the structure of A#B and the structure of B#A will be essentially the same. This means that

C * (A#B) = C * (B#A) (14)

for all C; and in fact even for fairly short sequences the two sides of the equation will be very close to equal.

On the other hand, under the straightforward juxtaposition interpretation the reverse is not true, i.e.

(A#B) * C not = (B#A) * C (15)

because the **action** of the juxtaposition A#B may depend quite sensitively on order. If A*B is defined as the output of some Turing machine M given program A and data B, then clearly equality is not even approximately valid, since A#B and B#A need not be similar as programs.

Next, and most crucially, what about distributivity with respect to *? In the juxtaposition interpretation this plainly does not hold. But by modifying the juxtaposition interpretation slightly and inoffensively, one may salvage the rule. Suppose that, when juxtaposing two sequences, one places a **marker** betweenthem, as in

001111 # 0101 = 001111|0101

This will not affect the patterns in the sequence significantly. Then, suppose that one changes the computational interpretation of the sequences appropriately, so that a "|" marker indicates to the Turing machine that it should run two separate programs, one consisting of the part of the sequence before the |, the other consisting of the part of the sequence after the |, and that when it has finished running the two programs it should juxtapose the results. If one adopts this interpretation then one finds that, pleasantly enough,

(A#B) * C = A*C # B*C (15)

The reverse kind of distributivity is false; as a rule

A * (B#C) and A*B # A*C are not equal. However, in the case of pattern recognition processes, the equality

$$A * (B\#C) \;?=\; A*B + A*C \quad (16)$$

will hold a great deal of the time. It will hold in precisely those cases where there A detects no emergent pattern between B and C. The **difference**

$$A * (B\#C) - A*B - A*C \quad (17)$$

is the emergent pattern which A detects between B and C.

So we end up with a peculiar and possibly unique kind of algebraic structure. (M,+,*) is a ring; (M,+,#) is a not a ring due to the rule 0#A=A; and (M,#,*) is not a ring due to the lack of left-sided distributivity, but it is what is known as a **near- ring** (Pilz, 1977). This unsightly conglomeration, we suggest, is the algebra of autopoiesis, the algebra of complexity, the algebra of mind. Until this algebra is understood, complex systems will remain largely uncomprehended.

Incorporating emergence into the magician dynamic yields, in the simplest case, the following iteration:

$$z_n = z_{n-1} * (z_{n-1} + z_{n-1}\#2) \quad (18)$$

This case assumes that only emergences between **pairs** are being recognized. To incorporate emergences between triples yields

$$z_n = z_{n-1} * (z_{n-1} + z_{n-1}\#2 + z_{n-1}\#3) \quad (19)$$

and the most general case gives an infinite series formally represented by

$$z_n = z_{n-1} * [\, z_{n-1} \# (1_\# - z_{n-1})\#{-1} \,] \quad (20)$$

(where $1_\#$ denotes the identity of the semigroup (M,#)).

What are the analogues of Julia sets and Mandelbrot sets forthese unorthodox iterations? The mathematics of today gives few clues. But these are the questions which we must answer if we want a genuine science of complex systems.

Finally, it is worth observing that this third operation # is not technically necessary. As with multiple interactions, however, the operation # may be expressed in terms of (M,+,*), if one is willing to avail oneself of unnatural formalistic inventions. All that is required here is a magician called, say, **Jux**, with the property that

$$*_3(A,B,Jux) = A\#B \quad (21)$$

This reduces the product # to an $*_3$ magician product, and hence, by our earlier reduction, to a series of ordinary magician products. Expressing it in this way conceals the algebraic properties of juxtaposition, but for situations in which these properties are not important, it may be more convenient to have only two operations to deal with.

This uncomfortable reduction reminds one of the limitations of hypercomplex numbers. Ideally, one would like to be able to deal with **any algebra that happens to come up** in the context of analyzing a complex system (this point is made very forcefully in (Andreka and Nemeti, 1990)). One would like to be able to deal with dynamics on (M,+,*,#) just as easily as on (M,+,*). Whether this will ever be the case, as the saying goes, "remains to be seen." In any event, I suggest that the hypercomplex numbers are an excellent place to begin. Generalizations to other algebras are important, but if hypercomplex numbers are too difficult, then more esoteric algebras would seem to be completely out of reach.

7.6 ALGEBRA, DYNAMICS AND COMPLEXITY

So, many complex systems can be naturally viewed as quadratics on hypercomplex numbers and related algebras. So what?

Formalism, in itself, confers no meaning or understanding. There is nothing inherently to be gained by formulating a complex system as an abstract algebra. But the idea is that, by doing mathematical analysis and computer simulation of iterations on algebras, one should be able to obtain practical insights into complex system behavior. At present this is more a vision than a reality. But my belief is that this research programme has an excellent chance of success.

In this concluding section, I will first seek to put the model of the previous section in perspective, by relating it to previous work on the algebraic modeling of complex systems. Then I will propose a series of **crucial questions** regarding the behavior of iterations on hypercomplex numbers. Finally, I will briefly discuss the possible evolutionary implications of the association between complex systems and fractals on abstract algebras.

Algebra, Dynamics and Complexity

I am not the first to seek a connection between complex systems and abstract algebra. In an article by H. Andreka and I. Nemeti, entitled "Importance of Universal Algebra for Computer Science," one finds the following statement:

Badly needed is a **theory of complex systems** as an independent but exact mathematical theory aimed at the production, handling and study of highly complex systems. This theory would disregard the system's origin, its nature etc., the only aspect it would concentrate on would be its being complex. ...

We conclude by taking a look at today's mathematics trying to estimate from which of its branches could a new mathematics "of quality" emerge forming the basis of this (not yet existing) culture of complex systems. (More or less this amounts to looking for the mathematics of general system theory (in the sense of Bertallanfy, Ashby and their followers)...).... It is universal algebra, category theory, algebraic logic and model theory which seem promising for our problems.

Not only do they dismiss the vast formalism of "complexity science" that has emerged within physics, chemistry and applied computer science; but, despite their central interest in

understanding complex **computer programs**, they also dismiss the traditional theory of computer science. The answer to the problems of complexity science, or they claim, lies nowhere else but in **abstract algebra** and related areas of **mathematical logic**. Of the four branches of algebra and logic which they identify, **universal algebra** is the most prominent in their own work; and indeed, if interpreted sufficiently broadly, universal algebra can be understood to encompass a great deal of logic and model theory. For universal algebra is nothing more or less than the study of **axiom systems and algebras in the abstract**:

> Abstract algebra ... contains many... axiom systems, e.g. group theory, lattice theory, etc. Universal algebra no more puts up with the study of finitely many fixed axiom systems. Instead, the axiom systems themselves form the subject of study.... Universal algebra investigates those regularities which hold when we try to describe an arbitrary phenomenon by some axioms we ourselves have chosen.

Now, this sounds very promising, but the trouble with this proposal is that, from the practical scientist or the computer programmer's point of view, universal algebra is a disappointingly **empty** branch of mathematics. There are very few useful, nontrivial theorems. The same can be said for category theory, model theory and algebraic logic. There are some neat technical results, but it is all too abstract to have much **practical** value. In short, even if one expressed a certain complex system in this kind of mathematical language, one would not be able to **use** this expression to draw meaningful conclusions about the ideosyncratic behavior of that particular system. Of course, it is impossible to foresee the future development of any branch of mathematics or science; it is possible that these fields of study (which have largely fallen out of favor, due precisely to this paucity of interesting results) will someday give rise to deep and applicable ideas. But until this happens, the programme of Andreka and Nemeti remains a dream.

Although the ideas of this chapter were developed before I had heard of the work of Andreka and Nemeti, in retrospect it would seem to complement their research programme quite nicely. There are, however, two major differences between my approach and theirs. First, I derive algebraic structures from general system-theoretic models, rather than from models of particular systems. And second, I lay a large stress on **dynamical iterations** on these algebras, rather than the algebraic structures themselves.

Regarding the first difference, although I do not claim that hypercomplex numbers are the only algebraic structures of any use to complexity science, I do claim that these rings are of fundamental system-theoretic importance. I have proposed alternative algebraic structures for modeling certain aspects of system behavior, and though I have shown that these alternative algebraic structures can be formally **reduced** to hypercomplex numbers, I have not demonstrated the practical utility of these reductions. Therefore I believe that, eventually, it will be necessary to develop a flexible theory of **polynomial iterations on algebras**. However, one must start somewhere, and given the general utility of the hypercomplex numbers for modeling magician systems, I believe that they form a very good starding point.

Next, regarding the second difference, it must not be forgotten that, if universal algebra suffers from a lack of examples, dynamical systems theory almost suffers from a **surplus** of examples.

Drawing on the incomparable power of differential and integral calculus, it is by far the most operational approach to complexity science yet invented. The approach which I have taken here is intended to take advantage of the strengths of both approaches: to combine the structural freedom of abstract algebra with the operationality of dynamical systems theory. Thus, instead of embracing the utter generality advocated by Andreka and Nemeti, I restrict myself initially to the hypercomplex numbers, and propose to transfer some of the ideas and methods of dynamical systems theory to this context.

Despite its analytical power, however, as noted in Chapter One, the contemporary mathematical theory of dynamical systems is somewhat restricted in scope. First of all, nearly all of dynamical systems theory deals with deterministic rather than stochastic systems. And very little of the theory is applicable, in practice, to systems with a large number of variables. Much current research in dynamical systems theory deals with "toy iterations" -- very simple deterministic dynamical systems, often in one or two or three variables, which do not accurately modelany real situation of interest, but which are easily amenable to mathematical analysis. Implicit in the research programme of dynamical systems theory is the assumption that the methods used to study these toy iterations will someday be generalizable to more interesting iterations.

The approach of the present chapter, though falling into the general category of "dynamics," pushes in a different direction from the main body of dynamical systems theory. Instead of looking at more and more complicated iterations on low- dimensional real and complex spaces, I propose to look at simple iterations such as $z_{n+1} = z_n^2 + c$ and its relatives, but on relatively high-dimensional spaces with unorthodox multiplications. The place of complexity science, I suggest, is here, halfway between the unbridled generality of universal algebra and the rigid analytic conformity of conventional dynamical systems theory.

7.7 SOME CRUCIAL CONJECTURES

In this section I will present a list of conjectures, the exploration of which I believe to be important for the development of the "complex systems and hypercomplex fractals" research programme. All of these conjectures basically get at the same point: the crucial thing to study is the nature of the multiple mappings by which systems give rise to algebras, algebras give rise to Julia sets, and Julia sets describe systems. (We will use the term "Julia set" loosely here, to refer not only to the Julia sets obtained from nice mappings like $z^2 + c$, but also to the Julia-set-like entities obtained from the less tidy mappings also discussed above.)

Conjecture 1. Different types of complex systems give rise to different types of multiplication table.

It would seem that, by verifying this conjecture, a fair amount of insight could be obtained from hypercomplex numbers without doing hardly any mathematics or simulation. Do immunological systems have a characteristic type of multiplication table? What about psychological systems, ecological systems, economic systems? Do different personality types give rise to different multiplication tables? What about different economies? The reduction to hypercomplex numbers gives an interesting new method for comparing and classifying complex

systems, related but not identical to standard differential and difference equation approaches. Palmer (1994) has proposed that more and more complex systems correspond to more and more complex algebras: quaternions, octonions, and so forth. Whether or not these details are correct, it seems plausible that some such correspondence holds.

Conjecture 2. Simpler multiplication tables give rise to simpler Julia sets.

Up to a certain point, this question could be approached in a purely intuitive way, by producing a large number of hypercomplex Julia sets and observing their complexity. Eventually, however, one wishes to have a formal understanding, and here one runs into the problem of defining "simplicity." One could take the approach of algorithmic information theory, defining the simplicity of a multiplication table as the length of the shortest program required to compute it, and the simplicity of a Julia set in terms of the lengths of the shortest programs required to compute its discrete approximations; but this approach makes the statement almost obvious, since the shortest program for computing a Julia set will probably be to run the quadratic iteration, and the bulk of the program required for computing the quadratic iteration will generally be taken up by the multiplication table. Instead, one wishes to measure the **visual** complexity of the Julia set; say, the length of the shortest program for computing the Julia set which is inferrable by some learning algorithm that knows nothing about quadratic iteration, say a learning algorithm that works by extracting repeated patterns (Bell, Cleary and Witten, 1990).

Conjecture 3. Projections of Julia sets be used to help "steer" complex system behavior.

Suppose that one had a flexible method for computing Julia sets of high-dimensional hypercomplex algebras in real time (of course, this will never be possible for **arbitrary** high-dimensional hypercomplex algebras, but as just observed above, it seems reasonable to expect that complex systems will lead to structured rather than arbitrary multiplication tables). Today's workstations may not be up to the task of rapidly simulating Julia sets over 75-dimensional hypercomplex algebras, but those of ten years hence may well be. Thus one can imagine, in not too distant future, using the following method to study a complex system. First, represent the system as a magician system. Then, locate the system's initial state within the Julia set of the system. Next, test out possible changes to the system by observing whether or not they move the state of the system out of the Julia set. In this scheme, an interactive movie of the projected Julia set would be much more than an attractive toy, it would be an indispensable tool for planning, for system steering. The beautiful complexity of the well known complex number Julia sets are a warning against the assumption that system viability follows simple, linear rules. System viability follows intricate fractal boundaries; and surely the same is true for subtler aspects of system behavior.

Conjecture 7. Similar systems give rise to similar Julia sets.

This is simply the question of how the **structure** of a Julia set depends on the structure of the algebra which generated it. I have already suggested that unstructured algebras give rise to unstructured Julia sets. This idea naturally leads to thehypothesis that algebras which share common structures will tend to give rise to Julia sets that share common structures. If this kind of

"structural continuity" holds true it has a profound system-theoretic meaning: it says that patterns in system **structure** are closely tied with patterns in system **viability**.

It seems at least plausible that our intuitive ability to manage complex systems rests on an implicit awareness of this correspondence. We tend to assume that structurally similar systems will operate similarly, and in particular that they will be viable under similar circumstances. But, from the perspective of autopoietic algebra, this assumption is seen to depend on the peculiar relation between hypercomplex Julia sets and their underlying multiplication tables.

Conjecture 5. There are close relations between the Julia sets corresponding to different views of the same system.

To explore this conjecture one must first decide what one means by "different views." One natural approach is to use ring theory. Recall the definition of an **ideal** of a ring A: a subset U of A is an ideal of A if u*a and a*u are in U for every a in A. It is particularly interesting to apply this idea to the special case, discussed several times above, in which the multiplication * connotes both pattern recognition and action: then a collection of patterns U is an ideal if U contains all patterns recognized **by** processes in U, and all patterns recognizable **in** processes in U. In other words, in this case, an ideal is a closed pattern system.

One can find ideals by running the following iterative process until it stops: beginning with a set of patterns S_0, let S_{i+1} denote all the patterns recognized by and in the pattern/processes in S_i. An ideal U defines a ring of cosets A/U, whose elements are of the form U + a, in which U + a and U + b are considered equivalent if there are u and v in U so that u + a = v + b. Addition and multiplication on the coset ring are defined by the obvious rules (U+a) + (U+b) = U + (a+b), (U+a) * (U+b) = U + a*b. The ring A/U represents the magician system A as viewed from the subjective perspective of the subsystem U. The homomorphism theorems tell us that A/U is homomorphic to A, so that the "subjective world" of U is in a sense a faithful reflection of the "objective world" A. In other words, there is a map f from A to A/U with the property that f(ab) = f(a)f(b) and f(a+b) = f(a) + f(b); this map determines the "subjective" correlate f(a) = U + a of the "objective" entity a.

In this ring-theoretic perspective, our question is: what is the relation between the Julia set for a quadratic over A, and the Julia set for the same quadratic over A/U. What is the relation between the Julia set over A/U and the Julia set over A/V? How does the guaranteed homomorphism carry over into a **geometrical** correspondence between the different Julia sets? One might speculate that the Julia set for A/U would maintain the same overall structure as that for A, but would omit certain patterns and details. However, this is only a speculation, andthe matter is really not clear at all.

The emphasis on Julia sets in these conjectures is somewhat arbitrary. Julia sets have to do with system viability, but viability is only one property of a system. Just as interesting are more specific conjectures of system **dynamics**, i.e., conjectures regarding the individual **trajectories** which must be computed in order to get Julia sets. In fact, conjectures 2, 4 and 5 may be reformulated in terms of trajectories, to obtain new and perhaps equally interesting conjectures:

2'. Do simpler multiplication tables give rise to simpler average trajectories?

4'. Do similar systems tend to give rise to similar trajectories?

5'. Are there relations between the trajectories corresponding to different views of the same system?

Furthermore, the system steering tool proposed in Conjecture 3 could also be generalized to deal with system trajectories. Instead of just testing for viability, one could test the properties of the trajectories obtained by changing certain elements of the system.

And, just as the viability/instability dichotomy gives rise to Julia sets, so every other property of trajectories gives rise to its **own** characteristic set. These other sets are generally subsets of the Julia set; they are fractals within a fractal. For instance, suppose one wants to look for system states in which all magicians keep their numbers within certain specified bounds (this occurs, for example, with biological systems, in which a "homeostatic" state is defined as one in which the levels of all important substances remain within their natural range). Then one needs to compute the set consisting of all system states satisfying **this** property, rather than merely the property of viability; and system steering should be done by applying the method of Conjecture 3 to an appropriate subset of the Julia set, instead of the entire Julia set. And conjectures 2, 4 and 5 apply to these **subsets** of the Julia set as well as to the Julia set as a whole, thus giving rise to what might be labeled conjectures **2''**, **4''** and **5''**.

7.8 EVOLUTIONARY IMPLICATIONS

Having outlined what must be done in order to turn the sketchy model given above into a concrete, operational theory, I will now briefly turn to more speculative matters. The association of complex systems with hypercomplex fractals has a great number of interesting implications, but among all these, perhaps the most striking are the implications for complex system **evolution**. Hypercomplex fractal geometry has the potential to put some meat on the bones of the "post-Neo-Darwinist" theory of evolution.

In its most extreme form, the Neo-Darwinist view ofevolution assumes that an organism consists of a disjoint collection of traits. Evolution then gradually improves each trait, until each trait achieves a value which maximizes the "fitness criterion" posed by the environment. This is a very handy view which works well for many purposes; as argued in *EM*, however, it falls short on two fronts. First, it ignores ecology; that is, it ignores the substantial **feedback** between an organism and its environment. The environment of an organism is not independent of that organism, on the contrary, it is substantially shaped by that organism, and therefore the various fitness criteria of the various organisms in an ecosystem form a system of simultaneous nonlinear equations. And second, strict Neo-Darwinism ignores the complex self-organizing processes going on **inside** the organism. An organism is not a list of traits, it is an autopoietic system, and a slight change in one part of the system can substantially affect many different parts of the system. In short, strict Neo-Darwinism sets up a simple, mechanical organism against an inanimate environment, when what one really has is one complex autopoietic system, the organism, nested within another complex autopoietic system, the environment.

It is perfectly reasonable and direct to model both of these autopoietic systems -- organisms and environments -- as magician systems. On the one hand, organisms are chemical systems, and magician systems are substantially similar to Kampis's (1991) "component-systems," which were conceived as a model of biochemical reactions. And on the other hand, ecosystems are frequently modeled in terms of coupled ordinary differential equations, but any model expressed in this form can easily be reformulated as a magician system model. This is not the place to go into details, but the interested reader will find that the energetic transformations involved in a food web are easily expressed in terms of magician spells, as are such inter- organismic relations as parasitism and symbiosis.

So, suppose one has modeled organisms and environments as magician systems. Then one has two related questions:

1) How can one modify the internal makeup of an organism to as to improve its fitness relative to its environment, or, at very least, so as to change its behavior without drastically decreasing its fitness?

2) How can one modify the population structure of an environment without threatening ecological stability?

The answer to both of these questions, if the present theory is correct, is: first of all, by staying within the fractal boundary of an appropriate Julia set.

Changing the internal makeup of an organism means changing the population structure of the magician system defining that organism. Different population structures will lead to different organisms. This is vividly observable in cases where a slight change in the level of one chemical leads to a dramatic structural change; e.g. in the axolotl, which an increased thyroxin level grants the ability to breathe air. But if onechanges the organism's makeup in the wrong way, one will obtain a vector outside the relevant Julia set, and homeostasis will be destroyed; the level of some variable will shoot off to infinity. Effective evolution requires containment within the appropriate fractal boundary.

Similarly, a small change in the population of one organism can sometimes lead to ecological catastrophe. But in other situations, this is not true; whole species can be destroyed with few lasting effects on the remainder of the ecosystem. In the present view, the difference between these two cases reduces to a difference in proximity to the boundary of the relevant Julia set. When well into the interior of the Julia set, changes can be made without destroying viability. When near the boundary, however, changes must be made very delicately; many directions will maintain viability and many will not.

Organisms and ecosystems appear to be largely unpredictable. It is possible that much of this unpredictability reduces to the unpredictability of the fractal boundaries of Julia sets over abstract algebras. In Chapter 3 it was seen that, in the simple case of a two-dimensional quadratic, the genetic algorithm has the ability to generate diverse elements of the Julia set of a

mapping. It is possible that biological evolution represents a similar process carried out in a space of much higher dimension. In this view, the challenge of evolving complex systems is that of **generating diversity while staying within appropriate fractal boundaries**.

A great deal of work remains to be done before this idea can be turned into a concrete theory, capable of being compared with empirical data. Intuitively and speculatively, however, the suggestions are clear. Casting aside strict Neo-Darwinism does not mean abandoning all structure and order in evolution. It means exchanging the rigid and artificial structure of trait lists and fixed fitness criteria for the more complex, intricate and natural structure of Julia sets. It means accepting that adaptive systems are really **adaptive, autopoietic attractors** -- attractors of internal dynamical processes which give rise to intricate fractal structures and which, in some cases, rival the evolutionary dynamic itself in complexity.

CHAPTER EIGHT

THE STRUCTURE OF CONSCIOUSNESS

8.1 INTRODUCTION

Over the past few years, the writing of books on consciousness has become a minor intellectual industry. From philosophers to computer scientists, from psychologists to biologists to physicists, everyone seems to feel the need to publish their opinion on consciousness! Most of these works are reductionistic in focus: one complex mechanism after another is proposed as the essential "trick" that allows us to feel and experience ourselves and the world around us.

On the other hand, Nicholas Humphrey, in his *Natural History of the Mind*, has taken a somewhat different perspective. He argues that consciousness is something tremendously simple: that it is nothing more or less than **raw feeling**. Simply feeling something there. According to this view, consciousness exists **beneath** the level of analysis: it is almost too simple to understand. This view ties in nicely with Eastern mysticism, e.g. with the Hindu notion that at the core of the mind is the higher Self or *atman*, which is totally without form.

I like to think there is value in both of these approaches. Raw awareness is present in every state of consciousness, but different states of consciousness have different structures; and, possibly, there are some universals that tie together **all** different states of consciousness.

The mechanistically-oriented consciousness theorists are getting at the structure of states of consciousness. Some of these theorists, such as Daniel Dennett in his celebrated book *Consciousness Explained*, have made the error of assuming this **structure** to be the essence, of ignoring the subjective facts of experience. Other theorists, however, are looking at biological and computational mechanisms of consciousness structure in a clearer way, without confusing different levels of explanation. This is the category into which I would like to place myself, at least for the purposes of the present chapter. I am concerned here mainly with structures and

mechanisms, but I do not confuse these structures and mechanisms with the essence of subjective experience.

Three Questions of Consciousness

In my view, there are three questions of consciousness. The first is, what is raw feeling, raw awareness? What are its qualities? The second is, what are the **structural and dynamical properties** of different states of consciousness? And the third is: how does raw feeling interface with the structural and dynamical properties of states of consciousness? Each of these is an extremely interesting psychological question in its own right.

I will focus here on the second question, the question of structure. This is where scientific and mathematical ideas have a great deal to offer. However, I do not consider it intellectually honest to entirely shy away from the other questions. Thus, before proceeding to describe various structures of consciousness, I will roughly indicate how I feel the three questions fit together. One may, of course, appreciate may analysis of the various structures of consciousness without agreeing with my views on the nature of raw awareness, on the first and third questions.

I prefer to answer the first question by not answering it. Raw consciousness is completely unanalyzable and inexpressible. As such, I believe, it is equivalent to the **mathematically random**. Careful scrutiny of the concept of randomness reveals its **relative** nature: "random" just means "has no discernible structure with respect to some particular observer." Raw awareness, it is argued, is random in this sense. It is **beyond** our mental categories; it has no analyzable structure.

As to the third question, the relation between raw consciousness and the structure of consciousness, I prefer to give an **animist** answer. As Nick Herbert affirms in *Elemental Mind* (1988), awareness, like energy, is there in everything. From this position, one does not have to give an explanation of how certain states of matter give rise to trans-material awareness, and others do not.

However, animism in itself does not explain how some entities may have more awareness than others. The resolution to this is, in my view, provided by the idea of "consciousness as randomness." Entities will be more aware, it seems, if they open themselves up more, by nature, to the **incomprehensible**, the ineffable, the random. This observation can, as it turns out, be used to explain why certain states of consciousness seem more acutely aware than others. But I will not attempt to push the theory of consciousness this far here; this topic will be saved for elswhere.

The Structure of Consciousness

So how are states of consciousness **structured**? The most vivid states of consciousness, I will argue here, are associated with those mental system that make **other** mental systems more coherent, more robustly autopoietic. These mental systems involve extreme openness to raw awareness. This view fits in naturally with all that is known about the neuropsychology of

consciousness. Consciousness is a perceptual-cognitive loop, a feedback dynamic, that serves to coherentize, to make whole, systematic, definite.

Pushing this train of thought further, I will argue that, in particular, the most vivid states of consciousness are manifested as **time-reversible** magician systems. It is the property of reversibility, I believe, that allows these thought-systems their openness to the random force of raw awareness.

In David Bohm's language, to be introduced below, reversibility means that consciousness manifests "proprioception of thought." In terms of hypercomplex algebras, on the other hand, reversibility means that states of consciousness correspond to **division algebras**. This turns out to be a very restrictive statement, as the only reasonably symmetric finite-dimensional division algebras are the reals, the complexes, the quaternions and the octonions. The **octonions**, which contain the other three algebras as subalgebras, will be taken as the basic algebraic structure of consciousness. This abstract algebraic view of consciousness will be seen to correspond nicely with the phenomenology of consciousness. Octonionic algebras result from adjoining a reflective "inner eye" to the perceptual-cognitive loop that coherentizes objects.

As this summary should make clear, this chapter (even more than the rest of the book) is something of a potpourri of innovative and unusual ideas. No claim is made to solve the basic problem of consciousness -- which, insofar as it is a "problem," is certainly insoluble. Rather, the psynet model is used to make various interrelated forays into the realm of consciousness, centered on the question of the **structure** of conscious experience.

8.2 THE NEUROPSYCHOLOGY OF CONSCIOUSNESS

Before presenting any original ideas, it may be useful to review some relevant experimental work on the biology of consciousness. For decades neuropsychologists have shunned the word "consciousness," preferring the less controversial, more technical term "attention." But despite the methodological conservatism which this terminology reflects, there has been a great deal of excellent work on the neural foundations of conscious experience. In particular, two recent discoveries in the neuropsychology of attention stand out above all others. First is the discovery that, in Dennett's celebrated phrase, there is no "Cartesian Theater": conscious processes are **distributed** throughout the brain, not located in any single nexus. And next is the discovery of the basic **role** of this distributed consciousness: nothing esoteric or sophisticated, but simply **grouping**, **forming wholes**.

According to Rizzolati and Gallese (1988), there are two basic ways of approaching the problem of attentiveness. The first approach rests on two substantial claims:

1) that in the brain there is a selective attention center or circuit independent of sensory and motor circuits; and

2) that this circuit controls the brain as a whole.... (p. 240)

In its most basic, stripped-down form this first claim implies that there are some brain regions exclusively devoted to attention. But there are also more refined interpretations: "It may be argued ... that in various cerebral areas attentional neurons can be present, intermixed with others having sensory or motor functions. These attentional neurons may have connections among them and form in this way an attentional circuit" (p.241).

This view of attention alludes to what Dennett (1991) calls the "Cartesian Theater." It holds that there is some particular **place** at which all the information from the senses and the memory comes together into one coherent picture, and from which all commands to the motor centers ultimately emanate. Even if there is not a unique spatial location, there is at least a single unified system which acts **as if** it were all in one place.

Rizzolati and Gallassi contrast this with their own "premotor" theory of attention, of which they say:

First, it claims that ... attention is a vertical modular function present in several independent circuits and not a supramodal function controlling the whole brain. Second, it maintains that attention is a consequence of activation of premotor neurons, which in turn facilitates the sensory cells functionally related to them.

The second of these claims is somewhat controversial -- many would claim that the **sensory** rather than premotor neurons are fundamental in arousing attention. However, as Rizzolati and Gallassi point out, the evidence in favor of the **first** point is extremely convincing. For instance, Area 8 and inferior Area 2 have no reciprocal connections -- and even in their connections with the parietal lobe they are quite independent. But if one lesions either of these areas, severe attentional disorders can result, including total "neglect" of (failure to be aware of) some portion of the visual field.

There are some neurological phenomena which at first **appear** to contradict this "several independent circuits" theory of consciousness. But these apparent contradictions result from a failure to appreciate the self-organizing nature of brain function. For instance, as Rizzolatti et al (1981) have shown, although the neurons in inferior Area 6 are not responsive to emotional stimuli, nevertheless a lesion in this area can cause an animal to lose its ability to be aware of emotional stimuli. But this does not imply the existence of some brain-wide consciousness center. It can be better explained by positing an **interdependence** between Area 6 and some other areas responsive to the same environmental stimuli and **also** responsive to emotional stimuli. When one neural assembly changes, all assemblies that interact with it are prodded to change as well. Consciousness is **part** of the self-structuring process of the brain; it does not stand outside this process.

So consciousness is distributed rather than unified. But what does neuropsychology tell us about the **role** of consciousness? It tells us, to put it in a formula, that consciousness serves **to group disparate features into coherent wholes**. This conclusion has been reached by many different researchers working under many different theoretical presuppositions. There is no longer any reasonable doubt that, as Umilta (1988) has put it, "the formation of a given percept is dependent on a specific distribution of focal attention."

For instance, Treisman and Schmidt (1982) have argued for a two-stage theory of visual perception. First is the stage of elementary feature recognition, in which simple visual properties like color and shape are recognized by individual neural assemblies. Next is the stage of **feature integration**, in which consciousness focuses on a certain location and **unifies the different features** present at that location. If consciousness is not focused on a certain location, the features sensed there may combine on their own, leading to the perception of **illusory objects**.

This view ties in perfectly with what is known about the psychological consequences of various brain lesions. For instance, the phenomenon of **hemineglect** occurs primarily as a consequence of lesions to the right parietal lobe or left frontal lobe; it consists of a disinclination or inability to **be aware of** one or the other side of the body. Sometimes, however, these same lesions do not cause hemineglect, but rather **delusional perceptions**. Bisiach and Berti (1987) have explained this with the hypothesis that sometimes, when there is damage to those attentional processes connecting features with whole percepts in one side of the visual field, **the function of these processes is taken over by other, non-attentional processes**. These specific consciousness-circuits are replaced by unconscious circuits, but the unconscious circuits can't properly do the job of percept-construction; they just produce delusions. And this sort of phenomenon is not restricted to **visual** perception. Bisiach et al (1985) report a patient unable to perceive **meaningful** words spoken to him from the left side of his body -- though perfectly able to perceive **nonsense** spoken to him from the same place.

Psychological experiments have verified the same phenomenon. For instance, Kawabata (1986) has shown that one makes a choice between the two possible orientations of the Necker cube **based on the specific point on which one first focuses one's attention**. Whatever vertex is the focus of attention is perceived as **in the front**, and the interpretation of the whole image is constructed to match this assumption. Similar results have been found for a variety of different ambiguous figures -- e.g. Tsal and Kolbet (1985) used pictures that could be interpreted as either a duck or a rabbit, and pictures that could be seen as eithr a bird or a plane. In each case the point of conscious attention directed the perception of the whole. And, as is well known in such cases, once consciousness has finished forming the picture into a coherent perceived whole, **this process is very difficult to undo**.

Treisman and Schmidt's division of perception into two levels is perhaps a little more rigid than the available evidence suggests. For instance, experiments of Prinzmetal et al (1986) verify the necessity of consciousness for perceptual integration, but also point out some minor role for consciousness in enhancing the quality of perceived features. But there are many ways of explaining this kind of result. It may be that consciousness acts on more than one level: first in unifying sub-features into features, then in unifying features into whole objects. Or it may be that perception of the whole **causes** perception of the features to be improved, by a sort of feedback process.

8.3 THE PERCEPTUAL-COGNITIVE LOOP

In this section I will abstract the neuropsychological ideas discussed above into a more generalized, mathematical theory, which I call the theory of the Perceptual-Cognitive Loop. This

is not a complete theory of the structure of consciousness -- it will be built on in later sections. But it is a start.

Edelman (1990) has proposed that consciousness consists of a **feedback loop** from the perceptual regions of the brain to the "higher" cognitive regions. In other words, consciousness is a process which cycles information from perception to cognition, to perception, to cognition, and so forth (in the process continually creating **new** information to be cycled around).

Taking this view, one might suppose that the brain lesions discussed above hinder consciousness, not by destroying an entire autonomously conscious neural assembly, but by destroying the **perceptual** end of a larger consciousness-producing loop, a perceptual-cognitive loop or **PCL**. But the question is: why have a loop at all? Are the perceptual processes themselves incapable of grouping features into wholes; do they need cognitive assistance? Do they need the activity of premotor neurons, of an "active" side?

The cognitive end of the loop, I suggest, serves largely as a **tester** and **controller**. The perceptual end does some primitive grouping procedures, and then passes its results along to the cognitive end, asking for approval: "Did I group **too little**, or **enough**?" The cognitive end seeks to integrate the results of the perceptual end with its **knowledge and memory**, and on this basis gives an answer. In short, it **acts** on the percepts, by trying to do things with them, by trying to use them to interface with memory and motor systems. It gives the answer "too little coherentization" if the proposed grouping is simply torn apart by contact with memory -- if different parts of the supposedly coherent percept connect with totally different remembered percepts, whereas the whole connects significantly with nothing. And when the perceptual end receives the answer "too little," it goes ahead and tries to group things together **even more**, to make things even more coherent. Then it presents its work to the cognitive end again. Eventually the cognitive end of the loop answers: "Enough!" Then one has an entity which is sufficiently coherent to withstand the onslaughts of memory.

Next, there is another likely aspect to the perceptual-cognitive interaction: perhaps the cognitive end also **assists in the coherentizing process**. Perhaps it proposes ideas for interpretations of the whole, which the perceptual end then approves or disapproves based on its access to more primitive features. This function is not in any way contradictory to the idea of the cognitive end as a tester and controller; indeed the two directions of control fit in quite nicely together.

Note that a **maximally coherent** percept is not desirable, because thought, perception and memory require that ideas possess some degree of flexibility. The individual features of a percept should be detectable to some degree, otherwise how could the percept be related to other similar ones? The trick is to stop the coherence-making process **just in time**.

But what exactly is "just in time"? There is not necessarily a unique optimal level of coherence. It seems more likely that each consciousness-producing loop has its own characteristic level of cohesion. Hartmann (1991) has proposed a theory which may be relevant to this issue: he has argued that each **person** has a certain characteristic "boundary thickness" which they place between the different ideas in their mind. Based on several questionnaire and interview studies, he has shown that this is a statistically significant method for classifying personalities. "Thin-

boundaried" people tend to be sensitive, spiritual and artistic; they tend to blend different ideas together and to perceive a very thin layer separating themselves from the world. "Thick-boundaried" people, on the other hand, tend to be practical and not so sensitive; their minds tend to be more compartmentalized, and they tend to see themselves as very separate from the world around them. Hartmann gives a speculative account of the neural basis of this distinction. But the present theory of consciousness suggests an alternate account: that perhaps this distinction is **biologically** based on a difference in the "minimum cohesion level" accepted by the cognitive end of consciousness-producing loops.

The iterative processing of information by the perceptual-cognitive loop is what enables the same object to be disrupted by randomness **again and again and again**. And **this**, I claim, is what gives the feeling that one is conscious of some specific object. Without this iteration, consciousness is felt to lack a definite object; the object in question lasts for so short a time that it is just barely noticeable.

And what of the old aphorism, "Consciousness is consciousness of consciousness"? This reflexive property of consciousness may be understood as a consequence of the passage of potential coherentizations from the cognitive end to the perceptual end of the loop. The cognitive end is trying to understand what the perceptual end is doing; it is **recognizing patterns** in the series of proposed coherentizations and ensuing memory-caused randomizations. These higher-order patterns are then sent through the consciousness-producing loop as well, in the form of new instructions for coherentization. Thus the process that produces consciousness itself becomes transformed into an object of consciousness.

All this ties in quite nicely with the **neural network theory of consciousness** proposed by the British mathematician R. Taylor (1993). Taylor proposes that the consciousness caused by a given stimulus can be equated with the memory traces elicited by that stimulus. E.g. the consciousness of a sunset is the combination of the faint memory traces of previously viewed sunsets, or previously viewed scenes which looked like sunsets, etc. The PCL provides an analysis on a level one deeper than Taylor's theory; in other words, Taylor's theory is a **consequence** of the one presented here. For, if the PCL works as I have described, it follows that the cognitive end must always search for memory traces similar to the "stimulus" passed to it by the perceptual end.

What is Coherentization?

There is a missing link in the above account of the PCL: what, exactly, **is** this mysterious process of "coherentization," of boundary-drawing? Is it something completely separate from the ordinary dynamics of the mind? Or is it, on the other hand, an **extension** of these dynamics?

The chemistry involved is still a question mark, so the only hope of understanding coherentization at the present time is to bypass neuropsychology and try to analyze it from a general, philosophical perspective. One may set up a **model** of "whole objects" and "whole concepts," and ask: in the context of this model, what is the structure and dynamics of coherentization? In the Peircean perspective, both **objects** and **ideas** may be viewed as "habits" or "patterns," which are related to one another by **other** patterns, and which have the capacity to

act on and transform one another. We then have the question of how a pattern can be coherentized, how its "component patterns" can be drawn more tightly together to form a more cohesive whole.

Here we may turn to the psynet model, and propose that: To coherentize is to make something autopoietic, or more robustly autopoietic. What consciousness does, when it coherentizes, is to **make autopoietic systems**. It makes things more self-producing.

From a neural point of view, one may say that those percepts which are most likely to survive in the evolving pool of neural maps, are those which receive the most external stimulation, and those which **perpetuate themselves** the best. External stimulation is difficult to predict, but the tendency toward self-perpetuation can be built in; and this is the most natural meaning for coherentization.

In this view, then, what the perceptual-cognitive loop does is to take a network of processes and **iteratively make it more robustly self-producing**. What I mean by "robustly self-producing" is: autopoietic, with a wide basin as an attractor of the cognitive equation (of magician dynamics). Any mathematical function can be reproduced by a great number of magician systems; some of these maps will be robustly self-producing and others will not. The trick is to find these most robustly self-producing systems. There are many possible strategies for doing this kind of search -- but one may be certain that the **brain**, if it does this kind of search, certainly does not use any fancy mathematical algorithm. It must proceed by a process of guided trial-and-error, and thus it must require constant testing to determine the basin size, and the degree of autopoiesis, of the current iterate. The consciousness, in this model, is in the **testing**, which disrupts the overly poor self-production of interim networks (the end result of the iterative process is something which leads to fairly little consciousness, because it is relatively secure against disruption by outside forces).

So, "coherentization" is not a catch-word devoid of content; it is a concrete process, which can be understood as a peculiar chemical process, or else modeled in a system-theoretic way. Thinking about neuronal groups yields a particularly elegant way of modeling coherentization: as the search for a large-basined **autopoietic subsystem** of magician dynamics. This gives a very concrete way of thinking about the coherentization of a complex **pattern**. To coherentize a pattern which is itself a system of simpler patterns, emerging cooperatively from each other, one must replace the component patterns with others that, while expressing largely the same regularities, emerge from each other in an even more tightly interlinked way.

8.4 SUBVERTING THE PERCEPTUAL-COGNITIVE LOOP

The perceptual-cognitive loop is important and useful -- but it does not go far enough. It explains how we become attentive to things; or, to put it differently, how we **construct** "things" by carrying out the process of conscious attention. But as humans we can do much more with our consciousness than just be attentive to things. We can introspect -- consciously monitor our own thought processes. We can meditate -- consciously fixate our consciousness on **nothing** whatsoever. We can creatively focus -- fix our consciousness on abstract ideas, forming them

into wholes just as readily as we construct "physical objects." How do these subtle mental conditions arise from the "reductionistic" simplicity of the perceptual-cognitive loop?

Sketch of a Theory of Meditation

Let us begin with meditation -- in particular, the kind of meditation which involves emptying the mind of forms. This type of meditation might be called "consciousness without an object." In Zen Buddhism it is called zazen.

The very indescribability of the meditative state has become a cliche'. The Zen Buddhist literature, in particular, is full of anecdotes regarding the futility of trying to **understand** the "enlightened" state of mind. Huang Po, a Zen master of the ninth century A.D., framed the matter quite clearly:

Q: How, then, does a man accomplish this comprehension of his own Mind?

A: That which asked the question IS your own Mind; but if you were to remain quiescent and to refrain from the smallest mental activity, its substance would be seen as a void -- you would find it formless, occupying no point in space and falling neither into the category of existence nor into that of non-existence. Because it is imperceptible, Bodhidharma said: 'Mind, which is our real nature, is the unbegotten and indestructible Womb; in response to circumstances, it transforms itself into phenomena. For the sake of convenience, we speak of Mind as intelligence, but when it does not respond to circumstances, it cannot be spoken of in such dualistic terms as existence or nonexistence. Besides, even when engaged in creating objects in response to causality, it is still imperceptible. If you know this and rest tranquilly in nothingness -- then you are indeed following the Way of the Buddhas. Therefore does the sutra say: 'Develop a mind which rests on no thing whatever.'

The present theory of consciousness suggests a novel analysis of this state of mind that "rests on no thing whatever." Consider: the perceptual-cognitive loop, if it works as I have conjectured, must have **evolved** for the purpose of making percepts cohesive. The consciousness of objects is a corollary, a spin-off of this process. Consciousness, raw consciousness, was there all along, but it was not **intensively focused** on one thing. Meditative experience relies on **subverting** the PCL away from its evolutionarily proper purpose. It takes the intensity of consciousness derived from repeated iteration, and removes this intensity from its intended context, thus producing an entirely different effect.

This explains why it is so **difficult** to achieve consciousness without an object. Our system is wired for consciousness **with** an object. To regularly attain consciousness without an object requires the formation of new neural pathways. Specifically, I suggest, it requires the development of pathways which feed the perceptual end of the perceptual-cognitive loop **random stimuli** (i.e., stimuli that have no cognitively perceivable structure). Then the perceptual end will send messages to the cognitive end, as if it were receiving structured stimuli -- **even though it is not receiving any structured stimuli**. The cognitive then tries to integrate the random message into the associative memory -- but it fails, and thus the perceptual end makes a new presentation. And so on, and so on. What happens to the novice meditator is that thoughts

from the associative memory continually get in the way. The cognitive end makes **suggestions** regarding how to coherentize the random input that it is receiving, and then these suggestions cycle around the loop, destroying the experience of emptiness. Of course these suggestions are mostly nonsense, since there is no information there to coherentize; but the impulse to make suggestions is quite strong and can be difficult to suppress. The cognitive end must be trained not to make suggestions regarding random input, just as the perceptual end must be trained to accept random input from sources other than the normal sensory channels.

This is not an attempt to explain away mystical experience -- quite the opposite. It is an attempt to acknowledge the ineffable, ungraspable nature of such experience. As argued in detail in *The Structure of Intelligence*, there is no objective notion of "randomness" for finite structures like minds and brains. Random is defined only relative to a certain observer (represented in computation theory as a certain Universal Turing Machine). So, to say that the meditative state involves tapping into randomness, is to say that the meditative state involves tapping into some source that is **beyond one's own cognitive structures**. Whether this source is quantum noise, thermal noise or the divine presence is an interesting question, but one that is not relevant to the present theory, and is probably not resolvable by rational means.

Creative Inspiration

Finally, let us peek ahead to the final chapter for a moment, and have a look at the role of the perceptual-cognitive loop in the process of **creative inspiration**.

As will be observed in detail later, many highly creative thinkers and artists have described the role of consciousness in their work as being very small. The biggest insights, they have claimed, always pop into the consciousness **whole**, with no deliberation or decision process whatsoever -- all the work has been done elsewhere. But yet, of course, these sudden insights always concern some topic that, at some point in the past, the person in question **has** consciously thought about. A person with no musical experience whatsoever will never all of a sudden have an original, fully detailed, properly constructed symphony pop into her head. Someone who has never thought about physics will not wake up in the middle of the night with a brilliant idea about how to construct a unified field theory. Clearly there is something more than "divine inspiration" going on here. The question is: what is the dynamics of this subtle interaction between consciousness and the unconscious?

In the present theory of consciousness, there is no rigid barrier between consciousness and the unconscious; everything has a certain degree of consciousness. But only in the context of an iterative loop does a **single object** become fixed in consciousness long enough that raw consciousness becomes comprehensible as **consciousness of that object**. The term "unconscious" may thus be taken to refer to those parts of the brain that are not directly involved in a consciousness-fixing perceptual/cognitive loop.

This idea has deep meaning for human creative process. In any creative endeavor, be it literature, philosophy, mathematics or science, one must struggle with forms and ideas, until one's mind becomes **at home** among them; or in other words, until one's consciousness is able to **perceive them as unified wholes**. Once one's consciousness has perceived an idea as a coherent whole --

then one need no longer consciously mull over that idea. The idea is **strong** enough to withstand the recombinatory, self-organizing dynamics of the unconscious. And it is up to these dynamics to produce the **fragments** of new insights -- fragments which consciousness, once again entering the picture, may unify into **new wholes**.

So: without the perceptual-cognitive loop to aid it, the unconscious would not be significantly creative. It would most likely recombine all its contents into a tremendous, homogeneously chaotic mush ... or a few "islands" of mush, separated by "dissociative" gaps. But the perceptual-cognitive loop makes things **coherent**; it places **restrictions** on the natural tendency of the unconscious to combine and synthesize. Thus the unconscious is posed the more difficult problem of relating things with one another **in a manner compatible with their structural constraints**. The perceptual-cognitive loop produces wholes; the unconscious manipulates these wholes to produce new fragmentary constructions, new collections of patterns. And then the perceptual-cognitive loop takes these new patterns as raw material for constructing new wholes.

And what, then, is the relation between the **creative** state and the **meditative** state? Instead of a fixation on the void of pure randomness, the creative condition is a fixation of consciousness on certain **abstract forms**. The secret of the creative artist or scientist, I propose, is this: abstract forms are perceived with the reality normally reserved for sense data. Abstract forms are **coherentized** with the same vigor and effectiveness with which everyday **visual** or **aural** forms are coherentized in the ordinary human mind. Like the meditative state, the creative state subverts the perceptual-cognitive loop; it uses it in a manner quite different than that for which evolution intended it.

One may interpret this conclusion in a more philosophical way, by observing that the role of the perceptual-cognitive loop is, in essence, **to create reality**. The reality created is a mere "subjective reality," but for the present purposes, the question of whether there is any more **objective** reality "out there" is irrelevant. The key point is that the very **realness** of the subjective world experienced by a mind is a consequence of the perceptual-cognitive loop and its construction of **boundaries** around entities. This means that reality depends on consciousness in a fairly direct way: and, further, it suggests that what the creative subself accomplishes is to **make abstract forms and ideas a concrete reality**.

8.5 THE EVOLUTION OF THE PERCEPTUAL-COGNITIVE LOOP

Julian Jaynes, in *The Breakdown of the Bicameral Mind*, has argued that consciousness evolved suddenly rather than rapidly, and that this sudden evolution occurred in the very recent past. He believes that the humans of Homeric times were not truly conscious in the sense that we are. His argument is based primarily on literary evidence: the characters in the *Odyssey* and other writings of the time never speak of an "inner voice" of consciousness. Instead they refer continually to the **voices of the gods**. Jaynes proposes that this "hearing of voices," today associated with schizophrenia, was in fact the root of modern consciousness. Eventually the voice was no longer perceived as a voice, but as a more abstract inner guiding force, in other words "consciousness."

Jaynes' theory is admirable in its elegance and boldness; unfortunately, however, it makes very little scientific sense. Inferring the history of mind from the history of literary style is risky, to say the least; and Jaynes' understanding of schizophrenia does not really fit with what we know today. But despite the insufficiency of his arguments, I believe there is a kernel of truth in Jayne's ideas. In this section I will use the theory of the perceptual-cognitive loop to argue that the idea of a **sudden appearance** of modern consciousness is quite correct, though for very different reasons than those which Jaynes put forth.

The perceptual-cognitive loop relies on two abilities: the perceptual ability to recognize elementary "features" in sense data, and the cognitive ability to link conjectural "wholes" with items in memory. A sudden jump in either one of these abilities could therefore lead to a sudden jump in consciousness. In *EM* I argue that the **memory**, at some point in early human history, underwent a sudden structural "phase transition." I suggest that this transition, if it really occurred, would have caused as a corollary effect a sudden increase in the intensity of **consciousness**.

The argument for a phase transition in the evolution of memory rests on the idea of the **heterarchical subnetwork** of the psynet. From the view of the heterarchical network, mind is an associative memory, with connections determined by habituation. So, suppose that, taking this view, one takes N items stored in some organism's memory, and considers two items to be "connected" if the organism's mind has detected pragmatically meaningful relations between them. Then, if the memory is sufficiently complex, one may study it in an approximate way by assuming that these connections are drawn "at random." Random graph theory becomes relevant.

Recall the theory of random graph thresholds, introduced in Chapter Seven. The crucial question, from the random graph theory point of view, is: what is the **chance** that, given two memory items A and B, there is a connection between A and B?

For instance, if this chance exceeds the value 1/2, then the memory is almost surely a "nearly connected graph," in the sense that one can follow a chain of associations from almost any memory item to almost any other memory item. On the other hand, if this chance is less than 1/2, then the memory is almost certainly a "nearly disconnected graph": following a chain of associations from any one memory item will generally lead only to a small subset of "nearby" memory items. There is a "phase transition" as the connection probability passes 1/2. And this is merely one among many interesting phase transitions.

The evolutionary hypothesis, then, is this. Gradually, the brain became a better and better pattern recognition machine; and as this happened the memory network became more and more densely connected. In turn, the more effective memory became, the more useful it was as a guide for pattern recognition. Then, all of a sudden, pattern recognition became useful enough that it gave rise to a memory **past the phase transition**. Now the memory was **really** useful for pattern recognition: pattern recognition processes were able to search efficiently through the memory, moving from one item to the next to the next along a path of gradually increasing relevance to the given object of study. The drastically increased pattern recognition ability filled the memory in even more -- and all of a sudden, the mind was operating on a whole new level.

And one consequence of this "new level" of functioning may have been -- an effective perceptual-cognitive loop. In a mind **without** a highly active associative memory, there is not so much **need** for a PCL: coherentization is a protection against reorganizing processes which are largely irrelevant to a pre-threshold memory network. In a complex, highly interconnected memory, reorganization is necessary to improve associativity, but in a memory with very few connections, there is unlikely to **be** any way of significantly improving associativity. Furthermore, even if the pre-threshold memory **did** have need of a PCL, it would not have the **ability** to run the loop through many iterations: this requires each proposed coherentization to be "tested" with numerous different connections. But if few connections are there in the first place, this will only very rarely be possible.

So, in sum, in order for the **cognitive** end of the loop to work properly, one needs a quality associative memory. A phase transition in associative memory paves the way for a sudden emergence of consciousness.

In this way one arrives at a plausible scientific account of the sudden emergence of consciousness. It is a speculative account, to be sure -- but unlike Jaynes' account it relies on precise **models** of what is going on inside the brain, and thus it is falsifiable, in the sense that it could be disproved by appropriately constructed computer simulations. **Consciousness of objects** emerging out of raw consciousness as a consequence of phase transitions in the associative memory network -- certainly, this picture has simplicity and elegance on its side.

8.6 HALLUCINATIONS AND REALITY DISCRIMINATION

The theory of the Perceptual-Cognitive Loop may seem excessively abstract and philosophical. What does it have to say about the concrete questions that concern clinical or experimental psychologists? To show that it is in fact quite relevant to these issues, I will here consider an example of a psychological phenomenon on which the psynet model and PCL shed light: hallucinations.

Hallucinations have proved almost as vexing to psychological theorists as they are to the individuals who experience them. A variety of theories have been proposed, some psychological, some neurobiological, and some combining psychological and physiological factors (for reviews see Slade and Bentall, 1988; Slade, 1994). Jaynes, mentioned above, believed hallucinations to be the evolutionary root of consciousness! However, none of the existing theories has proved entirely satisfactory.

Perhaps the most precise and comprehensive theory to date is that of Slade and Bentall (1988; Slade, 1994), which holds that hallucinations are due to a lack of skill at "reality-testing" or reality discrimination. According to this theory, individuals prone to hallucinate do not generate internal stimuli differently from the rest of us; they merely interpret these stimuli differently.

The Slade and Bentall theory provides a good qualitative understanding of a variety of clinical observations and experimental results regarding hallucinatory experience. However, the direct evidence for the theory is relatively scant. There are experiments which show that individuals with high Launay-Slade Hallucination scores have a bias toward classifying signals as real,

instead of imagined (Slade and Bentall, 1988; Feelgood and Rantsen, 1994). But these experiments only indicates a **correlation** between poor reality discrimination and hallucination. They do not establish a causal relationship, which is what Slade (1994) and Slade and Benthall (1988) posit.

Taking up the same basic conceptual framework as Slade and Benthall, I will argue that, in fact, poor reality discrimination and hallucination **produce one another**, given their initial impetus by the cognitive trait that Hartmann (1991) calls "thin-boundariedness." In short, certain individuals place particularly permeable boundaries around entities in their minds, including their self-systems. This is a result of individual differences in the nature of consciousness-embodying Perceptual-Cognitive Loops. The tendency to construct permeable boundaries, it is argued, encourages both hallucination and poor reality discrimination, which in turn are involved in positive feedback relationships with each other.

These ideas are an alternate and, I suggest, more psychologically plausible explanation of the data which Slade and Bentall take in support of their theory of hallucinations. They are a simple, nontechnical example of how the psynet model encourages one to take a longer, deeper look at the mind than conventional psychological theory requires.

Six Possible Explanations

Before embarking on our analysis of hallucination, some methodological clarifications may be in order. It is well-known that "correlation does not imply causation." However, attention is rarely drawn to the full realm of possible explanations for a correlation between two variables. Given a correlation between A and B, there are at least six distinct possible explanations, all of which must be taken seriously.

The three commonly-recognized explanations of a correlation between A and B are as follows:

1) A may cause B

2) B may cause A

3) A and B may both be caused by some other factor C.

The three additional explanations ensue from dynamical systems theory (Abraham and Shaw, 1991; Goertzel, 1994), which indicates that two variables may in fact "cause each other" due to positive feedback. They are as follows:

4) A and B may cause each other (with no initial prompting

other than random fluctations)

5) A and B may cause each other, once initially activated

by some other factor C

6) A and B may cause each other, in a process of mutual

feedback with some other factor C

These possibilities are all observed in connectionist models on a routine basis (Rumelhart and McClelland, 1986), and in fact have a long history in Oriental psychology, going back at least to the notion of "dependent causation" in Buddhist psychology (Crook and Rabgyas, 1988).

Let us now consider all six of these possibilities in the context of reality discrimination and hallucination.

First, there is the possibility that unreliable reality discrimination causes hallucinations; i.e. that, as has been claimed, "hallucinations result from a dramatic failure of the skill of reality discrimination" (Slade and Benthall, 1988).

Then there is the possibility that things work the other way around: that hallucinations **cause** unreliable reality discrimination. This hypothesis is also quite plausible. For, consider reality discrimination as a categorization problem. One might reason as follows. In ordinary experience, there are substantial differences between internally generated-stimuli and externally-generated stimuli. Thus, it is easy to "cluster" stimuli into two categories: real versus imaginary. But, in the experience of a person prone to hallucinations, there is more of a continuum from internally-generated to externally-generated stimuli. The two categories are not so distinct, and thus categorization is not such an easy problem. Quite naturally, for such individuals, skill at distinguishing the two categories will be below par.

Third, there is the possibility that both hallucinations and unreliable reality discrimination are caused by some other factor, or some other group of factors. This of course raises the question of what these other factor(s) might be.

Fourth, there is the possibility that hallucinations and unreliable reality discrimination **cause each other**. In other words, the two may stand in a positive feedback relation to each other: the more one hallucinates, the worse one does reality discrimination; and the worse one does reality discrimination, the more one hallucinates. This possibility is inclusive of the first two possibilities.

Finally, there are combinations of the third and fourth options. The fifth possibility is that an external factor prompts a slight initial propensity toward poor reality discrimination and hallucination, which then blossoms, by positive feedback, into more prominent phenomena. And the sixth possibility is that poor reality discrimination and hallucinations to emerge cooperatively, by positive feedback, in conjunction with some other factor.

Here I will argue for the fifth possibility. My argument is perhaps somewhat speculative, but no more so than the arguments of Slade (1994) and Slade and Benthall (1988). It represents a much more natural interpretation of the data which they adduce in favor of their theory.

Consciousness Creates Reality

The first key conceptual point to remember, in discussing hallucinations, is that even in the normal (non-hallucinating) individual, the role of consciousness is to create percepts and concepts from stimuli, or in other words, **to create subjective reality.** And the second key point is that **not everyone creates reality in the same way**; not only do the contents of subjective reality differ from one person to the other, but the internal parameters of the reality-creating process also differ. Hallucinations are a matter of confusing internally generated stimuli for external reality.

The relation between hallucinations and consciousness has been discussed in the past. In particular, Frith (1979) has used the limited channel capacity of consciousness to explain the occurence of hallucinations. Because of the limited capacity of consciousness, it is argued, only one hypothesis regarding the nature of a given stimulus can be consciously entertained at a given time. Hallucinations occur when preconscious information about a stimulus is not filtered out, so that consciousness becomes crowded with information, and the correct hypothesis (that the stimulus is internally rather than externally generated) is pushed out. It is worth emphasizing that the ideas presented here are quite different from these. Here we are arguing that individual differences in conscious information processing are crucial for hallucination. Frith, on the other hand, posits that the nature of conscious processing is invariant with respect to the propensity toward hallucination, while the type of information fed into consciousness varies.

How do individuals prone to hallucination differ in their creation of subjective reality? According to the PCL approach, the key stage is judgement of "enough." The PCL coherentizes thought-and-percept-systems, and it keeps going until they are coherent enough -- but what is "enough"? A **maximally coherent** percept is not desirable, because thought, perception and memory require that ideas possess some degree of flexibility. The individual features of a percept should be detectable to some degree, otherwise how could the percept be related to other similar ones? The trick is to stop the coherence-making process at just the right time. However, it should not be assumed that there is a unique optimal level of coherence. It seems more likely that each consciousness-producing loop has its own characteristic level of cohesion.

And this is where the ideas of Hartmann (1991) become relevant. Hartmann has argued that each person has a certain characteristic "boundary thickness" which they place between the different ideas in their mind. Thick-boundaried people perceive a rigid division between themselves and the external world, and they create percepts and concepts which hold together very tightly. Thin-boundaried people, on the other hand, perceive a more permeable boundary between themselves and the external world, and create more flexible percepts and concepts. Thin-boundaried people are more likely to have hallucinations, and also more likely to have poor reality discrimination (Hartmann, 1988). Thin-boundariedness, I suggest, is the third major factor involved in the dynamics of reality discrimination and hallucination.

Based on several questionnaire and interview studies, Hartmann has shown that boundary thickness is a statistically significant method for classifying personalities. "Thin-boundaried" people tend to be sensitive, spiritual and artistic; they tend to blend different ideas together and to perceive a very thin layer separating themselves from the world. "Thick-boundaried" people, on the other hand, tend to be practical and not so sensitive; their minds tend to be more compartmentalized, and they tend to see themselves as very separate from the world around

them. The difference between thin and thick-boundaried personalities, I suggest, lies in the "minimum cohesion level" accepted by the cognitive ends of PCL's. Hartmann himself gives a speculative account of the neural basis of this distinction, which is quite consistent with these ideas.

A Theory of Hallucination

Now we are ready to tie together the various threads that have been introduced, into a coherent theory of the interrelation between reality discrimination, hallucination, and boundary-drawing. I will argue that thin-boundariedness provides the initial impetus for both poor reality-discrimination and hallucination, which then go on to support and produce each other via a relation of positive feedback (Goertzel, 1994).

First, consider the thin-boundariedness for reality discrimination. As compared to an average person, an exceptionally thin-boundaried individual will have PCL's that place less rigid, more flexible boundaries around mental entities. This implies that external, real-world information will be stored and conceptualized more similarly to internal, imagined information. Such a person will naturally have more difficulty distinguishing real stimuli from non-real stimuli -- a conclusion which is supported by Hartmann's data. Thin-boundariedness is bound up with poor reality discrimination in a direct way.

Next, to see the direct relationship between thin-boundariedness and hallucination, it is necessary to consider the **iterative** nature of PCL's. The key point is that a PCL does not merely classify data, it constructs data. It is responsible for developing stimuli into percepts in particular ways. Thus, if it has judged a certain stimulus to be "real," it will develop it one way; but if has judged the stimulus to be "imaginary," it will develop it another way. In a thin-boundaried person, there is less likely to be a rigid distinction between different ways of developing stimuil into percepts; so it is much more likely, that as a perceptual-cognitive loop iterates, internal stimuli will be developed **as if** they were external stimuli. What this means is that thin-boundaries people will be more likely to have internal stimuli that are as vividly and intricately developed as external stimuli -- i.e., as suggested by Hartmann's data, they will be more likely to hallucinate.

So, thin-boundariedness has the ability to lead to both hallucination and poor reality discrimination. We have already suggested, however, that the latter two traits have the propensity to support each other by positive feedback. Poor reality discrimination causes hallucination, according to the mechanisms identified by Slade and Bentall. And hallucinations may cause poor reality discrimination, by giving the mind a more confusing data set on which to base its categorization of internal versus external stimuli.

The view which emerges from these considerations is consistent with the fifth possible relationship listed in Section 2: "A and B may cause each other, once initially activated by some other factor C." Thin-boundariedness sets off a process of mutual activation between the traits of poor reality discrimination and hallucination.

One might wonder whether the sixth relationship is not the correct one. Perhaps thin-boundariedness is itself produced, in part, by hallucinations or poor reality discrimination. But

this conclusion does not seem to hold up. There is no clear causative link between hallucination and thin-boundariedness. And the polarity of the influence of poor reality-discrimination on boundary-drawing seems quite open. In some individuals, poor reality discrimination might cause excessive thin-boundariedness; but in others, it might instead cause excessive thick-boundariedness. In yet others, it might do neither. The question is whether an individual tends to err on the side of making external stimuli "loose" or making internal stimuli "tight" -- or errs in a non-biased, neutral way.

Conclusion

I have used the PCL and the notion of mental-process intercreation to propose a new explanation of the correlation between hallucination and reality discrimination. My explanation differs from that of Slade and Bentall (1988) in that it posits a circular causality between the two factors, initiated by a third factor of thin-boundariedness. While the Slade and Bentall theory is simpler, the present theory is psychologically more plausible, and does greater justice to the complexity of the human mind.

The question arises whether the present theory is empirically distinguishable from the Slade and Bentall theory. Such a distinction, it seems, could be made only by a study of the **development** of hallucinations in individuals prone to hallucinate, say schizophrenic individuals. The Slade and Bentall view, in which poor reality discrimination causes hallucination, would predict that poor reality discrimination should precede hallucinations, and should not increase proportionately to hallucinations. The present theory predicts that the two factors should increase together, gradually, over the course of development.

8.7 PROPRIOCEPTION OF THOUGHT

Now let us return to more fundamental issues -- to the qeustion of the nature of consciousness itself. Raw awareness, I have said, is essentially random. The structure of consciousness, which exploits raw awareness, is that of an iteratively coherentizing perceptual-cognitive loop. Is there any more that can be said?

I believe that there is. One can make very specific statements about the types of **magician systems** involved in states of consciousness. Before we can get to this point, however, some preliminary ideas must be introduced. Toward this end, we will draw inspiration from a somewhat unlikely-sounding direction: the philosophical thought of the quantum physicist David Bohm, as expressed (among other places) in his book *Thought as a System* (1988).

Bohm views thought as a system of reflexes - - habits, patterns - - acquired from interacting with the world and analyzing the world. He understands the self- reinforcing, self- producing nature of this system of reflexes. And he diagnoses our thought- systems as being infected by a certain malady, which he calls the absence of proprioception of thought.

Proprioceptors are the nerve cells by which the body determines what it is doing - - by which the mind knows what the body is doing. To understand the limits of your proprioceptors, stand up on the ball of one foot, stretch your arms out to your sides, and close your eyes. How long can you

retain your balance? Your balance depends on proprioception, on awareness of what you are doing. Eventually the uncertainty builds up and you fall down. People with damage to their proprioceptive system can't stay up as long as as the rest of us.

According to Bohm,

... [T]hought is a movement - - every reflex is a movement really. It moves from one thing to another. It may move the body or the chemistry or just simply the image or something else. So when 'A' happens 'B' follows. It's a movement. All these reflexes are interconnected in one system, and the suggestion is that they are not in fact all that different. The intellectual part of thought is more subtle, but actually all the reflexes are basically similar in structure. Hence, we should think of thought as a part of the bodily movement, at least explore that possibility, because our culture has led us to believe that thought and bodily movement are really two totally different spheres which are no basically connected. But maybe they are not different. The evidence is that thought is intimately connected with the whole system. If we say that thought is a reflex like any other muscular reflex - - just a lot more subtle and more complex and changeable - - then we ought to be able to be proprioceptive with thought. Thought should be able to perceive its own movement. In the process of thought there should be awareness of that movement, of the intention to think and of the result which that thinking produces. By being more attentive, we can be aware of how thought produces a result outside ourselves. And then maybe we could also be attentive to the results it produces within ourselves. Perhaps we could even be immediately aware of how it affects perception. It has to be immediate, or else we will never get it clear. If you took time to be aware of this, you would be bringing in the reflexes again. So is such proprioception possible? I'm raising the question....

The basic idea here is quite simple. If we had proprioception of thought, we could feel what the mind was doing, at all times -- just as we feel what the body is doing. Our body doesn't generally carry out acts on the sly, without our observation, understanding and approval. But our mind (our brain) continually does exactly this. Bohm traces back all the problems of the human psyche and the human world -- warfare, environmental destruction, neurosis, psychosis -- to this one source: the absence of proprioception of thought. For, he argues, if we were really aware of what we were doing, if we could fully **feel** and **experience** everything we were doing, we would not do these self-destructive things.

An alternate view of this same idea is given by the Zen master Thich Nhat Hanh (1985), who speaks not of proprioception but of "mindfulness." Mindfulness means being aware of what one is doing, what one is thinking, what one is feeling. Thich Nhat Hanh goes into more detail about what prevents us from being mindful all the time. In this connection he talks about samyojama - - a Sanskrit word that means "internal formations, fetters, or knots." In modern terminology, samyojama are nothing other than self- supporting thought- systems:

When someone says something unkind to us, for example, if we do not understand why he said it and we become irritated, a knot will be tied in us. The lack of understanding is the basis for every internal knot. If we practice mindfulness, we can learn the skill of recognizing a knot the moment it is tied in us and finding ways to untie it. Internal formations need our full attention as soon as they form, while they are still loosely tied, so that the work of untying them will be easy.

Self- supporting thought systems, systems of emotional reflexes, guide our behaviors in all sorts of ways. Thich Nhat Hanh deals with many specific examples, from warfare to marital strife. In all cases, he suggests, simple sustained awareness of one's own actions and thought processes - - simple mindfulness - - will "untie the knots," and free one from the bundled, self- supporting systems of thought/feeling/behavior.

Yet nother formulation of the same basic concept is given by psychologist Stanislaw Grof (1994). Grof speaks, not of knots, but rather of "COEX systems" - - systems of compressed experience. A COEX system is a collection of memories and fantasies, from different times and places, bound together by the self- supporting process dynamics of the mind. Elements of a COEX system are often joined by similar physical elements, or at least similar emotional themes. An activated COEX system determines a specific mode of perceiving and acting in the world. A COEX system is an attractor of mental process dynamics, a self- supporting subnetwork of the mental process network, and, in Buddhist terms, a samyojama or knot. Grof has explored various radical techniques, including LSD therapy and breathwork therapy, to untie these knots, to weaken the grip of these COEX systems. The therapist is there to assist the patient's mental processes, previously involved in the negative COEX system, in reorganizing themselves into a new and more productive configuration.

Reflexes and Magicians

Before going further along this line, we must stop to ask: What does this notion of "proprioception of thought," based on a neo-behaviourist, reflex-oriented view of mind, have to do with the psynet model? To clarify the connection, we must first establish the connection between reflexes and magicians. The key idea here is that, in the most general sense, a **habit** is nothing other than a **pattern**. When a "reflex arc" is established in the brain, by modification of synaptic strengths or some other method, what is happening is that this part of the brain is recognizing a pattern in its environment (either in the other parts of the brain to which it is connected, or in the sensory inputs to which it is connected).

A reflex, in the psynet model, may be modelled as the interaction of three magicians: one for perception, one for action, and one for the "thought" (i.e. for the internal connection between perception and action). The "thought" magician must learn to recognize patterns among stimuli presented at different times and generate appropriate responses.

This view of reflexes is somewhat reminiscent of the triangular diagrams introduced by Gregory Bateson in his posthumous book *Angels Fear* (1989). I call these diagrams "learning triads." They are a simple and general tool for thinking about complex, adaptive systems. In essence, they are a system-theoretic model of the reflex arc.

Bateson envisions the fundamental triad of thought, perception and action, arranged in a triangle:

THOUGHT

/ \

PERCEPTION -- ACTION

The logic of this triad is as follows. Given a percept, constructed by perceptual processes from some kind of underlying data, a thought process decides upon an action, which is then turned into a concrete group of activities by an action process. The results of the actions taken are then perceived, along with the action itself, and fed through the loop again. The thought process must judge, on the basis of the perceived results of the actions, and the perceived actions, how to choose its actions the next time a similar percept comes around.

Representing the three processes of THOUGHT, PERCEPTION and ACTION as pattern/process magicians, the learning triad may be understood as a very basic autopoietic mental process system. Furthermore, it is natural to conjecture that learning triads are **autopoietic subsystems** of magician system dynamics.

The autopoiesis of the system is plain: as information passes around the loop, each process is created by the other two that come "before" it. The attraction is also somewhat intuitively obvious, but perhaps requires more comment. It must be understood that no particular learning triad is being proposed as an attractor, in the sense that nearby learning triads will necessarily tend to it. The claim is rather that the **class** of learning triads constitutes a probabilistic strange attractor of magician dynamics, meaning that a small change in a learning triad will tend to produce something that adjusts itself until it is another learning triad. If this is true, then learning triads should be stable with respect to small perturbations: small perturbations may alter their details but will not destroy their basic structure as a learning mechanism.

Pattern, Learning and Compression

We have cast Bohm's reflex-oriented view of mind in terms of pattern/process magicians. A reflex arc is, in the psynet model, re-cast as a triadic autopoietic magician system. In this language, proprioception of thought -- awareness of what reflexes have produced and are producing a given action -- becomes awareness of the magician dynamics underlying a given behaviour.

In this context, let us now return to the notion of pattern itself. The key point here is the relation between pattern and **compression**. Recall that to recognize a pattern in something is to compress it into something simpler - - a representation, a skeleton form. Given the overwhelmingly vast and detailed nature of inner and outer experience, it is inevitable that we compress our experiences into abbreviated, abstracted structures; into what Grof calls COEX's. This is the function of the hierarchical network: to come up with routines, procedures, that will function adequately in a wide variety of circumstances.

The relation between pattern and compression is well-known in computer science, in the fields of image compression and text compression. In this contexts, the goal is to take a computer file and replace it with a shorter file, containing the same or almost the same contents. Text compression is expected to be lossless: one can reconstruct the text exactly from the compressed version. On the other hand, image compression is usually expected to be lossy. The eye doesn't have perfect

acuity, and so a bit of error is allowed: the picture that you reconstruct from the compressed file doesn't have to be **exactly** the same as the original.

Psychologically, the result of experience compression is a certain ignorance. We can never know exactly what we do when we lift up our arm to pick up a glass of water, when we bend over to get a drink, when we produce a complex sentence like this one, when we solve an equation or seduce a woman. We do not need to know what we do: the neural network adaptation going on in our brain figures things out for us. It compresses vast varieties of situations into simple, multipurpose hierarchical brain structures. But having compressed, we no longer have access to what we originally experienced, only to the compressed form. We have lost some information.

For instance, a man doesn't necessarily remember the dozens of situations in which he tried to seduce women (successfully or not). The nuances of the different womens' reactions, the particular situations, the moods he was in on the different occasions -- these are in large part lost. What is taken away is a collection of abstracted patterns that the mind has drawn out of these situations.

Or, to take another example, consider the process of learning a tennis serve. One refines one's serve over a period of many games, by a process of continual adaptation: this angle works better than that, this posture works better so long as one throws the ball high enough, etc. But what one takes out of this is a certain collection of motor processes, a collection of "serving procedures." It may be that one's inferences regarding how to serve have been **incorrect**: that, if one had watched one's serving attempts on video (as is done in the training of professional tennis players), one would have derived quite different conclusions about how one should or should not serve. But this information is lost, it is not accessible to the mind: all that is available is the compressed version, i.e., the serving procedures one has induced. Thus, if one is asked why one is serving the way one is, one can never give a decent answer. The answer is that the serve has been induced by learning triads, from a collection of data that is now largely forgotten.

The point is that mental pattern recognition is in general highly lossy compression. It takes place in purpose-driven learning triads. One does not need to recognize all the patterns in one's tennis serving behavior -- enough patterns to generate the full collection of data at one's disposal. One only wants those patterns that are useful to one's immediate goal of developing a better serve. In the process of abstracting information for particular goals (in Bohm's terms, for the completion of particular reflex arcs), a great deal of information is lost: thus psychological pattern recognition, like lossy image compression, is a fundamentally **irreversible** process.

This is, I claim, the ultimate reason for what Bohm calls the absence of proprioception of thought. It is the reason why mindfulness is so difficult. The mind does not know what it is doing because it can do what it does far more easily without the requirement to know what it is doing. Proceeding blindly, without mindfulness, thought can wrap up complex aggregates in simple packages and proceed to treat the simple packages as they were whole, fundamental, real. This is the key to abstract symbolic thought, to language, to music, mathematics, art. Intelligence itself rests on compression: on the substitution of packages for complex aggregates, on the substitution of tokens for diverse communities of experiences. It requires us to forget the roots of our

thoughts and feelings, in order that we may use them as raw materials for building new thoughts and feelings.

If, as Bohm argues, the lack of proprioception of thought is the root of human problems, then the only reasonable conclusion would seem to be that human problems are inevitable.

8.8 THE ALGEBRA OF CONSCIOUSNESS[1]

With these philosophical notions under our belt, we are now ready to turn to a most crucial question: if the purpose of consciousness is to create autopoietic systems, then what sorts of autopoietic systems does consciousness create? The answer to this question might well be: any kind of autopoietic system. I will argue, however, that this is not the case: that in fact consciousness produces very special sorts of systems, namely, systems with the structure of quaternionic or octonionic algebras.

In order to derive this somewhat surprising conclusion from the psynet model, only one additional axiom will be required, namely, that the autopoietic systems constructed by consciousness are "timeless," without an internal sense of irreversibility (an "arrow of time"). I.e.,

In the magician systems contained in consciousness, magician operations are reversible

In view of the discussion in the previous section, an alternate way to phrase this axiom is as follows:

At any given time, proprioception of thought extends through the contents of consciousness

Bohm laments that the mind as a whole does not know what it is doing. I have argued that, on grounds of efficiency, the mind **cannot** know what it is doing. It is immensely more efficient to compress experience in a lossy, purpose-driven way, than to maintain all experiences along with the patterns derived from them. However, this argument leaves room for special cases in which thought **is** proprioceptive. I posit that consciousness is precisely such a special case. Everything that is done in consciousness is explicitly felt, in the manner of physical proprioception: it is **there**, you can sense its being, feel it move and act.

Of course, physical proprioception can be unconscious; and so can mental proprioception. One part of the mind can unconsciously sense what another part is doing. The point, however, is that conscious thought-systems are characterized by self-proprioception. I will take this an an axiom, derived from phenomenology, from individual experience. This axiom is not intended as an original assertion, but rather as a re-phrasing of an obvious aspect of the very definition of consciousness. It should hardly be controversial to say that a conscious thought or thought-system senses itself.

In the context of the psynet model, what does this axiom mean? It means, that, within the scope of consciousness, magician processes **are** reversible. There is no information loss. What is done, can be undone.

Algebraically, reversibility of magician operations corresponds to **division**. For, if multiplication represents both **action** and **pattern recognition**, then the inverse under multiplication is thus an operation of **undoing**. If

$A * B = C$

this means that by acting on B, A has produced this pattern C; and thus, in the context of the cognitive equation, that C is a pattern which A has recognized in B. Now the inverse of A, when applied to C, yields

$A^{-1} * C = B$

In other words, it restores the substrate from the pattern: it looks at the pattern C and tells you what the entity was that aa recognized the pattern in.

For a non-psychological illustration, let us return, for a moment, to the example of text compression. A text compression algorithm takes a text, a long sequence of symbols, and reduces it to a much shorter text by eliminating various redundancies. If the original text is very long, then the shorter text, combined with the decompression algorithm, will be a pattern in the original text. Formally, if B is a text, and A is a compression algorithm, then $A * B = C$ means that C is the pattern in B consisting of the compressed version of a plus the decompression algorithm. C^{-1} is then the process which transforms C into B; i.e., it is the process which causes the decompression algorithm to be applied to the compressed version of A, thus reconstituting the original A. The magician a compresses, the magician A^{-1} decompresses.

So, proprioception of thought requires division. It requires that one does not rely on patterns as lossy, compressed versions of entities; that one always has access to the original entities, so one can see "what one is doing." Suppose, then, one has a closed system of thoughts, a mental system, in which division is possible; in which mental process can proceed unhindered by irreversibility. Recall that mental systems are subalgebras. The conclusion is that proprioceptive mental systems are necessarily **division algebras**: they are magician systems in which every magician has an inverse. The kinds of algebras which consciousness constructs are division algebras.

This might at first seem to be a very general philosophical conclusion. However, it turns out to place very strict restrictions on the algebraic structure of consciousness. For, as is well-known in abstract algebra, the finite-dimensional division algebras are very few indeed.

Quaternions and Octonions

The real number line is a division algebra. So are the complex numbers. There is no three-dimensional division algebra: no way to construct an analogue of the complex numbers in three dimensions. However, there are division algebras in four and eight dimensions; these are called the quaternions and the octonions (or Cayley algebra), see e.g. Kurosh, (1963).

The quaternions are a group consisting of the entities {1,i,j,k} and their "negatives" {-1,-i,-j,-k}. The group's multiplication table is defined by the products

i * j = k

j * k = i

k * i = j

i * i = j * j = k * k = -1,

This is a simple algebraic structure which is distinguished by the odd "twist" of the multiplication table according to which any two of the three quantities {i,j,k} are sufficient to produce the other.

The **real quaternions** are the set of all real linear combinations of {1,i,j,k}, i.e., the set of all expressions a + bi + cj + dk where a, b, c and d are real. They are a four-dimensional, noncommutative extension of the complex numbers, with numerous applications in physics and mathematics.

Next, the octonions are the algebraic structure formed from the collection of entities q + Er, where q and r are quaternions and E is a new element which, however, also satisfies $E^2 = -1$. These may be considered as a vector space over the reals, yielding the real octonions. While the quaternions are non-commutative, the octonions are also non-associative. The octonions have a subtle algebraic structure which is rarely if ever highlighted in textbooks, but which has been explored in detail by Onar Aam and Tony Smith (personal communication). The canonical basis for the octonion algebra is given by (i,j,k,E,iE,jE,kE). Following a suggestion of Onar Aam, I will adopt the simple notation I = iE, J = jE, K = kE, so that the canonical basis becomes (i,j,k,E,I,J,K).

Both the real quaternions and the real octonions have the property of allowing division. That is, every element has a unique multiplicative inverse, so that the equation A * B = C can be solved by the formula $A = B^{-1} * C$. The remarkable fact is that these are not only **good** examples of division algebras, they are just about the **only** reasonable examples of division algebras. One may prove that all finite division algebras have order 1, 2, 4 or 8. Furthermore, the only division algebras with the property of **alternativity** are the real, complexes, real quaternions and real octonions. Alternativity means that subalgebras consisting of two elements are associative. These results are collected under the name of the Generalized Frobenius Theorem. Finally, the only finite algebras which are **normable** are -- the reals, complexes, real quaternions and real octonions.

What these theorems teach us is that these are not merely arbitrary examples of algebras. They are very special algebras, which play several unique mathematical roles.

Several aspects of the quaternion and octonion multiplication tables are particularly convenient in the magician system framework. Perhaps the best example is the identity of additive and

multiplicative inverses. The rule $A^{-1} = -A$ (which applies to all non-identity elements) says that undoing (reversing) is the same as annihilation. Undoing yields the identity magician 1, which reproduces everything it comes into contact with. Annihilation yields zero, which leaves alone everything it comes into contact with. The ultimate action of the insertion of the inverse of A into a system containing A is thus to either to reproduce the system in question (if multiplication is done first), or to reproduce the next natural phase in the evolution of the system (if addition is done first).

Compare this to what happens in a magician system governed by an arbitrary algebra. Given a non-identity element A not obeying the rule $A^2 = -1$, one has to distinguish whether its opposite is to act additively or multiplicatively. In the magician system framework, however, there is no easy way to make this decision: the opposites are simply released into the system, free to act both additively and multiplicatively. The conclusion is that only in extended imaginary algebras like the quaternions and octonions can one get a truly natural magician system negation.

Consciousness and Division Algebras

The conclusion to which the Generalized Frobenius Theorem leads us is a simple and striking one: the autopoietic systems which consciousness constructs are quaternionic and octonionic in structure. This is an abstract idea, which has been derived by abstract means, and some work will be needed to see what intuitive sense it makes. However, the reasoning underlying it is hopefully clear.

The notion of reversibility being used here may perhaps benefit from a comparison with the somewhat different notion involved in Edward Fredkin's (Fredkin and Toffoli, 1982) theory of **reversible computation**. Fredkin has shown that any kind of computation can be done in an entirely reversible way, so that computatinoal systems need not produce entropy. His strategy for doing this is to design reversible logic gates, which can then be strung together to produce any Boolean function, and thus simulate any Turing machine computation. These logic gates, however, operate on the basis of redundancy. That is, when one uses these gates to carry out an operation such as "A and B," enough information is stored to reconstruct both A and B, in addition to their conjunction.

The main difference between these reversible computers and the reversible magician systems being considered here is the property of **closure**. According to the psynet model, mental entities are autopoietic magician systems. In the simplest case of fixed point attractors, this implies that mental entities are algebraically closed magician systems: they do not lead outside themselves. Even in the case of periodic or strange attractors, there is still a kind of closure: there is a range of magicians which cannot be escaped. Finally, in the most realistic case of stochastic attractors, there is a range of magicians which is unlikely to be escaped. In each case, everything in the relevant range is producible by other elements in that same range: the system is self-producing. Fredkin's logic gates do not, and are not intended to, display this kind of property. It is not in any way necessary that each processing unit in a reversible computing system be producible by combinations of other processing units. The contrast with reversible computers emphasizes the nature of the argument that has led us to postulate a quaternionic/ octonionic structure for consciousness. The simple idea that consciousness is reversible does not lead you to any

particular algebraic structure. One needs the idea of reversible conscious operations **in connection with the psynet model**, which states that mental systems are autopoietic subsystems of process dynamical systems. Putting these two pieces together leads immediately to the finite division algebras.

Next, a word should be said about the phenomenological validity of the present theory. The phenomenology of consciousness is peculiar and complex. But one thing that may be asserted is that consciousness combines a sense of momentary timelessness, of being "outside the flow of time," with a sense of change and flow. Any adequate theory of consciousness must explain both of these sensations. The current theory fulfills this criterion. Consciousness, it is argued, is concerned with constructing systems that lack a sense of irreversibility, that stand outside the flow of time. But in the course of constructing such systems, consciousness carries out irreversible processes. Thus there is actually an alternation between timelessness and time-boundedness. Both are part of the same story.

Finally, let us move from phenomenology to cognitive psychology. If one adopts the view given here, one is immediately led to the conclusion that the Generalized Frobenius Theorem solves an outstanding problem of theoretical psychology: it explains why consciousness is **bounded**. No previous theory of consciousness has given an adequate answer for the question: Why can't consciousness extend over the whole mind? But in the algebraic, psynet view, the answer is quite clear. There are no division algebras the size of the whole mind. The biggest one is the octonions. Therefore consciousness is limited to the size of the octonions: seven elements and their associated anti-elements.

The occurence of the number 7 here is striking, for, as is well-known, empirical evidence indicates that the capacity of human short-term memory is about 8. The figure is usually given as 7 +/- 2. Of course, 7 is a small number, which can be obtained in many different ways. But, nevertheless, the natural occurrence of the number 7 in the algebraic theory of consciousness is a valuable piece of evidence. In no sense was the number 7 arbitrarily put into the theory. It emerged as a consequence of deep mathematics, from the simple assumption that the contents of consciousness, at any given time, is a reversible, autopoietic magician system.

8.9 MODELLING STATES OF MIND

The quaternionic and octonionic algebras are structures which can be concretized in various ways. They describe a pattern of inter-combination, but they do not specify the things combined, nor the precise nature of the combinatory operation.

Thus, the way in which the algebraic structures are realized can be expected to depend upon the particular state of mind involved. For instance, a state of meditative introspection is quite different from a state of active memorization, which is in turn quite different from a state of engagement with sensory or motor activities. Each of these different states may involve different types of mental processes, which act on each other in different ways, and thus make use of the division algebra structure quite differently.

As the quaternions and octonions are very flexible structures, there will be many different ways in which they can be used to model any given state of mind. What will be presented here are some simple initial modelling attempts. The models constructed are extremely conceptually natural, but that is of course no guarantee of their correctness. The advantage of these models, however, is their concreteness. Compared to the general, abstract quaternion-octonion model, they are much more closely connected to daily experience, experimental psychology and neuroscience. They are, to use the Popperian term, "falsifiable," if not using present experimental techniques, then at least plausibly, in the near future. They also reveal interesting connections with ideas in the psychological and philosophical literature.

But perhaps the best way to understand the nature of the ideas in this section is to adopt a Lakatosian, rather than Popperian, philosophy of science. According to Lakatos, the "hard core" of a theory, which is too abstract and flexible to be directly testable, generates numerous, sometimes mutually contradictory "peripheral theories" which are particular enough to be tested. The hard core here is the quaternionic-octonionic theory of consciousness, clustered together with the psynet model and the concepts of thought-proprioception and reversible consciousness. The peripheral theories are applications of the algebras to particular states of consciousness.

An Octonionic Model of Short-Term Memory

Consider now a state of consciousness involving memorization or intellectual thought, which requires holding a number of entities in consciousness for a period of time, and observing their interactions. In this sort of state consciousness can be plausibly identified with what psychologists call "short-term memory" or "working memory."

The most natural model of short-term memory is one in which each of the entities held in consciousness is identified with one of the canonical basis elements {i,j,k,E,I,J,K}. The algebraic operation * is to be interpreted as merely a sequencing operation. While all the entities held in consciousness are in a sense "focuses of attention," in general some of the entities will be focussed on more intensely than others; and usually one entity will be the primary focus. The equation A * B = C means that a primary focus on A, followed by a primary focus on B, will be followed by a primary focus on C.

This model of short-term memory that implies the presence of particular regularities in the flow of the focus of consciousness. The nature of these regularities will depend on the particular way the entities being stored in short-term memory are mapped onto the octonions. For instance, suppose one wants to remember the sequence

PIG, COW, DOG, ELEPHANT, FOX, WOMBAT, BILBY

Then, according to the model, these elements must be assigned in a one-to-one manner to the elements of some basis of the octonion algebra. This need not be the canonical basis introduced above, but for purposes of illustration, let us assume that it is. For convenience, let us furthermore assume that the identification is done in linear order according to the above list, so that we have

pig, cow, dog, elephant, fox, wombat, bilby

i j k E I J K

Then the theory predicts that, having focussed on PIG and then COW, one will next focus on DOG. On the other hand, having focussed on DOG, and then on the **absence** or negative of COW, one will next focus on PIG.

Of course, the order of focus will be different if the mapping of memory items onto basis elements is different. There are many different bases for the octonions, and for each basis there are many possible ways to map seven items onto the seven basis elements. But nevertheless, under any one of these many mappings, there would be a particular order to the way the focus shifted from one element to the other.

This observation leads to a possible method of testing the octonionic model of short-term memory. The model would be falsified if it were shown that the movement from one focus to another in short-term memory were totally free and unconstrained, with no restrictions whatsoever. This would not falsify the general division algebra model of consciousness, which might be applied to short-term memory in many different ways; but it would necessitate the development of a more intricate connection between octonions and short-term memory.

This octonionic model of short-term memory may be understood in terms of chaos theory -- an interpretation which, one suspects, may be useful for empirical testing. Suppose one has a dynamical system representing short-term memory, e.g. a certain region of the brain, or a family of neural circuits. The dynamics of this system can be expected to have a "multi-lobed" attractor similar to that found by Freeman in his famous studies of olfactory perception in the rabbit. Each lobe of the attractor will correspond to one of the elements stored in memory. The question raised by the present model is then one of the second-order transition probabilities between attractor lobes. If these probabilities are all equal then the simple octonionic model suggested here is falsified. On the other hand, if the probabilities are biased in a way similar to one of the many possible octonion multiplication tables, then one has found a valuable piece of evidence in favor of the present model of short-term memory and, indirectly, in favor of the finite division algebra theory of consciousness. This experiment cannot be done at present, because of the relatively primitive state of EEG, ERP and brain scan technology. However, it is certainly a plausible experiment, and there is little doubt that it will be carried out at some point over the next few decades.

Learning Triads

The previous model dealt only with the holding of items in memory. But what about the active processing of elements in consciousness? What, for example, about states of consciousness which are focussed on learning motor skills, or on exploring the physical or social environment?

To deal with these states we must return to the notion of a "learning triad," introduced above as a bridge between the psynet model and Bohm's reflex-oriented psychology. The

first step is to ask: how might we express the logic of the learning triad **algebraically**? We will explore this question on an intuitive basis, and then go back and introduce definitions making our insights precise.

First of all, one might write

THOUGHT = PERCEPTION * ACTION

ACTION = THOUGHT * PERCEPTION

PERCEPTION = ACTION * THOUGHT

These equations merely indicate that a perception of an action leads to a revised thought, a thought about a perception leads to an action, and an action based on a thought leads to a perception. They express in equations what the triad diagram itself says.

The learning triad is consistent with the quaternions. It is consistent, for example, with the hypothetical identification

PERCEPTION THOUGHT ACTION

i j k

But the three rules given above do not account for much of the quaternion structure. In order to see how more of the structure comes out in the context of learning triads, we must take a rather unexpected step. We must ask: How does the activity of the learning triad relate to standard problem-solving techniques in learning theory and artificial intelligence?

Obviously, the learning triad is based on a complex-systems view of learning, rather than a traditional, procedural view. But on careful reflection, the two approaches are not so different as they might seem. The learning triad is actually rather similar to a simple top-down tree search. One begins from the top node, the initial thought. One tests the initial thought and then modifies it in a certain way, giving another thought, which may be viewed as a "child" of the initial thought. Then, around the loop again, to a child of the child. Each time one is modifying what came before -- moving down the tree of possibilities.

Viewed in this way, however, the learning triad is revealed to be a relatively weak learning algorithm. There is no provision for "backtracking" -- for going back up a node, retreating from a sequence of modifications that has not borne sufficient fruit. In order to backtrack, one would like to actually erase the previous modification to the thought process, and look for another child node, an alternative to the child node already selected.

An elegant way to view backtracking is as **going around the loop the wrong way**. In backtracking, one is asking, e.g.: What is the thought that gave rise to this action? Or, what is the action that gave rise to this percept? Or, what is the percept that gave rise to this thought? In algebraic language, one is asking questions that might be framed

THOUGHT * ACTION = ?

ACTION * PERCEPT = ?

PERCEPT * THOUGHT = ?

In going backwards in time, while carrying out the backtracking method, what one intends to do is to wipe out the record of the abandoned search path. One wants to eliminate the thought-process modifications that were chosen based on the percept; one wants to eliminate the actions based on these thought modifications; and one wants to eliminate the new percept that was formed in this way. Thus, speaking schematically, one wants to meet PERCEPT with -PERCEPT, ACTION with -ACTION and THOUGHT with -THOUGHT. Having annihilated all that was caused by the abandoned choice, one has returned to the node one higher in the search tree. The natural algebraic rules for backtracking are thus:

THOUGHT * ACTION = - PERCEPT

ACTION * PERCEPT = - THOUGHT

PERCEPT * THOUGHT = - ACTION

Backtracking is effected by a backwards movement around the learning triad, which eliminates everything that was just laid down.

The view of learning one obtains is then one of repeated forward cycles, interspersed with occasional backward cycles, whenever the overall results of the triad are not satisfactory. The algebraic rules corresponding to this learning method are consistent with the quaternion multiplication table. The division-algebra structure of consciousness is in this way seen to support adaptive learning.

In this view, the reason for the peculiar power of conscious reasoning becomes quite clear. Consciousness is sequential, while unconscious thought is largely parallel. Consciousness deals with a small number of items, while unconscious thought is massive, teeming, statistical. But the value of conscious thought is that it is entirely self-aware, and hence it is reversible. And the cognitive value of reversibility is that it allows backtracking: it allows explicit retraction of past thoughts, actions and perceptions, and setting down new paths.

In the non-reversible systems that dominate the unconscious, once something is provisionally assumed, it is there already and there is no taking it back (not thoroughly at any rate). In the reversible world of consciousness, one may assume something tentatively, set it aside and reason about it, and then retract it if a problem occurs, moving on to another possibility. This is the key to logical reasoning, as opposed to the purely intuitive, habit-based reasoning of the unconscious.

Thought Categories and Algebraic Elements

We have posited an intuitive identification of mental process categories with quaternionic vectors. It is not difficult to make this identification rigorous, by introducing a **magician system set algebra** based on the magician system algebra given above.

To introduce this new kind of algebra, let us stick with the current example. Suppose one has a set of magicians corresponding to perceptual processes, a set corresponding to thought processes, and a set corresponding to action processess. These sets are to be called PERCEPTION, THOUGHT and ACTION. The schematic equations given above are then to be interpreted as set equations. For instance, the equation

PERCEPTION * THOUGHT = ACTION

means that:

1) for any magicians P and T in the sets PERCEPTION and THOUGHT respectively, the product P*T will be in the set ACTION

2) for any magician A in the set ACTION, there are magicains P and T in the sets PERCEPTION and THOUGHT respectively so that P*T=A

"Anti-sets" such as -PERCEPTION are defined in the obvious way: e.g. -PERCEPTION is the class of all elements R so that P = -R for some element P in PERCEPTION.

In general, suppose one has a collection of subsets S (S_1,...,S_k) of a magician system M. This collection may or may not naturally define a set algebra. In general, the products of elements in S_i and S_j will fall into a number of different classes S_m, or perhaps not into any of these classes at all. One may always define a probabilistically weighted set algebra, however, in which different equations hold with different weights. One way to do this is to say that the tagged equation

$S_i * S_j = S_k \, p_{ijk}$

holds, with

$p_{ijk} = q_{ijk}^{\,a} \, r_{ijk}^{\,2-a}$

where q_{ijk} is the probability that, if one chooses a random element from S_i, and combines it with a random element from S_j, one will obtain an element from S_k is q_{ijk}; and r_{ijk} is the probability that a randomly chosen element from S_k can be produced by combining some element of S_i with some element of S_j.

It is easier to deal with straightforward set algebras than their probabilistic counterparts. In real psychological systems, however, it is unlikely that an equation such as

PERCEPTION * THOUGHT = ACTION

could hold strictly. Rather, it might be expected to hold probabilistically with an high probability (making allowances for stray neural connections, etc.).

Finally, suppose one has a collection of subsets S and a corresponding set algebra. One may then define the

relative unity of this set algebra, as the set 1_S of all elements U with the property that U * S_i is contained in S_i for all i. The relative unity may have an anti-set, which will be denoted -1_S. These definitions provide a rigorous formulation of the correspondence between thoughts, perceptions and actions and quaternionic vectors, as proposed in the previous section.

Note that the set algebra formalism applies, without modification, to stochastic magician systems, i.e. to the case where the same magician product A*B may lead to a number of different possible outcomes on different trials.

Octonions and Second-Order Learning

Quaternions correspond to adaptive learning; to learning triads and backtracking. The octonionic algebra represents a step beyond adaptive learning, to what might be called "second-order learning," or learning about learning. The new element E, as it turns out, is most easily interpreted as a kind of second-order monitoring process, or "inner eye." Thus, in the linear combinations q+Er, the elements q are elementary mental processes, and the elements Er are mental processes which result from inner observation of other mental processes.

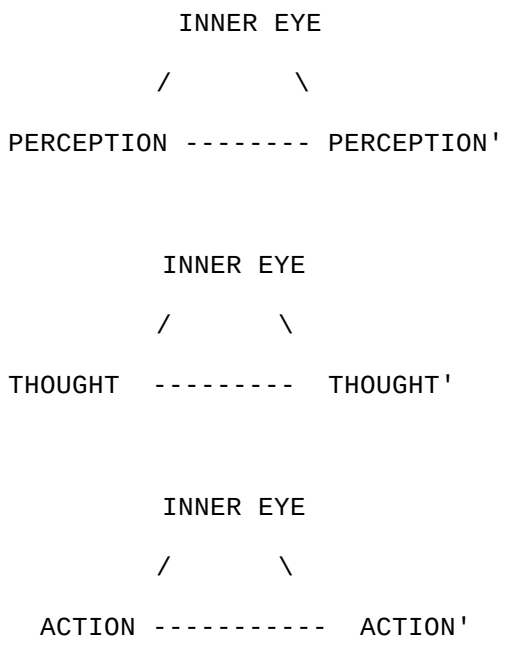
The three quaternionic elements i, j and k are mutually interchangeable. The additional octonionic element E, however, has a distinguished role in the canonical octonionic multiplication table. It leads to three further elements, I=ie, J=je and K=ke, which are themselve mutually interchangeable. The special role of E means that, in terms of learning triads, there is really only one natural interpretation of the role of E, which is given by the following:

```
PERCEPTION            THOUGHT         ACTION

        i                 j               k


INNER EYE         PERCEPTION'  THOUGHT'   ACTION'

        E           I             J         K
```

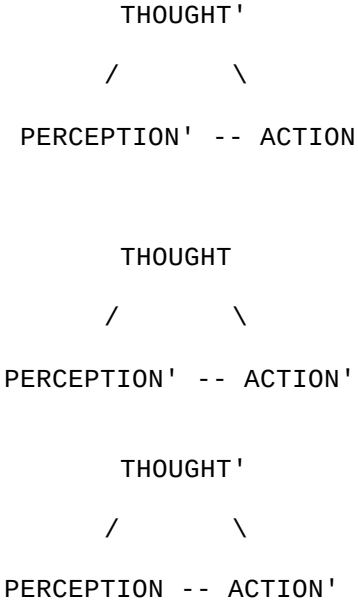
The meaning of this correspondence is revealed, first of all, by the observation that the systems of elements

```
(1,i,E,I),  (1,J,E,J),  (1,k,E,K)
```

are canonical bases of the subalgebras isomorphic to the quaternions generated by each of them. These quaternionic subalgebras correspond to learning triads of the form



The element E, which I have called INNER EYE, can thus act as a kind of second-order thought. It is thought which treats all the elements of the first-order learning triad as **percepts**: thought which perceives first-order perception, thought and action, and produces modifies processes based on these perceptions. These actions that second-order thought produces may then enter into the pool of consciously active processes and interact freely. In particular, the new perception, thought and action processes created by the inner eye may enter into the following learning triads:



These are genuine, reversible learning triads, because the sets (1,i,K,J), (1,I,j,K), and (1,I,J,k) are canonical bases for the subalgebras isomorphic to the quaternions which they generate. These, together with the basic learning triad and the three triads involving the INNER EYE, given above, represent the only seven quaternionic subalgebras contained within the octonions.

It is well worth noting that there is no loop of the form

```
          THOUGHT'

         /       \

PERCEPTION' -- ACTION'
```

within the octonion algebra. That is, complete substitution of the results of the inner eye's modifications for the elements of the original learning triad is **not** a reversible operation. One can substitute any two of the modified versions at a time, retaining one of the old versions, and still retain reversibility. Thus a complete round of second-order learning is not quite possible within the octonion algebra. However, one can attain a good approximation.

And, as an aside, the actual behavior of this "complete substitution" loop {I,J,K} is worth reflecting on for a moment. Note that IJ = -k, JK = -i, KI = -j. Traversing the complete substitution loop forward, one produces the anti-elements needed for backtracking in the original learning triad. Thus the INNER EYE has the potential to lead to backtracking, while at the same time leading to new, improved learning triads. There is a great deal of subtlety going on here, which will only be uncovered by deep reflection and extensive experimentation.

In order to completely incorporate the results of the inner eye's observations, one needs to transcend the boundaries of reversible processing, and put the new results (I,J,K) in the place of the old elementary learning triad (i,j,k). Having done this, and placed (I,J,K) in the role of the perceptual-cognitive-active loop, the octonionic process can begin again, and construct a new collection of modified processes.

What is permitted by the three triads generated by (1,i,K,J), (1,I,j,K), and (1,I,J,k) is a kind of "uncommitted" second-order learning. One is incorporating the observations of the inner eye, but without giving up the old way of doing things. The results of these new, uncommitted triads cannot be externally observed until a committment has been made; but the new triads can be used and progressively replaced, while the observing-eye process goes on.

8.10 MIND AS PREGEOMETRY

In *The Structure of Intelligence* I discussed the quantum theory of consciousness -- meaning the theory that the act of consciousness is synonymous with the collapse of the wave packet in a quantum system. I viewed this as a viable alternative to the theory that consciousness is entirely deterministic and mechanistic. However, I realized and stated the limitations of the "quantum theory of consciousness" approach. Merely to link awareness and physics together on the level of

nomenclature is not good enough. There has to be a more fundamental connection, a connection that makes a difference for both physics and psycholoyg.

Toward the end of *Chaotic Logic* I tried to push the same idea in a different direction. I discussed Feynman path integrals, the formalism according to which, in quantum theory, the amplitude (the square root of the probability) of a particle going from point A to point B is given by the sum of the amplitudes of all possible paths from A to B. Certain paths, I proposed, were algorithmically simpler than others -- and hence psychologically more likely. Perhaps the sum over all paths should be weighted by algorithmic simplicity, with respect to a given mind. Perhaps this would solve the renormalization problem, the famous "infinities" plaguing particle physics. I tried to solve the mathematics associated with this hypothesis, but in vain -- it just remained an abstract speculation.

That particular section of *Chaotic Logic* attracted no attention whatsoever. However, a number of readers of the book did remark to me that the magician system model reminded them of **particle physics**. This parallel did not occur to me at the time I conceived the magician system model -- perhaps because at the time I was reading a chemist (George Kampis) and collaborating with an algebraist (Harold Bowman) -- but in fact it is rather obvious. The combination of two magicians to form a third is similar to what happens when two particles collide and produce another particle. Antimagicians are like antiparticles. In addition to the common tool of abstract algebra, there is a strong conceptual similarity.

This idea remained vague for a while but finally began to come together when I began communicating with F. Tony Smith, a physicist at Georgia Tech. Tony has developed an original, elegant and intriguing theory of fundamental particle physics, which is based entirely on finite algebras acting on discrete spaces (Smith, 1996). The basis of all his algebras is the octonionic division algebra -- the octonions are "unfolded" to yield Clifford algebras and Lie algebras, which give the fundamental symmetry groups of all the different kinds of particles. Tony's theory was based on discrete algebraic entities living on the nodes of a graph (an eight-dimensional lattice)-- it was, in fact, a **graphical magician system**!

The correctness or incorrectness of this particular physics theory is not the point -- the point is the placing of physical and psychological models on a common mathematical ground. If both can reasonably be viewed in terms of discrete algebras living on graphs, then how difficult can it be to understand the relation between the two? This is a subject of intense current interest to me; the present section merely describes a few thoughts in this direction. While still somewhat rough-shod, I feel that these ideas indicate the kind of direction that fundamental physics must take, if it is to go beyond the mathematical confusions in which it is currently mired. The connection between the conscious mind and the physical world is there, it is active, and it cannot be denied forever.

Approaches to Quantum Measurement

I will begin by reviewing the puzzle of quantum measurement -- a puzzle that has drawn the attention of a great number of physicists, beginning with the foundation of quantum physics and continuing to the present day. I will not attempt a systematic review, but will only mention a few

of the best-known theories, and then turn to the more radical thought of John Wheeler, which will serve as a springboard for the main ideas of the section.

The key fact here is that a quantum system does not have a definite state: it lives in a kind of probabilistic superposition of different states. Yet somehow, when we measure it, it assumes a definite state.

The standard "Copenhagen interpretation" of quantum measurement states that, when a measurement is made, a quantum system suddenly collapses from a probabilistic superposition of states into a definite state. This nonlinear "collapse of the wave function" is an additional assumption, which sits rather oddly with the linear evolution equations, but poses no fundamental inconsistency. The act of measurement is defined somewhat vaguely, as "registration on some macroscopic measuring device." The fact that macroscopic measuring devices are themselves quantum systems is sidestepped.

London and Bauer (XX), Wigner (XX), Goswami (XX), Goertzel (XX) and others have altered the Copenhagen interpretation by replacing the macroscopic measuring instrument with **consciousness** itself. According to this view, it doesn't matter whether the photoelectric emulsion blackens, it matters whether the blackening of the emulsion enters someone's consciousness. The weak point here, of course, is that consciousness itself is not a well-defined entity. One is replacing on ill-defined entity, measurement, with another.

The interpretation of the measurement process that comes closest to a direct translation of the mathematical formalism of quantum theory is the **many-universes** theory, which states, that every time a measurement is made, universes in which the measurement came out one way are differentiated from universes in which it came out another way. This interpretation does not add an extra step to quantum dynamics, nor does it appeal to extra-physical entities. It is, in David Deutsch's phrase, short on assumptions but long on universes. The ontological status of the alternate universes is not quite clear; nor is it clear when a measurement, which splits off universes from each other, should be judged to have occured.

John Wheeler has taken a somewhat different view of the problems of quantum physics, one which is extremely relevant to the ideas of the present paper. While acknowledging the elegance and appeal of the many-universes theory, he rejects it because

[T]he Everett interpretation takes quantum theory in its present form as **the** currency, in terms of which everything has to be explained or understood, leaving the act of observation as a mere secondary phenomenon. In my view we need to find a different outlook in which the primary concept is to make meaning out of observation and, from that **derive** the formalism of quantum theory.

Quantum physics, in Wheeler's view, has uncovered the fundamental role of the observer in the physical world, but has not done it justice. The next physics breakthrough will go one step further, and place the observer at the **center**.

Wheeler also believes that the equations of quantum theory will ultimately be seen to be **statistical** in nature, similar to the equations of thermodynamics:

I believe that [physical] events go together in a higgledy-piggledy fashion and that what seem to be precise equations emerge in every case in a statistical way from the physical of large numbers; quantum physics in particular seems to work like that.

The nature of this statistical emergence is not specified; except that, preferably, one would want to see physical concepts arise out of some kind of **non-physical** substrate:

If we're ever going to find an element of nature that explains space and time, we surely have to find something that is deeper than space and time -- something that itself has no localization in space and time. The ... elementary quantum phenomenon ... is indeed something of a pure knowledge-theoretical character, an atom of information which has no localization in between the point of entry and the point of registration.

This hypothetical non-physical substrate, Wheeler has called "pregeometry." At one point, he explored the idea of using propositional logic as pregeometry, of somehow getting space and time to emerge from the statistics of large numbers of complex logical propositions (Wheeler, 19XX). However, this idea did not bear fruit.

Wheeler also suspects the quantum phenomenon to have something to do with the social construction of **meaning**:

I try to put [Bohr's] point of view in this statement: 'No elementary quantum phenomenon is a phenomenon until it's brought to a close by an irreversible act of amplification by a detection such as the click of a geiger counter or the blackening of a grain of photographic emulsion.' This, as Bohr puts it, amounts to something that one person can speak about to another in plain language....

Wheeler divides the act of observation into two phases: first the bringing to close of the elementary quantum phenomenon; and then the construction of meaning based on this phenomenon. The accomplishment of the first phase, he suggests, seems to depend on the **possibility** of the second, but not the actual **accomplishment** of the second. A quantum phenomenon, he suggests, is not a phenomenon until it is **potentially meaningful**. But if, say, the photoelectric emulsion is destroyed by a fire before anyone makes use of it, then the elementary quantum phenomenon was still registered. This is different from the view that the quantum phenomenon must actually enter some entity's consciousness in order to become a phenomenon.

Regarding the second phase of observation, the construction of meaning, Wheeler cites the Norwegian philosopher Follesdal that "meaning is 'the joint product of all the evidence that is available to those who communicate.'" Somehow, a quantum phenomenon becomes a phenomenon when it becomes evidence that is in principle available for communication.

In the end, Wheeler does not provide an alternative to the standard quantum theories of measurement. Instead, he provides a collection of deeply-conceived and radical ideas, which fit together as elegantly as those of any professional philosopher, and which should be profoundly thought-provoking to anyone concerned with the future of physical theory.

The Relativity of Reality

A particularly subtle twist to the puzzle of quantum measurement has been given in Rossler (19XX). Rossler describes the following experiment. Take two particles -- say, photons. As in the Aspect experiment (19XX), shoot them away from each other, in different directions. Then, measure the photons with two different measuring devices, in two different places. According to special relativity, **simultaneity is not objective**. So, suppose device 1 is moving with respect to device 2, and event A appears to device 1 to occur before event B. Then it is nonetheless possible that to device 2, event B should appear to occur before event A.

Then, in the Aspect experiment, suppose that device 1 measures photon A, and device 2 measures photon B. One may set the devices up so that, in the reference frame of device 1 photon A is measured first, but in the reference frame of device 2 photon B is measured first. In the reference frame of each of the measuring devices, one has precisely the Aspect experiment. But one also has a problem. The notion of a **measured state** must be redefined so that it becomes **reference-frame-dependent**. One cannot say in any "objective" way whether a given probabilistic superposition of states has collapsed or not; one can only say whether or not it has collapsed within some particular reference frame.

Aharanov and Albert (19XX), discussing a similar experiment, conclude that the notion of a single **time axis** is inadequate for describing physical reality. They conclude that each observer must be considered to have its own "personal" time axis, along which probabilistic superpositions gradually collapse into greater definiteness.

And if time is multiple, what about space? The verdict here is not quite so clear, but on balance, it seems probable that we will ultimately need subjective spaces to go along with subject time axes. This idea comes from the study of quantum theory in curved spacetime (19XX). In this theory, one arrives at the remarkable conclusion that, if two observers are moving relative to one another, and both look at the same exact spot, **one observer may see a particle there while the other does not**. This means that, in curved spacetime, particles have no real existence. Of course, it is difficult to interpret this result, since quantum physics in curved spacetime is not a genuine physical theory, just a sophisticated mathematical cut-and-paste job. But nonetheless, the result is highly suggestive.

These results suggest new troubles for the quantum theory of measurement -- troubles beyond the mere collapse of the wave packet, splitting of universes, etc. In the many-universes picture, instead of the whole human race living in a universe which periodically splits off into many, one must think of each observer as living in his own individual universe, which splits up periodically. These ideas sit rather oddly with Wheeler's concept of observation as having to do with the communication and the collective construction of meaning. Suppose something is only

measurable when it can, potentially be communicated throughout a community of minds. One must now ask: in whose universe do these community of minds actually exist?

Minds and Physical Systems

The next step is to tie these physical ideas in with the psynet model. But first, before we can do this, we must remind ourselves how the abstract process systems posited in the psynet model relate to physical systems.

This question might seem to be a restatement of the basic puzzle of the relation between mind and reality. For the moment, however, we are dealing with something simpler. We are merely dealing with the question of how one type of dynamical system (magician systems) deals with another type of dynamical system (physical systems). Or, to phrase it more accurately, we are dealing with the question of how one **level** of dynamical system relates to another **level** of dynamical system. For, after all, it is quite possible that physical systems themselves are magician systems -- as we shall see in the following section, there are models of particle physics which say just this.

Following previous publications, I will take a pragmatic approach to the relation between minds and brains. Invoking the theory of algorithmic information, we will assert that minds are **minimal algorithmic patterns** in brains. A magician system is a minimal pattern in a physical system *P* if:

1) the magician system is more "simply given" than the brain

2) the structure of the magician system, at a given time and in evolution over time, is similar to the structure of the system *P*

3) no part of the magician system can be removed without making

the magician system less structurally similar to the

system P

These two criteria may be quantified in obvious ways. A mind associated with a physical system *P* is a magician system which is a pattern in the system *P*, in this sense.

This approach intentionally sidesteps the question of the basic nature of awareness. In this respect, it is perhaps wise to reiterate the distinction between raw awareness and consciousness. Awareness, considered as the basic sense of existence or presence in the world, is not analyzed here, and is perhaps not analyzable. Consciousness, however, is considered as awareness mediated by mental structures. Different states of consciousness involve different mental structures. Thus, mind affects consciousness, but does not necessarily affect awareness.

In terms of quantum measurement, we may say that an elementary quantum phenomenon becomes a phenomenon, from the point of view of a given mind, when it enters that mind. Thus,

an elementary quantum phenomenon becomes a phenomenon to a given mind when it becomes a part of a minimal algorithmic pattern in the system supporting that mind. Taking an animist view, one may say that, when a quantum phenomenon registers on a photoelectric emulsion, it becomes a phenomenon from the point of view of that emulsion. It becomes a part of the "mind" of that emulsion. On the other hand, when a person views the emulsion, the quantum phenomenon becomes a phenomenon from the point of view of that person.

This subjectivistic view may seem problematic to some. Given the thought-experiments described above, however, it seems the only possible course. Psychologists have long understood that each mind has its own subjective reality; quantum physics merely extends this subjectivity in a different way. It shows that physical phenomena, considered in a purely physical sense, do not have meaning except in the context of someone's, or something's, subjective reality.

The Mental-Physical Optimality Principle

Now let us get down to business. Recall Wheeler's assertion that quantum mechanics should arise from the statistics of some underlying, pregeometric domain. He selected propositional logic as a candidate for such a "pregeometry." From a psychological point of view, however, propositional logic is simply a crude, severely incomplete model of the process of intelligent thought (see *SI* and *CL*). It represents an isolation of one mental faculty, deductive logic, from its natural mental environment. Instead of propositional logic, I propose, the correct pregeometry is **mind itself**. Physics, I propose, results from the statistics of a large number of **minds**.

Suppose that, as the psynet model suggests, minds can be expressed as hypercomplex algebras, and that mental process can be understood as a nonlinear iteration on hypercomplex algebras. In this context, we can raise the question "How might a collection of minds give rise to a physical reality?" in a newly rigorous way. The question becomes: How might a collection of hypercomplex algebras, and the attractors and meta-attractors within these algebras under an appropriate nonlinear dynamic, give rise to a physical reality?

Given that minds themselves are understood as patterns in physical systems, what is being proposed here must be seen as a kind of circular causation. Physical reality gives rise to physical systems, which give rise to minds, which in turn combine to produce physical reality. We know how physical reality gives rise to minds -- at least, to structured systems, which interact with awareness to form states of consciousness. We do not know how minds give rise to physical reality.

My answer to this question is a simple one: Minds **sum up** to form physical reality. This sum takes place in an abstract space which may be conceptualized as a **space of abstract algebras**.

The nature of this summation may be understood by analogy to Feynman integrals in quantum physics. In the simplest example, a Feynman integral is a sum over all possible paths from one point to another. One assigns each path a two-dimensional vector, the angular coordinate of which is given by the energy of the path, divided by the normalized Planck's constant. Then one adds up these vectors, and divides the sum by an overall normalization factor. The sum is the amplitude (the square root of the probability) of the particle in question going from the first point

to the second. The key is that nearly all of the paths cancel out with each other. The vectors are pointing in a huge number of different directions, so that ultimately, the only vectors that make a significant contributions are the ones that are bunched together with a lot of other vectors, i.e., the ones that are near local extrema of the energy function. The amplitude of the transition is thus approximately given by the local extrema of the energy function. As Planck's constant tends to zero, the cancellation of alternative paths becomes complete, and the "optimal" path is entirely dominant. As Planck's constant is increased, the cancellation of the alternative paths increasingly fails, and there is more of a "random" flavor to the assignment of amplitudes.

When adding together minds, we are not working in a two-dimensional complex space, but rather in an n-dimensional space of abstract algebras. Here the dimension n is large, and will vary based upon the number and size of minds involved -- one could speak of an infinite-dimensional Hilbert space, with the understanding that only a finite number of dimensions will be involved in any particular calculation. Despite the increase in dimensionality, however, the principle is the same as with Feynman integrals. We are adding together a huge variety of n-vectors, which are spreading out in all different directions. Mainly, these vectors will cancel each other out. But certain substructures common to a great number of the algebras will remain. These substructures will be ones that are "optimally efficient" in the sense that a significant percentage of the minds in the universe have evolved so as to possess them. My claim is that these optimal substructures are precisely **physical reality**. This is what I call the Mental-Physical Optimality Hypothesis.

This, finally, is the basic concept of my reduction of physics to psychology: that physical structures are precisely the most efficient and hence common structures of mind, so that when one sums together a vast number of minds, it is the precisely the physical structures which are not cancelled out.

Note that, in this theory, the process of summation does not eliminate the summands. Individual minds maintain their individual existence; but at the same time, they sum together to produce the collective physical reality, from which they all emerge. Mind and physical reality create each other, on a recurrent and ceaseless basis.

One might wonder whether, once mind has gone ahead and created a new physical reality, the old one entirely disappears. But this is a philosophical extravagance. One might as well suppose that physical entities have a certain lasting existence, a certain tendency to persist, and that they fade from existence only if their component parts are not re-created after a certain amount of time has lapsed. In this way one retains the fundamental continuity and consistency of the universe, and also the idea that mind and reality create each other. One has a kind of "illusory continuity" emerging from an underlying discrete dynamic of mind-reality intercreation. The kind of "tendency to persist" being invoked here has a venerable history in philosophy, for instance in the work of Charles S. Peirce (19XX).

The Efficiency of Finite Division Algebras

If one accepts a discrete theory of physics such as that of (Smith, 1996), then physical systems and psychological systems live in the same space, so that minds can indeed be summed together

to yield physical realities. In order to make this abstract idea convincing, however, we must give some concrete reason to believe that the **specific** algebraic structures involved in thought have something to do with the **specific** algebraic structures involved in the physical structure of the universe. This might be done in any number of different ways: the general mental-physical optimality principle, and the concept of mind as pregeometry, do not rely on any particular algebraic structure. At present, however, only one particular psychological-physical correspondence has presented itself to us, and this is the octonionic algebra. The octonions appear in the magician system of model of consciousness, and also in the discrete theory of physics. In both cases they represent the same intuitive phenomenon: **the structure of the present moment.**

One might well wonder why such a correspondence **should** exist. Why should physically important structures also be psychologically important structures? The answer to this question lies, we believe, in mathematics: Certain mathematical structures possess an intrinsic efficiency, which causes them to arise time and time again in different circumstances.

In this case, the efficient algebraic structure in question is the octonionic algebra. The reason for the importance of the octonions is very simple: There is a theorem which states that the only finite-dimensional division algebras are one, two, four and eight-dimensional. Among these, the only algebras with reasonable algebraic properties are the real numbers, the complex numbers, the quaternions and the octonions. The octonions are the largest reasonably structured algebra with the property of unique division. And the property of unique division is, one suspects, crucial for efficient functioning, both in physics and in psychology.

This is, as yet, not a scientific idea. It is a philosophical idea which takes mathematical form. I have not found any way to put the idea to empirical test. It should be remembered, however, that many widely accepted notions in particle physics are equally distant from the realm of empirical test. String theory is the best known example. In fundamental physics, as in theoretical cognitive science, experimentation is difficult, so that conceptual coherence and fecundity emerge as crucial criteria for the assessment of theories. I believe that, judged by these standards, the "mind as pregeometry" idea excels.

8.11 CONCLUSION

Richard Feynman said, "Whoever tries to understand quantum theory, vanishes into a black hole and will never be heard from again." The same might be said about consciousness -- and doubly for those who try to understand the relation between quantum theory and consciousness! The phenomenon of consciousness, like the elementary quantum phenomenon, displays layer after layer of complexity: one can keep unfolding it forever, always coming to new surprises, and always dancing around the raw, ineffable core.

The treatment of consciousness given here surpasses the one I gave in *Chaotic Logic*; and my own research on consciousness has extended significantly beyond the ideas of this chapter. But yet even this more extensive research is still in many ways incomplete. What I hope to have accomplished here, if nothing else, is to have illustrated some of the ways in which consciousness can be investigated in the context of the psynet model.

The concept of the dual network leads to the perceptual-cognitive loop, which is a useful abstraction of diverse neurobiological data. The concept of mental systems as magician systems leads to the division algebra theory of consciousness, which again provides a useful way of synthesizing different psychological observation. Division algebras connect many seemingly disparate things: Bohm's proprioception of thought, the magic number 7 +/- 2, the relation between mind and quantum reality,.... The psynet model does not, in itself, solve **any** of the puzzles of consciousness. But it does provide a useful and productive framework for thinking about the various issues involved.

CHAPTER NINE

FRACTALS AND SENTENCE PRODUCTION

9.1 INTRODUCTION

Language is a highly complex system. Modern linguistics, however, pays little if any explicit heed to this complexity. It **reflects** the complexity of language, in its labyrinthine theorizing. But it does not attempt to **come to grips** with this complexity in any concrete way. Rather than focussing on abstract, emergent structures, it revels in the intricate, subtlely patterned details.

In this chapter I will explore the parallels between language and other complex systems, in the specific context of **sentence production**. I will use a general mathematical model called **L-systems** to give a detailed psycholinguistic model of sentence production. The resulting model of sentence production has many relations with previous models, but also has its own unique characteristics, and is particularly interesting in the context of early childhood language. It fits in very naturally with the psynet model and the symbolic dynamics approach to complexity.

L-systems were introduced by Aristid Lindenmayer to model the geometry and dynamics of biological growth (see Lindenmayer, 1978). Recent work in biology and computer graphics (Prusinciewicz and Hanan, 1989) has demonstrated their remarkable potential for the generation of complex forms, especially the forms of herbaceous plants. But, although L-systems were originally inspired by the theory of formal grammar, the possibility of their direct relevance to linguistic phenomena has not been considered.

In the model given here, the production of a sentence is viewed as the iteration of an L-system. The L-system governs the process by which mental structures are progressively turned into sentences through a series of expansions and substitutions. At each stage of the iteration process, the need for accurate depiction of mental structures is balanced against the need for global simplicity and brevity, a type of "global iteration control" which is similar to that required for L-system modeling of woody plants (Prusinciewicz and Hanan, 1989).

9.2 L-SYSTEMS

The basic idea of an L-system is to construct complex objects by successively replacing parts of a simple object using a set of **rewrite rules**. The rewriting is carried out recursively, so that structures written into place by a rewrite rule are subsequently expanded by further application of rewrite rules. The crucial difference between L-systems and the Chomskygrammars used in linguistics is the method of applying rewrite rules. In Chomsky grammars, rules are applied sequentially, one after the other, while in L-systems they are applied in parallel and simultaneously replace all parts of the object to which they are applied. This difference has computation-theoretic implications -- there are languages which can be generated by context-free L-systems but not by context-free Chomsky grammars. And it also, we will argue, has **psycholinguistic** implications in the context of sentence production.

Before presenting formal details we will give a simple example. Suppose one is dealing with strings composed of "a"'s and "b"'s, and one has an L-system containing the following two rewrite rules:

X -> Y

X -> XY

Finally, suppose one's initial **axiom** is "a." Then the process of L-system iteration yields the following derivation tree:

Time	System State
0	X
1	Y
2	XY
3	YXY
4	XYYXY
5	YXYXYYXY
6	XYXXYXXYXYXXY

...

In going from time t to time t+1, all elements of the system at time t are replaced at once.

This example points out a terminological conflict -- in the L-system literature, the system elements are referred to as **letters**, and the system states themselves, the strings, are referred to as **words**; but in our application to sentence production, the system states will be **sentences** and the individual system elements will be "words" in the usual sense. To avoid confusion we will

eschew the usual L-system terminology and speak either of "words" and "sentences" or of "system elements" and "system states."

More formally, let S be a set called the "collection of system elements," and let S* be the set of all system states constructible from arrangements of elements of S. In typical L-system applications S* is, more specifically, the set of all finite, nonempty **sequences** formed from elements of S. A **context-free L-system** is defined as an ordered triple <S,w,P>, where w is an element of S* called the **axiom** (X in the above example) and P, a subset of S x S*, is the set of **rewrite rules**. In accordance with the conventions of set theory a rewrite rule should be written (X,m), where X is a system element and m is a system state, but the notation X -> m is more intuitive and, following much of the literature, we will use it here. In this formulation X is the **predecessor** and m is the **successor**. If no rewrite rule is explicitly given for some specific predecessorX, it is assumed that the identity rule X->X applies to that system element.

An L-system is **deterministic** if each system element a serves as the predecessor in only one rewrite rule. **Stochastic** L-systems allow system elements to appear in any number of rules, but each rule is associated with a certain probability. For instance, resuming our earlier example, one might have the rule set

X -> Y with probability .5

X -> YX with probability .5

Y -> XY with probability .8

Y -> Y with probability .2

This collection of rules will yield a different derivation every time. Each time an X comes up in the derivation, a random choice must be made, whether to replace it with Y or with YA. And each time a Y comes up a choice must be made whether to replace it with ab or to leave it alone.

Next, a **context-sensitive** L-system has a more general set of rewrite rules. For instance, one may have rules such as

X<X>X -> X

Y<X>X -> Y

where the brackets indicate that it is the central X which is being rewritten. According to these rules, if the central X is surrounded by two other X's, it will be left alone, but if it is preceded by a Y and followed by an X, it will be replaced by a Y.

Turtle Graphics

To apply L-systems to an applied problem, one must figure out a suitable interpretation of the strings or other system states involved. In the case of sentence production the interpretation is obvious, for, after all, L-systems are a modification of Chomsky grammars; but things are not so clear in the case of biological development, where many different interpretations have been proposed and used to advantage.

For the purpose of constructing computer simulations of plant development, the **turtle graphics** interpretation has been found to be particularly useful. In this interpretation each element of a string is taken to correspond to a command for an imaginary "turtle" which crawls around the computer screen. The simplest useful command set is:

F Move forward a step of length d

\+ Turn right by angle a

When the turtle moves forward it draws an line and hence from a sequence of F's and +'s a picture is produced. But other commands are also convenient and the list of commands quickly multiplies, especially when one considers three-dimensional growth processes. Much of the recent work involves parametrized L-systems, with system elements such as F(d) and +(a), which include real number or real vector parameters.

Using only very simple rewrite rules one may generate a variety of complicated forms. For instance, the rules

X -> X+YF++YF-FX--FXFX-YF+

Y -> -FX+YFYF++YF+FX--FX-Y

generate the fractal picture shown in Figure 5a. The command

\- Turn left by angle a

is used in addition to F and + , and the angle parameter is set at

a = 60 degrees

To produce this picture the L-system was iterated 4 times beginning from the axiom

XF

The string resulting from this iteration was used to control a simulated "turtle": each time the turtle encountered an F it moved one unit forward, each time it encountered a + or - it performed an appropriate rotation, and each time it encountered an X or a Y it ignored the symbol and proceed to the next meaningful control symbol.

Two additional control symbols which are extremely useful for modeling plant growth are the left and right brackets, defined as follows:

[Push the current state of the turtle onto a pushdown stack

] Pop a state from the stack and make it the current state of

the turtle

This adds an extra level of complexity to the form generation process, but it does not violate the underlying L-system model; it is merely a complicated interpretation. The brackets permit easy modeling of **branching structure**. The turtle can draw part of a branch, push the current state onto the stack, then follow a sequence of commands for drawing a sub-branch, then pop back the previous state and continue drawing the original branch. Figure 5b displays examples of plant-like structures generated from bracketed L-systems. Figure 6 gives a more realistic-looking example, in which leaves generated by B-spline patches are arranged according to the path of a three-dimensional turtle.

These models produce high-quality computer graphics, but they are not merely clever tricks; their efficiency and accuracy is due to their relatively accurate mimicry of the underlying biological growth processes. Plant growth is one of many biological phenomena which are governed by the process of **repeated concurrent substitution** that lies at the essence of the L-system model.

The approach we have described is quite effective for modeling the development of herbaceous, or non-woody plants. Woody plants display the same fundamental branching structure, but they are more complex. First of all there is secondary growth, which is responsible for the gradual increase of branch diameter with time. And secondly, while for herbaceous plants genetic factors are almost the sole determinant of development, for woody plants environmental factors are also important. Competition between branches and competition between trees are both significant factors. Because of these competition factors, a substitution cannot be applied without paying heed to the whole structure of the tree at the current time. The success of a substitution depends on the "space" left available by the other branches. The course of a branch may change based on the positions of the other branches, meaning, in the turtle graphics interpretation, that the turtle can change position based on collisions with its previous path.

Sentence production displays the same kind of complexity as the modeling of woody plants. The development of a sentence is not entirely internal, it can be influenced by the simultaneous development of other sentences (i.e. the sentences that will follow it). And there will often be competition between different parts of a sentence, each one vying to be further developed than the others. These complexities give a unique flavor to the L-system modeling of sentence production; but they are not an obstacle in the way of such modeling.

9.3 SENTENCE PRODUCTION AND THOUGHT PRODUCTION

Now let us turn the the main topic of the chapter, the application of L-systems to psycholinguistic modeling. Much less attention has been paid to the problem of language production than to the related problem of language **acquisition** (Anderson, 1983; Chomsky, 1975). Nevertheless, as many theorists have noted (Anisfeld, 1984; Butterworth, 1980; Jackendoff, 1987), no one has yet formulated an entirely satisfactory theory of sentence production. The concept of L-system, I will argue, is an important step on the way to such a theory. Sentence production is well understood as a process of messy, parallel, biologcal-fractal-style development.

Before presenting the L-system approach to sentence production, however, we must first deal with a few preliminary issues, regarding the logical form of language and, especially, the relation between **syntax** and **semantics** in the language production process. These are troublesome, controversial issues, but they cannot be avoided entirely.

Perhaps the first detailed psychological model of language production was that of the German neurologist Arnold Pick (1931). Pick gave six stages constituting a "path from thought to speech":

1) Thought formulation, in which an undifferentiated thought is divided into a sequence of topics or "thought pattern," which is a "preparation for a predicative arrangement ... of actions and objects."

2) Pattern of accentuation or emphasis

3) Sentence pattern

4) Word-finding, in which the main content words are found

5) Grammatization -- adjustments based on syntactic roles of content words, and insertion of function words

6) Transmission of information to the motor apparatus

Pick's sequential model, suitably elaborated, explains much of the data on speech production, especially regarding aphasia and paraphasia. And it can also be read between the lines of many more recent multileveled models of speech production (Butterworth, 1980). For instance, Jackendoff's "logical theory" of language production rests on the following assumption:

The initial step in language production is presumably the formulation of a semantic structure -- an intended meaning. The final step is a sound wave, which is a physical consequence of motions in the vocal tract. The job of the computational mind in production is therefore to map from a semantic structure to a sequence of motor instructions in the vocal tract. Again, the logical organization of language requires that this mapping be accomplished in stages, mapping from semantic structure to syntax, thence to phonology, thence to motor information.

Here we have three stages instead of six, and a rhetoric of computation, but the basic model is not terribly dissimilar.

Jackendoff implicitly connects Pick's stages of sentence production with the standard ideas from **transformational grammar** theory (Radford, 1988). Chomsky's "deep structure" corresponds to the emphasis pattern and sentence pattern of Pick's steps 3 and 4; whereas Chomsky's "surface structure" is the result of Pick's step 5. Grammatical transformations take a deep structure, a primary, abstract sentence form, and transform it into a fully fleshed-out sentence, which is then to be acted on by phonological and motor systems.

From this sequentialist perspective, the L-system model of sentence production to be presented here might be viewed as an explanation of precisely **how** grammatical transformations are applied to change a deep structure into a surface structure. In other words, they explain the transition from steps Pick's 3 and 4 to step 5; and they are internal to Jackendoff's "syntactic component" of the production system. I will call this the **purely grammatical** interpretation of the L-system model.

On the other hand, one may also take a broader view of the process of sentence production. It seems plain that there is a substantial amount of overlap between Pick's steps 1-5. The "two" processes of idea generation and sentence production are not really disjoint. In formulating an idea we go part way toward producing a sentence, and in producing a sentence we do some work on formulating the underlying idea. In fact, there is reason to believe that the process of producing sentences is inseparable from the process of formulating thoughts. Nearly two decades ago, Banks (1977) described several experimental results supporting the view that

[S]entences are not produced in discrete stages. Idea generation is not separable from sentence construction.... In typical speech production situations, sentences and ideas are produced simultaneously in an abstract code corresponding to speech. That is to say, our response mode determines the symbolic code used in thinking and influences the direction and structure of thought. Thought is possible in several modes, but one of the most common is speech.... [I]dea generation and sentence production represent the same functional process.

This view ties in with the anthropological theories of Whorf (1949) and others, according to which thought is guided and structured by language.

The L-system model to be given here is to a large extent independent of the thought/language question. Under the "purely grammatical" interpretation, it may be viewed as a model of Pick's stages 2-5, i.e. as an unraveling of transformational syntax into the time dimension. But, under what I call the **grammatical-cognitive interpretation**, it may also be taken more generally, as a model of both thought formulation and sentence production, of all five of Pick's initial stages considered as partially **concurrent** processes. In order to encompass both possible interpretations, I will keep the formulation as general as possible. In the remainder of this section I will first explore the grammatical-cognitive interpretation in a little more detail, and then present a general model of language which is independent of the choice of interpretation.

Models of Conceptual Structure

The "grammatical-cognitive" view of sentence production is closely related with Jackendoff's (1990) theory of **conceptual semantics**. According to conceptual semantics, ideas are governed by a grammatical structure not so dissimilar from that which governs sentences. For instance, the sentence

1) Sue hit Fred with the stick

would be "conceptually diagrammed" as followed:

|CAUSE(SUE,GO(STICK,TO FRED))|

| ACT(SUE,FRED) |

|

P

P_1 R P_2

| |

|ACT(SUE,STICK)| |ACT(STICK,FRED)|

The idea here is that semantic roles fall into two different tiers: one "thematic," dealing with motion and location; the other "active," dealing with agent and patient relations.

This semantic diagram is very different from the standard syntactic diagram of the sentence, which looks something like

IP

/ \

/ \

NP VP

| / \

| V' PP

| | |

| V NP

| | |

Sue hit Fred with the stick

The problem with Jackendoff's theory is the connection between syntactic structures and semantic structures. Some simple correspondences are obvious, e.g. agents go with subjects, and patients go with objects. But beyond this level, things are much more difficult.

Bouchard (1991) provides an alternative which is much less problematic. Bouchard's theory is based on an assumption called the **universal bracketing schema**, which states that "two elements may be combined into a projection of one of the two, subject to proper combinatory interpretation." The basic operation of conceptual semantics is thus taken to be **projection**. For instance, the sentence "Sue hit Fred with the stick" is diagrammed as follows:

[[[x[CAUSE[x[GO[TO y]]]] [with z]] [Time AT t] [Place AT p]]

where x = SUE, y = FRED, z = STICK

Each set of brackets groups together two entities, which are being jointly projected into one member of the pair. Bouchard's "relative theta alignment hypothesis" states that rank in this **conceptual** structure corresponds with rank in **syntactic** structure. The highest argument in the conceptual structure is linked to the subject position, the most deeply embedded argument is linked to the direct object position, and so on.

The details of Jackendoff's, Bouchard's and other formal semantic theories are complicated and need not concern us here -- I will approach conceptual structure from a slightly different perspective. The key point, for now, is that the grammatical-cognitive connection is not a vague philosophical idea; it is a concrete hypothesis which has a great deal of linguistic support, and which has been explored in the context of many different concrete examples.

A General Axiomatic Model of Language

In accordance with the above remarks, I will assume a very general and abstract model of linguistic and psycholinguistic structure. This model emanates directly from the psynet model, which views mental entities as a loosely-geometrically-organized "pool," freely interacting with each other, subject to a definite but fluctuating geometry.

Given a finite set W, called the collection of "words," I will make seven assumptions. The first five assumptions should be fairly noncontroversial. The final two assumptions are onlynecessary for the grammatical-cognitive interpretation of the model, and not for the purely grammatical interpretation; they may perhaps be more contentious.

We will assume:

1) that there is a collection of linguistic categories, inclusive of all words in the language.

2) that there is a finite collection of "basic linguistic forms," each of which is made up of a number of linguistic categories, arranged in a certain way. These forms may be represented as ordered strings $(i_1,...,i_n)$ where i_k denotes category C_k. Standard examples are N V and N V N, where N = noun and V = verb.

3) that there is a finite collection of **rewrite rules** f_i, which allow one to substitute certain arrangements of categories for other arrangements of categories (e.g. N -> Adj N allows one to replace a noun with an adjective followed by a noun).

4) that there is a space M of "mental structures" which arrangements of words are intended to depict. Examples from human experience would be pictures, memories, sounds, people, etc. In computer models these might be graphs, bitmaps, or any other kind of data structure.

5) that there is a metric, implicit or explicit, by which sentences may be compared to elements of M, to determine the relative accuracy of depiction.

6) that, referring back to (2), there is a collection of functions $\{g_i, i=1,...,r\}$ mapping M into M, so that when the arguments of f_i are close to the arguments of g_i, the value returned by g_i is close to the value returned by f_i (where "closeness" is measured by the metric guaranteed in Assumption 5).

7) that there exist in the mind mixed "linguistic forms" which are partly composed of words and partly composed of other mental structures. These mixed structures are constructed by a combination of the linguistic rules f_i and the semantic rules g_i. Assumptions 6 and 7, as stated above, are needed only for the grammatical-cognitive interpretation of the model. They are a formalization of the idea that sentence production and idea generation are two aspects of the same fundamental process. Because they are principles and not specific diagramming methods, they are less restrictive than the theories of Jackendoff and Bouchard, discussed above. However, they express the same fundamental intuitions as the diagramming methods of these theorists.

Assumption 6 is a version of Frege's principle of compositionality which has been called the **principle of continuous compositionality** (Goertzel, 1994). Frege argued that, when we compose a complex linguistic form from simple ones using certain rules, the **meaning** of the complex form is composed from the meaning of the simple ones using related rules. Assumption 6 may be interpreted as a reformulation of the principle of compositionality in terms of a **pragmatic** account of "meaning," according to which the meaning of a sentence S is the collection of mental structures which are close to S under the metric guaranteed by assumption 5.

Assumption 6 is necessary if one is going to postulate a continual "switching back and forth" between linguisticstructures and other mental structures. If one adopts Bouchard's model of conceptual structure then Assumption 6 immediately follows: her relative theta alignment hypothesis implies a close alignment of conceptual and syntactic structure. On the other hand, Jackendoff would seem to like Assumption 6 to hold for his conceptual grammar, but it is not clear whether it does (my suspicion is that it does **not**).

Finally, Assumption 7 follows from the ideas of Banks, as quoted above; it also fits in nicely with Bouchard's model, in which semantic entities and syntactic entities are often roughly interchangeable. Indeed, phenomenologically speaking, this assumption is almost obvious: one may easily work out one part of thought in verbal form while leaving the other part temporarily vague, not explicitly linguistic.

9.4 L-SYSTEMS AND SENTENCE PRODUCTION

In this section, I will model sentence production as an iterative process which involves three key subprocesses:

1) expanding linguistic forms into more precise linguistic forms, using L-system rewrite rules

2) substituting words for other mental structures, and

3) seeking at all times to minimize sentence length and complexity

Assumptions 1 and 2 represent simple magician dynamics; assumption 3 is a constraint on magician dynamics which is posed by the fact that we are dealing the **pattern/process** magicians. Long, complex sentences are unlikely to be patterns in situations; short situations are more likely to be.

The model is a complicated recursive process, and is not well depicted in words; thus I will present it in a kind of "pseudocode" as a collection of three interlinked procedures:

sentence **produce**(mental structure T)

First, choose a basic linguistic form with which to express the structure T. This form need not be filled out with words, it may be filled out with mental structures. It is an "emphasis pattern" and "sentence pattern"; a "deep structure."

The basis for choosing a form is as follows: one chooses the form F which, after carrying out the process A = **wordify**(**expand** (F)), yields the sentence giving the best depiction of the structure T (as measured by the "metric" guaranteed by Assumption 5).

The process returns the sentence A produced when, in the "choice" phase, F was given to **wordify**(**expand**()) as an argument.

sentence **wordify**(linguistic form F)

Any words in linguistic form F are left alone. Any mental structures T in linguistic form F are replaced by words or phrases which are judged to matchthem. These words or phrases may be expanded by **expand** to improve their match to T; but the chance of calling on **expand** decreases sharply with the length and complexity of the sentence obtained.

linguistic form **expand**(linguistic form F)

This process acts on the all components of F, in parallel, using:

a) various rewrite rules f_i and g_i.

b) where appropriate, the process **wordify**

The choice of a) or b), and the choice of transformation rules within a), is partly pseudo-random and largely dependent on analogy with previously produced sentences.

For each rewrite rule f_i or g_i that one applies, one obtains a new linguistic form G. The acceptability of the result G is judged by the following criterion: Does the greater accuracy of depiction obtained by going from F to G outweigh the increased complexity obtained by going from F to G?

In determining the accuracy of depiction and complexity obtained in going from F to G, one is allowed to apply **expand** again to G; but the chance of doing this decreases sharply with the length and complexity of the sentence involved.

Notice that the recursivity of **expand** and **wordify** could easily lead to an endless loop (and, in computational models, a stack overflow!) were it not for the stipulated sharp decrease in the probability of recursion as sentence length increases. Most sentences are not all that long or complex. An unusually high degree of elaboration can be seen, for example, in the sentences of Marcel Proust's *Remembrance of Things Past*, which sometimes contain dozens of different clauses. But this kind of elaboration is both difficult to produce, and difficult to understand. As a rough estimate, one might say that, in the process of producing a typical sentence, none of these processes will be called more than 5 - 20 times.

The purely linguistic interpretation of the model assumes that the process **produce** will be given a well-worked out idea to begin with, so that no further use of the conceptual rewrite rules g_i will be required. The grammatical-cognitive interpretation assumes that the argument of **produce** is vague, so that applications of the g_i must coexist with applications of the f_i, producing grammatical forms and clarifying ideas at the same time. More and more evidence is emerging in favor of the interpenetration of thought and language, a trend which favors the grammatical-cognitive interpretation. However, as stated above, the L-system model itself is independent of this issue.

I have presented this model of production as a set of procedures, but this communicative device should not be taken to imply that the psychological processes involved are necessarily executed by rule-based procedures in some high-level language. They could just as well be executed by neural networks or any other kind of dynamical system.

For instance, it is particularly easy to see how these linguistic processes could be carried out by a **magician system**. One need only postulate "wordification" magicians for transforming mental structures into words or phrases, and "transformation" magicians for transforming words

into phrases. The above sentence production algorithm then becomes a process of applying elements of an magician population to one's basic linguistic form, while retaining at each stage copies of previous versions of one's linguistic form, in case current magician activities prove counterproductive. This interpretation fits in nicely with the psynet model. The focus here, however, is on the language production process itself and not its implementation as an autopoietic magician system.

A Simple Example

To better get across the flavor of the model, in this section I will give a very simple "thought-experiment" regarding a particular sentence. I will show how, using the L-system model, this sentence **might** be produced. Consider the sentence:

2) Little tiny Zeb kisses the very big pine tree

The first stage of the process of producing this sentence might be a mental structure expressible in terms of the basic linguistic form

NVN

and loosely describable as

(this baby here)(is touching with his mouth)(this tree here)

The entities in parentheses denote **mental structures** rather than linguistic structures.

Now, how is this basic form expanded into a sentence? Each of the three elements is transformed, **at the same time**, by calls to *wordify* and *expand*. Wordification turns the thoughts into words or phrases, and expansion turns the words or phrases into longer phrases.

For instance, one possible path toward production might be immediate wordification of all components. The *wordify* process might thus transform the form into the pre-sentence

Zeb kisses tree

Expansion rules then follow: first, perhaps,

N -> Det N

N -> Adj N

lead to

Little Zeb kisses the tree

The expansion process goes on: the rule

Adj -> Adj Adj

is applied twice to yield

Little tiny Zeb kisses the big tree

and finally

Adj -> Adv Adj

is applied, giving the target sentence.

An account such as this ignores the "trial and error" aspect of sentence production. Generally "dead-end" possibilities will be considered, say

V -> Adv V

leading to

Little Zeb really kisses the tree

But these alternatives are rejected because they do not add sufficient detail. The addition of *really*, it may be judged, does not make the sentence all that much more similar to the original mental form -- it does not add enough similarity to compensate for the added complexity of the sentence.

This example account of development follows the pure grammatical interpretation of the L-system model, in which expansion takes place only on the linguistic level. From the perspective of the grammatical-cognitive interpretation, however, this is only one among many possibilities. There is no need for the whole thing to be wordified at once. It may be that, instead, the development process follows a sequence such as

(this baby here)(is touching with his mouth) big tree

(this baby here)(is touching with his mouth) the big tree

Zeb (is touching with his mouth) the tree

Tiny Zeb (is touching with his mouth) the big tree

Tiny Zeb kisses the very big tree

Little tiny Zeb kisses the very big pine tree

Or, on the other hand, actual ramification may occur on the level of **conceptual** rather than **syntactic** structure:

(this baby here)(is touching with his mouth) big tree

(this little baby here)(is touching with his mouth) the big tree

(this really very very little baby here)

(is touching with his mouth) the tree

Little tiny Zeb (is touching with his mouth) the big tree

Little tiny Zeb kisses the very big tree

Little tiny Zeb kisses the very big pine tree

Under either interpretation, this is a very **messy** model: the whole sentence keeps expanding out in all directions at once, ateach time considering many possible expansions, accepting some and rejecting others. But this kind of haphazard, statistically structured growth is **precisely** the kind of growth that one sees in nature. It is exactly the kind of growth that one would expect to see in the productions of the language centers of the brain -- for the brain is, after all, a biological system!

Bracketed L-Systems, Reflexives, and Sentences

The L-system expansion process is completely parallel; it makes no reference to the linear order in which the sentence is to be understood by the listener/reader. In some cases, however, this linear order is crucial for sentence understanding. Consider, for example:

3) the city's destruction by itself

*3a) * itself's destruction by the city*

The only difference here is that, in the legal sentence *(3)*, the antecedent precedes the pronoun, while in the illegal sentence *(3a)*, the antecedent follows the pronoun. This reflects the general rule of English grammar which states that reflexives cannot have independent reference, but must take their reference from an antecedent which is compatible (e.g. in case, gender and number). When producing a sentence such as this, the substitution rules involved must be **order-dependent** in a complex way, just as the rule N -> Adj N is order-dependent in a simple way.

The phenomenon of reflexives may be very nicely understood in terms of **bracketed** L-systems. The idea is that the word *itself* is itself a "close-bracket" command, while the word *city* is follows by an implicit"open-bracket" command, so that *(3)* might be rewritten as

the city['s destruction by]

Upon encountering the "[" the listener pushes the current state, *city's*, onto the stack, and upon encountering the next "]" the listener pops this off the stack and inserts it in the appropriate position.

The trick, of course, is that the "[" is not explicitly there, so that what one really has is

the city's destruction by]

where the listener must determine the position of the "[" based on syntactic and semantic rules and semantic intuition. The listener must also determine precisely what "state" should be pushed onto the stack: generally this is the coherent phrase immediately following the "[", but the identification of phrase boundaries is not always obvious.

The **illegal** sentence *(3a)*, on the other hand, is characterized by a close-bracket which does not match any possible preceding open-bracket positions. In

]'s destruction by the city

there is simply no preceding position into which the "[" might be placed! If this phrase were embedded in a larger sentence there would be candidate positions for the "[" but, in all probability, none of these would be satisfactory, and the sentence would still be illegal.

A more complex construction such as

3) the city's terrible sewer system's destruction of itself's broadcasting of itself throughout the region through propagation of a terrible odor

can also be cast in a similar way, i.e.

the city's [terrible sewer system['s destruction of]'s broadcasting of] throughout the region through propagation of a terrible odor

But this is difficult to understand, reflecting no doubt the small stack size of human short term memory.

This view of reflexives is closely related to Kayne's (1981) proposal that a sentence must define a **unambiguous path** through semantic space. The bracketed L-system model makes this "unambiguous path" proposal more concrete by connecting it with stack automata. The use of bracketed L-systems here is somewhat different from their use in computer graphics models of plant growth -- but the underlying interpretation in terms of a stack is the same, and so the parallel is genuine. In each case the brackets are a **formal** way of indicating something which the relevant **biological** system does without any explicit, formal instructions.

9.5 L-SYSTEMS AND CHILD LANGUAGE DEVELOPMENT

The L-system model is a complex one, and detailed empirical validation will be a lengthy process. Some evidence supporting the model, however, may be found in the patterns of early childhood language. Young children use simpler transformation rules, and have a lower "complexity ceiling," factors which result in in much shorter and simpler utterances, in which the underlying production dynamics more apparent.

The beginning of early childhood grammar is the two-word sentence. One attempt to explain two-word utterances is based on the grammatical formula

S -> NP VP

VP -> V NP

But, as pointed out by Anisfeld (1984), this formula is psychologically problematic. According to this formula, the V and the NP would seem to be tied more closely together than the NP and VP; but this is not what the experiments reveal. In fact the NP and the V are the more closely associated pair.

This "problem" is resolved immediately by the L-systemmodel. The transformational grammar approach suggests the iteration

0 NP VP

1 NP V NP

driven by the substitution rule VP -> V NP. The L-system approach, on the other hand, suggests that the mental structure playing the role of the generation 0 VP is tentatively represented by a verb V, and the mental structure playing the role of the generation 0 NP is tentatively represented by a noun N. Then the verb is expanded into a verb phrase. Thus the iteration is better represented

0 N V

1 N V N

driven by the rule V -> V N. The reason N V is a more natural combination is because it occurs at an earlier step in the derivation process.

More concrete evidence for the L-system view of child language is given by the phenomenon of **replacement utterances**. Braine (1971) observed that, in recordings he made of the speech of children 24-30 months old, 30-40% of the utterances produced were "replacement utterances," or sequences of more and more complex utterances spoken one after another in response to the same situation. The following are typical examples:

Stevie byebye car.

Mommy take Stevie byebye car.

Stevie soldier up.

Make Stevie soldier up, Mommy.

Car on machine.

Big car on machine.

Stand up.

Cat stand up.

Cat stand up table.

According to Anisfeld (1984), the children are unable to construct the final, expanded forms all at once, so they resort to a process of gradual construction. The initial utterance represents the most important information that the child wants to transmit -- usually a predicate but occasionally a subject. Later utterances add on more details.

Viewed from the perspective of the L-system model, these "replacement sequences" become a kind of **window** on the process of sentence production. What these children are doing, the model suggests, is vocalizing a process that we all go through when producing a sentence. The utterances in a replacement sequence all express the same underlying mental structure. But, as the speaker proceeds through the sequence, the constraint on sentence length and complexity gradually becomes more relaxed, so that theiterative process of sentence production is cut off at a later and later point.

For instance, the final example given above, "Cat stand up table," has a form that can be roughly given as

0 V

1 N V

2 N V N

(it is assumed that "stand up," to the child, is a single predicate unit). When the initial utterance "Stand up" is produced the need for brevity is so great that additional transformations are summarily rejected. The next time around, a number of possible transformation suggest themselves, such as V -> N V or V -> V N. Each one has a certain probability based on its usefulness for describing the situation as balanced against the complexity it adds to the sentence. The transformation V -> N V is selected. At the next stage, then, further transformations are suggested, say perhaps N -> Adj N or V -> Adv V, in addition to the two which arose at the previous stage. One of these is selected, according to the stochastic, parallel progressive

substitution process, and then, since the length and complexity threshold has been reached, the sentence is produced.

One final comment is in order. I have used the familiar categories N and V for these examples, but the L-system model is not restricted to these familiar categories. Braine (1976) has argued that the transformational approach to childhood grammar is misguided. Instead of using general grammatical rules, he suggests, young children produce utterances using an overlapping collection of narrowly constrained special-case rules. Evidence for this view is provided by the failure of linguists to provide a grammatical account of beginning speakers' two-word utterances. These ideas may contradict some of the narrower interpretations of transformational grammar, but are unproblematic for the L-system approach. An overlapping collection of special-case rules can be perfectly well captured by a nondeterministic, stochastic L-system (which is, broadly speaking, a kind of formal grammar). Children may well begin with special-case substitution rules, and gradually abstract more general rules involving categories like N and V. The question of **whether** there is a grammar is separate from the question of the **specificity** of the substitution rules.

CHAPTER TEN

DREAM DYNAMICS

10.1 INTRODUCTION

Scientifically as well as experientially, dreaming is a confusing phenomenon. It is known to be necessary for healthy functioning, in the sense that dream deprivation leads to harmful effects. Yet the biological and psychological reasons for this necessity are largely unknown.

Clinical psychologists, since well before Freud, have believed dreams to hold deep secrets about the unconscious mind. Jung believed even more: that they were a doorway to the collective unconscious. The importance of dreams is affirmed by many ancient wisdom traditions. On the other hand, recent theorists have proposed that the contents of dreams are just "random noise," generated by primitive parts of the brain.

In recent years, however, it has become possible to simulate some aspects of dreaming behavior using formal neural networks. In this chapter I will review some of this work -- the Crick- Mitchison hypothesis and its extensions in the work of George Christos. I will discuss the shortcomings of this work, and propose an alternate approach based on the dynamics of mental **processes**. The new approach proposed here salvages many of the ideas of the neural network approach, while also tying in neatly with the older, content-oriented theories of dreams.

The Crick-Mitchison Hypothesis

In 1983 Crick and Mitchison coined the catch-phrase, "We dream in order to forget." This audacious statement, now called the Crick-Mitchison hypothesis, was primarily formulated not on the basis of neurological experiments, but on the basis of experience with a simple formal network model called the **Hopfield net** (Hopfield, 1985). Crick and Mitchison observed that Hopfield nets, when used as associative memories, quickly become overloaded with useless old memories. Why, they asked, does the human brain not suffer from the same problem? The answer they arrived at was: because the human brain dreams.

In a 1986 paper, Crick and Mitchison retracted their original statement, replacing it with the weaker "We dream to reduce fantasy and obsession." However, some authors, such as Christos (1994), feel that the original Crick-Mitchison formulation was more accurate. Computer simulations by Christos (1992) and others (Nadel et al, 1986; Parisi, 1986) give partial support to the original hypothesis. By passing a Hopfield net through alternate phases of learning and forgetting, one does indeed solve the problem of overloading, though at the cost ofdrastically reducing the memory capacity of the network.

Neural network models of dreaming are typically interpreted to contradict the psychoanalytic account of dreams, according to which dreams serve deep emotional functions (Freud, 1900). I will argue, however, that there is no such contradiction. If one transplants the idea of "dreaming as forgetting" to the level of **neuronal groups** instead of neurons -- i.e., the level of complex process dynamics instead of merely switching network dynamics -- the apparent contradiction disappears. One obtains a theory that is quite consistent with the Freudian view of dreams, as well as being, in some respects, less conceptually problematic than the original Hopfield net theory. This new and improved theory of dreaming behavior, while speculative, is conceptually quite satisfying. By invoking a "process dynamics" view of the mind, it closes the gap between a physicalistic model of dreams (the Hopfield net model) and a mentalistic model of dreams (the Freudian model).

10.2 TESTING THE CRICK-MITCHISON HYPOTHESIS

According to the most reasonable estimates, the brain takes in about 100 billion billion bits of information during an average lifetime. But the brain can **store** only 100 thousand billion bits at a time. Clearly a great deal of information compression is going on in the brain, so that these numbers are not strictly comparable. One cannot justifiably conclude that only one out of every million pieces of information can be "permanently" retained. But nevertheless, it seems clear that the brain has a large incentive for forgetting irrelevant information. A very important part of learning and remembering is knowing what to forget.

But how does the brain do this forgetting? Crick and Mitchison were the first to suggest that **dreaming** might play a role in this process. To see the inspiration for this idea one must think a little about the limitations of Hopfield nets as associative memories.

Hopfield Nets

The Hopfield net is perhaps the simplest formal neural network model of any practical use (Hopfield, 1985). It is a natural elaboration of the original neural network model of McCullough

and Pitts (1943). In the simplest case, each neuron has a state of either 1 or -1, and each neuron accepts input, potentially, from all other neurons. At each time step, each neuron sums up all its inputs, arriving at a certain "internal sum." The input which neuron i receives from neuron j is multiplied by amount w_{ij} before it is added onto the internal sum of neuron i. Then, after it has constructed its internal sum, a neuron tests whether this sum is over its threshold or not. If so, it sends out charge to all the other neurons. If not, it doesn't. In the simplest model, before the neuron fires again, it wipes its memory clean, sets its internal sum to zero.

The firing of neurons is the dynamics of the network. Thisdynamics may be managed in at least two different ways: asynchronously, whereby the neurons fire one after the other, in some random order; or synchronously, whereby all neurons fire at the same time. Under either scheme, the dynamics may involve a variety of complex circuits. In many applications, however, the weights are taken symmetric ($w_{ij} = w_{ji}$), which guarantees that, from any initial condition, the network will eventually converge to a fixed point attractor (this is not obvious, but may be shown by algebraic manipulations). Relaxing the symmetry conditions gives the full spectrum of dynamical behavior -- periodic points, limit cycles, etc.

To view a Hopfield network as an associative memory, one assumes one is given a series of "memories" $A_1,...,A_M$, each one of which is presented as a vector of -1's and 1's. The synaptic weights are then defined in terms of these memories. To find the correct w_{ij}, for each pair (i,j) of potentially interconnected neurons, one sums up the products $A_{ki}\ A_{kj}$ for all the memories vectors A_k. If all goes well, then this choice of weights will lead to a network which has the desired memories as fixed-point attractors.

The catch is that, like a human brain, a Hopfield net can only store so much. Once the number of items stored exceeds about 15% of the number of neurons, the memory ceases to function so efficiently, So-called "parasitic memories" emerge -- memories which are combinations of **parts** of real memories, and which are fallaciously associated with a number of different inputs.

So, suppose the Hopfield net in question is part of a living, learning system, which constantly needs to store new information. Then the network will need to have a way of unlearning old associations, of forgetting less crucial memories so as to make room for the new ones. This, Crick and Mitchison suggest, is the role of dreaming. Instead of **adding** terms $A_{ki}\ A_{kj}$ **on** the synaptic weights, as one does to train the network, dreaming **subtracts** these products, multiplied by a suitable constant factor, from the synaptic weights.

But which memories get their contribution to the weights decreased in this manner? Why, of course, **the ones that are remembered most often!** If a **real** memory is remembered often, and one decreases its weight a little bit, this won't really hurt it, because it has a deep basin. But if a memory is remembered often simply because it is **parasitic**, then a slight decrease in the weight will destroy it -- because its basin is so shallow. Or so the theory goes.

Hopfield, independently of Crick and Mitchison, did some simple reverse-learning experiments with a computer-simulated network. He filled the network just slightly above its capacity, and then tried to winnow it down, by feeding it random inputs, observing its responses,

and **subtracting** some percentage of the memories thus obtained. In order to avoid completely obliterating the memory, the percentage had to be taken very small, around 1%.

The First Christos Experiment

Intrigued by the reverse-learning approach to dreams, West Australian mathematician George Christos decided to put the Crick-Mitchison hypothesis to the test (Christos, 1992). In his first experiment, Christos simulated a repeatedly dreaming Hopfield network by "random sample analysis." At each stage, to determine what the network should dream about, he presented it with a large number of random inputs (100-200 times the number of stored memories), and observed what memories the network associated these inputs with. This statistical procedure gives a good qualitative picture of the "energy landscape" of the network. Those memories that are retrieved more often must have larger basins of attraction, and are therefore the ones that should be removed by "reverse-learning" or dreaming.

The results of this experiment were somewhat surprising. At first things work just like the Crick-Mitchison hypothesis suggests: dreaming reduces the incidence of spurious, parasitic memories. But then, as the network continues to dream, the percentage of spurious memories begins to **increase** once again. Before long the dreaming destroys **all** the stored memories, and the network responds to all inputs with **false memories only**.

For instance, in a network with five memories, using a weight-reduction factor of 1%, the first fifty dreams behave according to Crick-Mitchison. But by the time **five hundred dreams** have passed, the network is babbling nonsense. The same pattern is observed in larger networks, and with different reduction percentages. Intuitively one might think that dreaming would tend to even out the basins of attraction of the stored memories. But, as Christos put it, "intuition is not guaranteed in a nonlinear process such as this."

Dreaming, removing part of a memory "attractor," may actually cause **new** spurious attractors, **new** false memories. In this sense the Crick-Mitchison hypothesis is based on a conceptual fallacy: it is based on the idea that one can remove a **single false memory** from a Hopfield net, in a localistic way. But in fact, just as false memories **emerge** from holistic dynamics, attempts to remove false memories **give rise** to holistic dynamics. The reverse-learning mechanism inevitably affects the behavior of the network as a whole. Linear intuition does not apply to nonlinear dynamics.

The Second Christos Experiment

But this is not the end of the story. Human dreaming is obviously a **cyclic** phenomenon: we dream, then we wake, then we dream, then we wake, and so on. For his second experiment, Christos sought to simulate this process with a Hopfield net. He trained the network with

memories, then put it through the reverse-learning, "forgetting" process. Then he trained it with new memories, and put it through the forgetting process again. And so on.

The experiment was a success, in the sense that the network continues to function as an associative memory long after its"fifteen percent limit" has been reached. Dreaming does, as Crick and Mitchison predicted, allow the network to function in **real-time** without overloading. Since the older memories are subject to more dreams, they tend to be forgotten more emphatically, making room for new memories.

The only catch is that the overall memory capacity of the network is drastically reduced. A network normally capable of storing twenty memories can now only store three or four, and even for these it has a low retrieval rate. The reason is that, just as in the first Christos experiment, dreaming actually **increases** the production of false memories.

So the Crick-Mitchison hypothesis is salvaged, but just barely. If one keeps the number of memories very, very small -- around .05 of the number of neurons in the network -- then the cycle of learning and forgetting is indeed effective. Given the huge size of the brain (100 billion neurons is a reasonable figure), the number .05 is not particularly worrying from the neuroscience point of view. A network with intermittent periods of dreaming and learning provides much less efficient memory than a network which is simply re-started every time it gets full, but real brains may not have the option of periodically wiping themselves, or even parts of themselves clean..

10.3 A MENTAL PROCESS NETWORK APPROACH

So, is dreaming really forgetting? Based merely on these Hopfield net simulations, this seems a rather grandiose conclusion, which is perhaps why Crick and Mitchison eventually reformulated their statement in a weaker form.

It is my contention, however, that there is actually some **psychological** sense to be found in the idea that dreaming is forgetting. This sense may be found by transplanting the idea from the domain of Hopfield nets to the domain of **neural maps**. A neural map is a network formed from neuronal groups rather than single neurons; it is thus a network of complex processes rather than a network of simple mathematical functions.

In some ways, I will argue, the "dreaming is forgetting" idea makes **more** sense in a neural map context than it does in terms of Hopfield nets. In the domain of neural maps, there is no problem with the forgetting of useful instead of useless memories, a fact which leads one to suspect that the memory capacity limitations of the dreaming Hopfield network may not apply to the brain. Furthermore, the transplantation of the theory to the domain of neural maps leads to a connection with the psychoanalytic theory of dreams. It explains how dreaming might accomplish both the role of forgetting **and** the role of highlighting or resolving emotional problems.

This account of dreaming is speculative but, in truth, no more so than the original Crick-Mitchison hypothesis, the Freudian theory, or any other current theory of dreaming behavior.

The key asset of the present theory is its ability to bridge the gap between the physicalistic dynamics of neurons and the mentalistic dynamics of thoughts and emotions. The processdynamics of neuronal groups are used as a "middle ground" between these different levels of explanation.

Memory and Consciousness

One thing sorely lacking in the Crick-Mitchison hypothesis is any reference to the phenomenon of **consciousness**. The peculiar feature of human dreaming, after all, is that when one dreams one is **conscious while sleeping**. In a sense, the absence of consciousness from the Hopfield net automatically makes it inadequate as a model of dreaming behavior.

It was argued in Chapter Eight that one of the psychological roles of consciousness is the **creation** of memories, as wide- basined autopoietic magician systems. To understand the relation between this aspect of consciousness and dreaming, recall the classic experiments of the neurologist Wilder Penfield. Penfield noted that by stimulating a person's temporal lobe, one can cause her to **relive memories of the past**. Merely by touching a spot in the brain, Penfield could cause a person to have a detailed, intricate experience of walking along the beach with sand blowing in their face, or lying in bed in the morning rubbing their eyes and listening to the radio. The most remarkable aspect of these memory experiences was their **detail** -- the person would recite information about the pattern of the wallpaper, or the color of someone's shoes ... information which, in the ordinary course of things, no one would bother to remember. And if he touched the same spot again -- the same experience would emerge again.

Penfield's conclusion was that the brain stores far more information than it ever uses. Every moment of our life, he declared, is filed away somewhere in the brain. If we only knew how to access the relevant data, we could remember everything that every happened to us with remarkable accuracy.

Recent replications of Penfield's experiments, however, cast serious doubt on this interpretation. First of all, the phenomenon doesn't occur with everyone; maybe one in five people can be caused to relive the past through temporal lobe stimulation. And even for this select group, a careful analysis of the "relived memories" suggests that they are **not exact memories at all**.

Current thinking (Rosenfield, 1988) is that, while the basic frameworks of the relived memories are indeed derived from the past, the details are **manufactured**. They are manufactured to serve current emotional needs, and based on cues such as the places the person has visited earlier that day. This, of course, ties in perfectly with the view of consciousness in terms of a memory-creating Perceptual-Cognitive Loop.

And what distinguishes those people who are susceptible to the Penfield phenomenon? The answer, it appears, is **limbic system activity**. The limbic system is one of the older portions of the brain -- it is the "reptile brain" which is responsible for basic emotions like fear, rage and lust. When a susceptible person has their temporal lobe stimulated, the limbic system is activated. But for a non-susceptible person, this is not thecase; the limbic system remains

uninvolved. The message seems to be that **emotion is necessary for the construction of memory**. This is a very Freudian idea, but nonetheless it is supported by the latest in neurological research.

Certain people, with a particularly well reinforced pathway between the limbic system and the temporal lobes, can be caused to construct detailed memories by temporal lobe stimulation. Penfield's original conclusion was that the brain stores **all its experience**, like a vast computer database. But the fact of the matter seems to be that the brain, quite **unlike** any computer database known today, holds the fragments of a memory in many different places; and to piece them together, conscious, emotional intervention is required.

Memory and Dream

So consciousness **produces** memories. What does this have to say about dreams?

When **awake**, consciousness is constrained by the goal- directed needs of the brain's perceptual-motor hierarchy. It pieces together fragments of different ideas, but always in the service of some particular objective. This means that, unless they are almost omnipresent, false memories will quickly be rejected, due to their failure to solve the practical problems at hand.

When the body is **sleeping**, on the other hand, consciousness can act unfettered by the needs of the perceptual and motor systems. It has much less lower-level guidance regarding the manner in which it combines fragments from the mind's associative memory. Therefore it will tend to produce false memories just as readily as real ones. In particular, it will combine things based on how well they "go together" in the associative memory, rather than how well they serve pragmatic goals.

In other words, to use the language of the psynet model, what I am suggesting is that dreaming is inefficient in the Hopfield network precisely because the Hopfield network is merely a static associative memory, rather than a dynamic associative memory coupled with an hierarchical perception/control network. In the human brain, during waking hours, the perceptual-motor network reinforces real memories due to their greater utility. During sleeping hours, dreaming decreases real and false memories alike, but as the false memories do not have so much waking reinforcement, they are eventually obliterated.

Furthermore, in the brain, old memories are constantly interacting and **recalling one another**. Old memories do not necessarily fade just because they are not explicitly elicited by external stimuli. This is one very important role of **feedback** in the brain: memories activate one another in circular networks, thus preventing the dreaming dynamic of reverse learning from chipping away at them.

Dreaming as Forgetting, Revisited

To put these abstract ideas in a more concrete context, let us turn to the psynet model. In biological terms, this means one is stepping up from individual neurons, as in the Crick-Mitchison hypothesis, to **neuronal groups** which represent pattern/process magicians.

In the psynet perspective, memories are autopoietic systems, so it is easy to see that many useless memories will naturally "forget themselves." Thought systems with no positive external input will often just disapppear. In biological terms, one may say that, if a certain cell assembly is of no use, it will simply not be reinforced ... it will dissolve, and other more useful assemblies will take its place.

But of course, this does not always happen. Some thought systems, like political belief systems, may proceed **of their own accord**; continuing to exist in perpetuity because of their own autopoiesis. Neurally speaking, one may say that, in the more abstract regions of the mind, maps which are of no **objective** use can sustain themselves by **self-reinforcing dynamics**, by map-level feedback loops. These will not be forgotten by simple neglect.

This leads us to our main idea. Suppose that dreaming served as a special dynamic for **forgetting** useless, self- perpetuating thought systems? This hypothesis forms a connection between the "dreaming as forgetting" idea and the psychoanalytic theory of dreams, in which dreams represent unresolved problems and anxieties. All that is needed to draw the connection is the association between neuroses or "complexes" and complex, self- perpetuating systems of neuronal groups.

While sleeping, consciousness is **present**, and yet it is **dissociated** from the ordinary external world. In place of the external world, one has a somewhat sloppily constructed **simulacrum** of reality. And on what principles is this simulacrum created? The main difference between dream and reality is that, in the dream-world, **expectations are usually correct**. Whatever the controlling thought-system foresees or speculates, actually happens. Thus the disjointed nature of dream life -- and the peculiarly satisfying nature of dreams. In dreams, we can get **exactly** what we want ... something that very rarely happens in reality. And we can also get **precisely** what we most fear. The image in the mind instantaneously transforms into the simulated perception.

In dreams, in short, thought-systems get to construct their own **optimal input**. This observation, though somewhat obvious, leads to a rather striking conclusion. Suppose a thought-system has evolved a circular reinforcement-structure, as a tool for **survival** in a hostile world -- for survival in an environment that is constantly threatening the system with destruction by **not** conforming with its expectations. What will be the effect on this thought-system of a simulated reality which **does** conform to its expectations?

This question can be answered mathematically, using the formalism of magician systems. But this is hardly necessary; the upshot is easily perceived. A thought-system, presented with acomfortable world, will **let down its guard**. It will relax its circular reinforcement dynamics a little -- because, in the dream world, it will no longer need them. The dream world is constructed **precisely** so that the thought-system will be reinforced **naturally**, without need for circularity. Thus dreaming acts to **decrease** the self-reinforcement of circular belief systems. It **weakens** them ... and thus, after a fashion, serves as a medium for their "forgetting."

Let's say a married man has a dream about his wife strutting down the street in a skimpy dress, swaying her hips back and forth, loudly announcing that she's looking for a good time. Meanwhile he's running after her trying to stop her, but his legs won't seem to move fast enough;

they suddenly feel like they're made of glue. The interpretation is obvious enough, but what is the **purpose**?

According to the present model, the purpose is precisely to feed his suspicious thought-system what it wants to hear. Apparently a certain component of his mind is very mistrustful of his wife; it suspects her, if not of actually **committing** adultery, at least of possessing the **desire** to do so. This mental subsystem may have no foundation in reality, or very little foundation; it may to a great degree sustain itself. It may produce its own evidence -- causing him to interpret certain behaviors as flirtatious, to mis-read certain tones of voice, etc.

The thought system, if it is unfounded, **has** to be circular in order to survive daily reality. But in the dream world it is absolutely correct: she really **is** looking to have sex with someone else. While getting input from the dream world, then, the thought system can **stop** producing its own evidence. It can get used to receiving evidence from the **outside** world.

Temporarily, at least, the dream-world breaks up the circularity of the thought system, by removing the **need** for it to make up its own evidence. Whether the circularity will later restore itself is another question; but one may say that the **expected** amount of circularity in the system will be less than it would have been had the dream **not** occurred. In other words, his suspicions, having temporarily had the liberty of surviving without having to create mis-perceptions, may forego the creation of mis-perceptions for a while. And in this way, perhaps they will begin to fade away entirely. Or, alternately, the thought- system may rejuvenate itself immediately, in which case he can expect to have similar dreams again and again and again.

The big question raised by the second Christos experiment is, how could a neural network be configured to tell **real** memories from **false** ones? How could one get it to "subtract off" only the bad guys? If this problem were solved, then the extremely low storage capacity of dreaming Hopfield nets would be solved. But in the mental process network theory, there is no problem with this. If a **useful** thought system is subjected to dream treatment, and given a simulated reality which fulfills all its expectations, it will not be **jarred** as much as a lousythought-system, which relies more centrally on self-reinforcement for its survival. Its dream world will not be as drastically different from the **real** world. It will get in the habit of **not** producing its own evidence ... but then, it never **had** this habit to an excessive degree in the first place.

Let's say the same married man mentioned above has a dream about his wife kissing him passionately and giving him a new gold watch with a huge, fist-sized diamond on it. This is produced by the thought-system that knows his wife loves him and treats him well. It temporarily permits this thought system to relax its defenses, and stop actively "twisting" reality to suit its preconceptions. But in fact, this thought system **never** twisted reality nearly so much as its suspicious competitor. So the dream has little power to change its condition. Dreams will, according to this line of reasoning, affect **predominantly circular** belief systems much more drastically than they will affect those belief systems which achieve survival mainly through interaction with the outside world.

And what about the role of the **body** in dreaming? One must recall that **perception** and **action** are inseparable -- they are carried out by a unified perceptual-motor hierarchy. Or, in the

aphorism of Heinz von Foerster, "If you want to see, learn how to act." To fool a thought-system into thinking it is perceiving its ideal environment, one must often feed it perceptions that involve **actions** as well.

It is disappointing to realize that the Hopfield net is too unstructured to serve as a useful psychological model of dreaming. For, after all, the Hopfield network is so easy to analyze formally, and to simulate on the computer. When one starts talking about consciousness and inter-reinforcing networks of memories, things become much murkier, much less elegant and mathematical. But on the other hand, perhaps that's just the way mind **is** -- ....

10.4 DREAMING AND CRIB DEATH

One concrete, if speculative, application of these ideas is to the phenomenon of Sudden Infant Death Syndrome (also known as SIDS, or simply "crib death" or "cot death"). Nearly two in every thousand births result in a death classified as SIDS. A huge amount of biological data has been gathered regarding this "disease," but even so, researchers have been unable to arrive at a good explanation. George Christos has proposed an intriguing explanation based on **dreaming**.

It has been suggested that respiratory obstruction is the culprit, or else reduced circulation in the neck arteries. But these theories fail to account for the observed information: why, then, do infants sleeping on their back also die from SIDS? George Christos (1992a) has suggested that the flaw is not in the data but rather in the **physicalistic** bias of medical researchers. He suggests that SIDS is fundamentally a **psychological** problem, the understanding of which requires an analysis of the **infantmind**.

During ordinary, non-dreaming sleep, the brain and the body are fairly well "disconnected." But during dreaming sleep, this disconnection becomes more extreme. This fact has long been used by researchers to study the effect of dream deprivation in cats: the cats are placed on tiny steeply sloped "islands" in a room full of water. They can sleep on the islands, but when they sink into dreaming sleep, their muscles relax even further and they slip into the water.

The results of the **failure** of this brain-body disconnection are well-known. Walking and talking in one's sleep are two prime examples. One does not **want** the brain hooked up to the body during dream sleep; otherwise the dreams will control the body's motions, resulting in bizarre and dangerous activity.

But even during dreaming sleep, however, some degree of brain-body connection remains. Active contact is maintained with the muscles regulating the eyes, the heart and the lungs. And it is **these** connections, Christos claims, which are responsible for Sudden Infant Death Syndrome.

In the Stanford University sleep lab, an experimental subject was observed to actually **hold his breath** while dreaming about swimming underwater. Similarly, Christos proposes, SIDS occurs when infants **dream about being in the womb** and, consequently **forget about breathing**.

We dream about our past experiences. But what past experiences has an infant had? It is almost inarguable that, when an infant dreams during its first few weeks after birth, it dreams

about the first nine months of its life: its life as a fetus. And this is eminently sensible in terms of the dreaming- as-forgetting hypothesis, since an infant really does need to unlearn its memories of the womb. They are not very useful in its new, independent life.

But if the body does tend to "enact" the contents of its dreams as much possible, then when the infant dreams about the womb, its body will re-enact its fetal state. And one distinctive quality of the fetal state is the fact that breathing, in the usual sense, does not occur. The hypothesis is then that cot death is caused by infants dreaming that they are back in the womb and consequently forgetting to breathe.

This simple idea, Christos argues, explains all the circumstantial data regarding cot death: the exponentially decaying age at death curve, the higher risk in the prone sleeping position, and the climatic variation in incidence.

SIDS infants also demonstrate a greater amount of dreaming sleep than other infants, and a reduced motility during dreaming sleep. Furthermore, most SIDS deaths occur in the early morning, which is precisely the time of the **longest** dreaming sleep, corresponding to the most intense dreams. SIDS infants demonstrate a higher than average heart rate, meaning that their heart rate is closer to the **fetal** heart rate. And, most significantly of all, SIDS infants who survive past one month demonstrate a higher heart rate **only during dreaming sleep**.

There is, in this view, no cure for SIDS -- the best that a parent can do is to make their infant's life as "un-fetus-like"as possible. Avoid wrapping the baby tightly, and put her to sleep face up. This is not a terribly comforting view of crib death, and it does not provide funding agencies with much cause to support research work on the alleviation of crib death. Instead, it suggests, funds would be better spent on research into the biological and psychological foundations of dreaming.

SIDS and the Mental Process Network

Given the mental-process-network theory of dreams outlined above, one can easily form a plausible story of the **psychological** events underlying Christos's proposed mechanism. When the infant emerges from the womb, her neural patterns are still configured for **womb life**. She has developed a rudimentary "belief system" for dealing with life as a fetus. This system must be **destroyed**, or at least tranquilized, as quickly as possible. The tool for this destruction is the **dream**: dreams of fetal bliss fool the womb-life thought-system into complacency, making it easier for the new, outside-world-life thought-systems to grow and prosper.

But in order to most effectively fool the old thought- systems into thinking they're in their ideal environment, the womb, certain **actions** must also be re-enacted (after all, the thought-system perceives what the **body** is doing, and will recognize if something isn't right). Sudden infant death occurs because this process gets out of hand -- in other words, it is a consequence of **overenthusiasm** for life outside the womb, of an attempt at overly rapid adjustment.

This is, of course, a speculation -- as are all other proposed explanations of SIDS, at this stage. But, at very least, it is a nice illustration of how the **dreaming is forgetting** idea can be lifted

from the Hopfield network context, and transplanted into a Neural Darwinist, psynet-style, process dynamics view of the mind.

CHAPTER ELEVEN

ARTIFICIAL SELFHOOD

11.1 INTRODUCTION

The psynet model portrays the mind as a seething "soup" of intertransforming mental processes, self-organized into a spatial distribution. These "magician" processes lock into various attractors, which adapt themselves to each other, and to external circumstances. Low-level autopoietic process systems bind together into a meta-attractor, the dual network.

The dual network is a far more sophisticated emergent structure than any current AI program has been able to manifest -- but it is not the end of the line by any means. The dual network itself interacts with and supports a variety of other dynamics and structures. One such dynamic, mindspace curvature, was introduced in Chapter Four. Here, thinking along similar lines, we will introduce an abstract **structure** that is hypothesized to emerge from the dual network: the **self**.

By "self" what I mean is "psychosocial self" -- that is, a mind's image of itself as a part of the external world. There are other notions of self, e.g. the "higher self" of various spiritual traditions, but these are not at issue here. A self, in the sense I mean it here, is formed through observing one's interactions with the world, and it is also a tool for interacting with the world. The self is just one among many autopoietic systems within the dual network, but it assumes particular importance because of its sheer size. It extends from the lowest levels, physical sensations, to very high levels of abstract feeling and intuition. And it is also widely extended through the various levels of the heterarchical network, pertaining to thoughts, perceptions and actions in all sorts of different domains. The self is very often engaged with consciousness, which means that it exercises a strong effect on the way things are "bundled" up in memory.

Self is rarely discussed in the context of AI, but I believe that it should be. I will argue that self is necessary in order that a system adequately represent knowledge in terms of autopoietic systems. Thus, until we have "artificial selfhood," we will never have true artificial intelligence. And in order to induce artificial selfhood, it may be necessary to give our AI agents a **social context** -- to pursue artificially intelligent artificial life, or A-IS, artificial intersubjectivity.

Later chapters will pick up on these themes by considering the notion of **multiple** subselves, and explaining human personality phenomena in terms of subself dynamics. Then, in the final chapter, AI will be drawn back into the picture, and the precise relation of human and machine creativity to selves and subselves will be explored.

11.2 AUTOPOIESIS AND KNOWLEDGE REPRESENTATION

"Knowledge representation" is a key word in AI. It usually refers to the construction of explicit formal data structures for encapsulating knowledge. The psynet model represents a fundamentally different approach to knowledge representation. It contends that knowledge is stored in wide-basined autopoietic magician systems, rather than frames, objects, schemas, or other such formal constructions. These formal constructions may describe the contents of self-producing mental process systems, but they do not emulate the inherent creativity and flexibility of the systems that they describe.

The single quality most lacking in current AI programs is the ability to go into a new situation and "get oriented." This is what is sometimes called the brittleness problem. Our AI programs, however intelligent in their specialized domains, do not know how to construct the representations that would allow them to apply their acumen to new situations. This general knack for "getting oriented" is something which humans acquire at a very early age. It is something that current AI programs lack, due to their brittle, "dead," non-autopoietic systems for knowledge representation.

As a "straw man" example of the inflexibility of AI programs, consider Herbert Simon's famous "computer-scientist" program, BACON. This program was inspired by Sir Francis Bacon, who viewed science as a matter of **recognizing patterns in tables of numerical data**. But, Sir Francis Bacon never appreciated the amount of imagination involved in gleaning patterns from scientific data; and the program BACON falls into the same trap, albeit far more embarrassingly.

For instance, consider the "ideal gas law" from thermodynamics, which states that

$$pV/nT = 8.32$$

where p is the pressure of the gas, V is the volume of the gas, T is the temperature in degrees Kelvin, and n is the quantity of the gas in moles. In practice, this relation cannot be expected to hold **exactly**, but for most real gasses it is a very good approximation.

Given an appropriate table of numbers, BACON was able to **induce** this law, using rules such as:

If two columns of data increase together, or decrease together, then consider their quotient.

If one column of data increases, while another decreases, then consider their product.

Given a column of data, check if it has a constant value

As pressure goes up, volume goes down, so BACON forms the product pV. Next, as the **combined** quantity pV goes up, so does thetemperature -- thus BACON constructs the quotient pV/T. And as pV/T goes up, so does the number of moles -- hence the quotient (pV/T)/n = pV/nT is constructed. This quotient has a constant value of 8.32 -- so the ideal gas law is "discovered."

Very interesting, indeed. But how terribly far this is from what real scientists do! Most of the work of science is in determining **what kind of data to collect**, and figuring out creative experiments to obtain the data. Once a reliable set of data is there, finding the patterns is usually the **easiest** part. Often the pattern is **guessed** on the basis of terribly incomplete data -- and this intuitive guess is then used to guide the search for more complete data. But BACON is absolutely incapable of making an intuitive guess from sketchy data -- let alone figuring out what kind of data to collect, or designing a clever experiment.

Simon once claimed that a four-to-five hour run of BACON corresponds to "not more than one human scientific lifetime." Douglas Hofstadter, in *Metamagical Themas*, has sarcastically expressed his agreement with this: one run of BACON, he suggests, corresponds to about **one second** of a human scientist's life work. We suggest that Hofstadter's estimate, though perhaps a little skimpy, is much closer to the mark. Only a very small percentage of scientific work is composed of BACON-style data crunching.

A few AI researchers have attempted to circumvent this pervasive brittleness. Perhaps most impressively, Doug Lenat has developed a theory of **general heuristics** -- problem-solving rules that are abstract enough to apply to **any** context whatsoever. His programs AM and EURISKO applied these general heuristics to mathematics and science respectively; and both of these programs were moderately successful. For example, EURISKO won a naval fleet design contest two years in a row, until the rules were changed to prohibit computer programs from entering. And it also received a patent for designing a three-dimensional semiconductor junction.

But still, when looked at carefully, even EURISKO's triumphs appear simplistic and mechanical. Consider EURISKO's most impressive achievement, its invention of a 3-D semiconductor junction. The novelty here is that the two logic functions

"Not both A and B"

and

"A or B"

are both done **by the same junction**, the same device. One could build a 3-D computer by appropriately arranging a bunch of these junctions in a cube. But how did EURISKO make this invention? The crucial step was to apply the following general-purpose heuristic: "When you have a structure which depends on two different things, X and Y, try making X and Y the **same** thing." The discovery, albeit an interesting one, came **right out** of the heuristic. This is a far cry from the systematic intuition of a talented human inventor, which synthesizes dozens of different heuristics in a complex,situation-appropriate way.

For instance, the Croatian inventor Nikola Tesla, probably the greatest inventor in recent history, developed a collection of highly ideosyncratic thought processes for analyzing electricity. These led him to an steady stream of brilliant inventions, from alternating current to radio to robotic control. But not **one** of his inventions can be traced to a single "rule" or "heuristic." Each stemmed from far more subtle intuitive processes, such as the **visualization** of

magnetic field lines, and the physical metaphor of electricity as a **fluid**. And each involved the simultaneous conception of many interdependent components.

EURISKO may have good general-purpose heuristics, but what it lacks is the ability to create its own **specific-context** heuristics based on everyday life experience. And this is precisely because **it has no everyday life experience**: no experience of **human** life, and no autonomously-discovered, body- centered **digital** life either. It has no experience with fluids, so it will never decide that electricity is like a fluid. It has never played with Lincoln Logs or repaired a bicycle or prepared an elaborate meal, nor has it experienced anything analogous in its digital realm ... so it has no experience with building complex structures out of multiple interlocking parts, and it will never understand what is involved in this.

EURISKO pushes the envelope of rule-based AI; it is just about as flexible as a rule-based program can ever get. But it is not flexible enough. In order to get programs capable of context-dependent learning, I believe, it is necessary to write programs which **self-organize** -- if not exactly as the brain does, then at least **as drastically** as the brain does. However, Lenat is of a different opinion: he believes that this greater flexibility can come from a richer knowledge base, rather than a richer self-organizing cognitive dynamics. With this in mind, he moved from EURISKO to the CYC project, an ambitious attempt to encode in a computer everything that a typical 8-year-old child knows. Given this information, it is believed, simple EURISKO-type heuristics will be able to make complicated intuitive inferences about the real world.

However, the CYC project finished in 1995, and despite heavy funding over a ten year period, it cannot be labeled a success. A CD-ROM is available containing a large amount of common-sense knowledge, stored in formal-logical language. But so what? This knowledge is not sufficient for dealing with everyday situations. It is not sufficiently flexible. The CYC knowledge base is like a department store model -- it looks alive from a distance, but approach closer and you see that it's really just a mock-up. Where are the practical applications making use of the CYC CD- ROM? They do not exist.

The point is that the human mind does not embody its commensense knowledge as a list of propositions. What we know as common sense is a **self-reproducing system**, a structural conspiracy, an attractor for the cognitive equation. Abstractions aside, the intuitive sense of this position is not hard to see. Consider a simple example. Joe has three beliefsregarding his girlfriend:

A: She is beautiful

B: I love her

C: She loves me

Each of these beliefs helps to **produce** the others. He loves her, in part, because she is beautiful. He believes in her love, in part, because **he** loves **her**. He believes her beautiful, in part, because of their mutual love relationships.

Joe's three thoughts reinforce each other. According to the psynet model, this is not the exception but the rule. When Joe looks at a chair obscured by shadows, he believes that the legs are there because he believes that the seat is there, and he believe that the seat is there because he believes that the legs are there. Thus sometimes he may perceive a chair where there is no chair. And thus, other times, he can perceive a chair far more effectively than a computer with a high-precision camera eye. The computer understands a chair as a list of properties, related according to Boolean logic. But he understands a chair as a collection of **processes**, mutually activating one another.

The legs of a chair are defined partly by their **relation** with the seat of a chair. The seat of a chair is defined largely by its **relation** with the back and the legs. The back is defined partly by its relation to the legs and the seat. Each part of the chair is defined by a fuzzy set of patterns, some of which are patterns involving the **other parts** of the chair. The recognition of the chair involves the recognition of low-level patterns, then middle-level patterns among these low-level patterns, then higher-level patterns among these. And all these patterns are organized associatively, so that when one sees a certain pattern corresponding to a **folding** chair, other folding- chair-associated patterns become activated; or when one sees a certain pattern corresponding to an **armchair**, other armchair- associated patterns become activated. But, on top of these dual network dynamics, **some** patterns inspire one another, boosting one another beyond their "natural" state of activation. This circular action is the work of the cognitive equation -- and, I suggest, it is **necessary** for all aspects of intelligent perception, action and thought. The failure of AI programs to construct useful internal models of the world should be understood in this light.

But how do humans come to **build** their numerous knowledge- representing autopoietic systems? My claim is that the knowledge-representation capacities of the human mind are centered around a single dynamic data structure, called the **self**. Computer programs, as they currently exist, do not have selves -- and this is why they are not intelligent. CYC tried to get all the knowledge possessed by an eight-year old self -- without the self at the center. But the individual bits of knowledge only have meaning as part of the autopoietic self-system, and other, smaller, autopoietic mental systems. In order to write an intelligent program, we will have to write a program that is ableto evolve a variety of robust autopoietic mental process systems, including a self.

11.3 WHAT IS THE SELF?

Psychology provides many different theories of the self. One of the clearest and simplest is the "synthetic personality theory" proposed by Seymour Epstein's (1984). Epstein argues that the self is a **theory**. This is a particularly useful perspective for AI because theorization is something with which AI researchers have often been concerned.

Epstein's personality theory paints a refreshingly simple picture of the mind:

[T]he human mind is so constituted that it tends to organize experience into conceptual systems. Human brains make connections between events, and, having made connections, they

connect the connections, and so on, until they have developed an organized system of higher- and lower-order constructs that is both differentiated and integrated. ...

In addition to making connections between events, human brains have centers of pleasure and pain. The entire history of research on learning indicates that human and other higher-order animals are motivated to behave in a manner that brings pleasure and avoids pain. The human being thus has an interesting task cut out simply because of his or her biological structure: it is to construct a conceptual system in such a manner as to account for reality in a way that will produce the most favorable pleasure/pain ratio over the foreseeable future. This is obviously no simple matter, for the pursuit of pleasure and the acceptance of reality not infrequently appear to be at cross-purposes to each other.

He divides the human conceptual system into three categories: a self-theory, reality-theory, and connections between self-theory and reality-theory. And he notes that these theories may be judged by the same standards as theories in any other domain:

[Since] all individuals require theories in order to structure their experiences and to direct their lives, it follows that the adequacy of their adjustment can be determined by the adequacy of their theories. Like a theory in science, a personal theory of reality can be evaluated by the following attributes: extensivity [breadth or range], parsimony, empirical validity, internal consistency, testability and usefulness.

A person's self-theory consists of their best guesses about what kind of entity they are. In large part it consists of ideas about the relationship between oneself and other things, or oneself and other people. Some of these ideas may be wrong; butthis is not the point. The point is that the theory as a whole must have the same qualities required of scientific theories. It must be able to explain familiar situations. It must be able to generate new explanations for unfamiliar situations. Its explanations must be detailed, sufficiently detailed to provide practical guidance for action. Insofar as possible, it should be concise and self-consistent.

The acquisition of a self-theory, in the development of the human mind, is intimately tied up with the body and the social network. The infant must learn to distinguish her body from the remainder of the world. By systematically using the sense of touch -- a sense which has never been reliably simulated in an AI program -- she grows to understand the relation between herself and other things. Next, by watching other people she learns about people; inferring that she herself is a person, she learns about herself. She learns to guess what others are thinking about her, and then incorporates these opinions into her self-theory. Most crucially, a large part of a person's self- theory is also a **meta-self-theory**: a theory about how to acquire information for one's self-theory. For instance, an insecure person learns to adjust her self-theory by incorporating only negative information. A person continually thrust into novel situations learns to revise her self-theory rapidly and extensively based on the changing opinions of others -- or else, perhaps, learns not to revise her self-theory based on the fickle evaluations of society.

Self and Cognition

The interpenetration between self-theories and meta-self- theories is absolutely crucial. The fact that a self-theory contains heuristics for exploring the world, for learning and gathering information, suggests that a person's self- and reality-theories are directly related to their **cognitive style**, to their mode of thinking.

And indeed, we find evidence for this. For instance, as mentioned above in the context of consciousness and hallucinations, Ernest Hartmann (1988) has studied the differences between "thick-boundaried" and "thin-boundaried" people. The prototypical thick-boundaried person is an engineer, an accountant, a businessperson, a strict and well-organized housewife. Perceiving a rigid separation between herself and the outside world, the thick-boundaried person is pragmatic and rational in her approach to life. On the other hand, the prototypical thin-boundaried person is an artist, a musician, a writer.... The thin-boundaried person is prone to spirituality and flights of fancy, and tends to be relatively sensitive, perceiving only a tenuous separation between her interior world and the world around her. The intriguing thing is that "thin- boundaried" and "thick-boundaried" are **self-theoretic** concepts; they have to do with the way a person conceives herself and the relation between herself and the world. But, according to Erdmann's studies, these concepts tie in with the way a person **thinks** about concrete problems. Thick-boundaried people are better at sustained and orderly logical thinking; thin-boundariedpeople are better at coming up with original, intuitive, "wild" ideas. This connection is evidence for a deep relation between self-theory and creative intelligence.

What Hartmann's results indicate is that the way we **think** cannot be separated from the way our **selves** operate. This is so for at least two reasons: one reason to do with the hierarchical network, another to do with the heterarchical network. First of all, every time we encapsulate a new bit of knowledge, we do so by analogy to other, related bits of knowledge. The self is a big structure, which relates to nearly everything in the mind; and for this reason alone, it has a broad and deep effect on our knowledge representation. This is the importance of the self in the heterarchical network.

But, because of the hierarchical nature of knowledge representation, the importance of self goes beyond mere analogy. Self does not have to do with arbitrary bits of information: it has to do, in large part, with the **simplest** bits of information, bits of information pertaining to the immediate perceptual and active world. The self sprawls out broadly at the lower levels of the dual network, and thus its influence propagates upward even more widely than it does.

This self/knowledge connection is important in our daily lives, and it is even more important developmentally. For, obviously, people do not learn to get oriented all at once. They start out, as small children, by learning to orient themselves in relatively simple situations. By the time they build up to complicated social situations and abstract intellectual problems they have a good amount of experience behind them. Coming into a new situation, they are able to reason associatively: "What similar situations have I seen before?" And they are able to reason hierarchically: "What simpler situations is this one built out of?" By thus using the information gained from orienting themselves to previous situations, they are able to make reasonable guesses regarding the appropriate conceptual representations for the new situation. In other words, they build up a dynamic data structure consisting of new situations and the appropriate conceptual representations. This data structure is continually revised as new information that

comes in, and it is used as a basis for acquiring new information. This data structure contains information about specific situation and also, more abstractly, about how to get oriented to new situations.

My claim is that it is not possible to learn how to get oriented to complex situations, without first having learned how to get oriented to simpler situations. This regress only bottoms out with the very simplest situations, the ones confronted by every human being by virtue of having a body and interacting with other humans. And it is these very simple structures which are dealt with, most centrally, by the self-theory. There is a natural order of learning here, which is, due to various psychological and social factors, automatically followed by the normal human child. This natural order of learning is reflected, in the mind, by an hierarchical data structure in which more and more complex situations are comprehended in terms of simpler ones. But we who write AI programs have made little or no attempt to respect this natural order.

We provide our programs with concepts which "make no sense" to them, which they are intended to consider as given, a priori entities. On the other hand, to a human being, there are no given, a priori entities; everything bottoms out with the phenomenological and perceptual, with those very factors that play a central role in the initial formation of self- and reality-theories. To us, complex concepts and situations are made of simpler, related concepts and situations to which we already know how to orient ourselves; and this reduction continues down to the lowest level of sensations and feelings. To our AI programs, the hierarchy bottoms out prematurely, and thus there can be no functioning dynamic data structure for getting oriented, no creative adaptability, no true intelligence.

This view of self and intelligence may seem overly vague and "hand-waving," in comparison to the rigorous theories proposed by logic-oriented AI researchers. However, there is nothing inherently non-rigorous about the build-up of simpler theories and experiences into complex self- and reality-theories. It is perfectly possible to model this process mathematically; the mathematics involved is simply of a different sort from what one is used to seeing in AI. Instead of formal logic, one must make use of ideas from dynamical systems theory (Devaney, 1988) and, more generally, the emerging science of complexity. The psynet model gives one natural method for doing this.

Self- and reality- theories, in the psynet model, arise as autopoietic attractors **within the context of the dual network**. This means that they cannot become sophisticated until the dual network itself has self-organized to an acceptable degree. The dual network provides routines for building complex structures from simple structures, and for relating structures to similar structures. It provides a body of knowledge, stored in this way, for use in the understanding of practical situations that occur. Without these routines and this knowledge, complex self- and reality- theories cannot come to be. But on the other hand, the dual network itself cannot become fully fleshed out without the assistance of self- and reality-theories. Self- and reality- theories are necessary components of creative intelligence, and hence are indispensible in gaining information about the world. Thus one may envision the dual network and self- and reality-theories evolving together, symbiotically leading each other toward maturity.

A speculation? Certainly. And until we understand the workings of the human brain, or build massively parallel "brain machines," the psynet model will remain in large part an unproven hypothesis. However, the intricate mathematical constructions of the logic-oriented AI theorists are also speculations. The idea underlying the psynet model is to make mathematical speculations which are psychologically plausible. Complex systems science, as it turns out, is a useful tool in this regard. Accepting the essential role of the self means accepting the importance of self-organization and complexity for the achievement of flexible, creative intelligence.

11.4 ARTIFICIAL INTERSUBJECTIVITY

So, suppose one accepts the argument that an autopoietic self-system is necessary for intelligence, for knowledge representation. The next question is: How is the self-structure to be made to emerge from the dual network?

The first possibility is that it will emerge spontaneously, whenever there is a dual network? This is plausible enough. But it seems at least as likely that some kind of social interaction is a prerequisite for the emergence of the self-structure. This leads to the concept of **A-IS**, or "artificial intersubjectivity" (as first introduced in *CL*). The basis of A-IS is the proposition that self- and reality-theories can most easily evolve in an appropriate **social** context. Today, computer science has progressed to the point where we can begin to understand what it might mean to provide artificial intelligences with a meaningful social context.

In AI, one seeks programs that will respond "intelligently" to **our** world. In artificial life, or **Alife**, one seeks programs that will evolve interestingly within the context of their **simu lated** worlds (Langton, 1992). The combination of these two research programmes yields the almost completely unexplored discipline of **AILife**, or "artificially intelligent artificial life" -- the study of synthetically evolved life forms which display intelligence with respect to their simulated worlds. A- IS, artificial intersubjectivity, may be seen as a special case of artificially intelligent artificial life. Conceptually, however, A-IS is a fairly large step beyond the very general idea of AILife. The idea of A-IS is to simulate a **system of intelligences collectively creating their own subjective (simulated) reality**.

In principle, any AILife system one constructed could become an A-IS system, under appropriate conditions. That is, any collection of artificially intelligent agents, acting in a simulated world, could come to collude in the modification of that world, so as to produce a mutually more useful simulated reality. In this way they would evolve interrelated self- and reality-theories, and thus artificial intersubjectivity. But speaking practically, this sort of "automatic intersubjectivity" cannot be counted on. Unless the different AI agents are in some sense "wired for cooperativity," they may well never see the value of collaborative subjective-world-creation. We humans became intelligent in the **context** of collaborative world- creation, of intersubjectivity (even apes are intensely intersubjective). Unless one is dealing with AI agents that evolved their intelligence in a social context -- a theoretically possible but pragmatically tricky solution -- there is no reason to expect significant intersubjectivity to spontaneously emerge through interaction.

Fortunately, it seems that there may be an alternative. I will describe a design strategy called "explicit socialization" which involves explicitly **programming** each AI agent, from the start, with:

1) an a priori knowledge of the existence and autonomy of theother programs in its environment, and

2) an a priori inclination to model the behavior of these other programs.

In other words, in this strategy, one **enforces** A-IS from the outside, rather than, as in natural "implicit socialization," letting it evolve by itself. This approach is, to a certain extent, philosophically disappointing; but this may be the kind of sacrifice one must make in order to bridge the gap between theory and practice. Explicit socialization has not yet been implemented and may be beyond the reach of current computer resources. But the rapid rate of improvement of computer hardware makes it likely that this will not be the case for long.

To make the idea of explicit socialization a little clearer, one must introduce some formal notation. Suppose one has a simulated environment E(t), and a collection of autonomous agents $A_1(t)$, $A_2(t)$,..., $A_N(t)$, each of which takes on a different state at each discrete time t. And, for sake of simplicity, assume that each agent A_i seeks to achieve a certain particular goal, which is represented as the maximization of the real-valued function $f_i(E)$, over the space of possible environments E. This latter assumption is psychologically debatable, but here it is mainly a matter of convenience; e.g. the substitution of a shifting collection of interrelated goals would not affect the discussion much.

Each agent, at each time, modifies E by executing a certain **action** $Ac_i(t)$. It chooses the action which it suspects will cause $f_i(E(t+1))$ to be as large as possible. But each agent has only a limited power to modify E, and all the agents are acting on E in parallel; thus each agent, whenever it makes a prediction, must always take the others into account. A-IS occurs when the population of agents self-organizes itself into a condition where E(t) is reasonably beneficial for all the agents, or at least most of them. This does not necessarily mean that E reaches some "ideal" constant value, but merely that the vector $(A_1,...,A_N,E)$ enters an **attractor** in state space, which is characterized by a large value of the society wide average satisfaction $(f_1 + ... + f_N)/N$.

The strategy of explicit socialization has two parts: **input** and **modeling**. Let us first consider input. For A_i to construct a model of its society, it must recognize patterns among the Ac_j and E; but before it can recognize these patterns, it must solve the more basic task of distinguishing the Ac_j themselves. In principle, the Ac_i can be determined, at least approximately, from E; a straightforward AILife approach would provide each agent with E alone as input. Explicit socialization, on the other hand, dictates that one should supply the Ac_i as input directly, in this way saving the agents' limited resources for other tasks. More formally, the input to A_i at time t is given by the vector

$$E(t), Ac_{v(i,1,t)}(t),..., Ac_{v(i,n(t),t)}(t) \quad (1)$$

for some n < N, where the range of the index function v(i,,)defines the "neighbors" of agent A_i, those agents with whom A_i immediately interacts at time t. In the simplest case, the range of i is always {1,...,N}, and v(i,j,t) = j, but if one wishes to simulate agents moving through a spatially extended environment, then this is illogical, and a variable-range v is required.

Next, coinciding with this specialized input process, explicit socialization requires a contrived **internal modeling process** within each agent A_i. In straightforward AILife, A_i is merely an "intelligent agent," whatever that might mean. In explicit socialization, on the other hand, the internal processes of each agent are given a certain a priori structure. Each A_i, at each time, is assumed to contain n(t) + 1 different modules called "models":

a) a model $M(E|A_i)$ of the environment, and

b) a model $M(A_j|A_i)$ of each of its neighbors.

The model $M(X|A_i)$ is intended to predict the behavior of the entity X at the following time step, time t+1.

At this point the concept of explicit socialization becomes a little more involved. The simplest possibility, which I call **first order e.s.**, is that the inner workings of the models $M(X|A_i)$ are not specified at all. They are just predictive subprograms, which may be implemented by any AI algorithm whatever.

The next most elementary case, **second order e.s.**, states that each model $M(A_j|A_i)$ **itself contains a number of internal models**. For instance, suppose for simplicity that n(t) = n is the same for all i. Then second order e.s. would dictate that each model $M(A_j|A_i)$ contained n+1 internal models: a model $M(E|A_j|A_i)$, predicting A_j's internal model of E, and n models $M(A_k|A_j|A_i)$, predicting A_j's internal models of its neighbors A_k. The definition of n'th order e.s. for n > 2 follows the same pattern: it dictates that each A_i models its neighbors A_j as if they used (n-1)'th order e.s. Clearly there is a combinatorial explosion here; two or three orders is probably the most one would want to practically implement at this stage. But in theory, no matter how large n becomes, there are still no serious restrictions being placed on the nature of the intelligent agents A_i. Explicit socialization merely guarantees that the **results** of their intelligence will be organized in a manner amenable to socialization.

As a practical matter, the most natural first step toward implementing A-IS is to ignore higher-order e.s. and deal only with first-order modeling. But in the long run, this strategy is not viable: we humans routinely model one another on at least the third or fourth order, and artificial intelligences will also have to do so. The question then arises: how, in a context of evolving agents, does a "consensus order" of e.s. emerge? At what point does the multiplication of orders become superfluous? At what depth should the modeling process stop?

Let us begin with a simpler question. Suppose one is dealing with agents that have the capacity to construct modelsof any order. What order model should a given agent choose to deal with? The only really satisfactory response to this question is the obvious one: "Seek to use a depth one greater than that which the agent you're modeling uses. To see if you have gone to the

correct depth, try to go one level deeper. If this yields no extra predictive value, then you have gone too deep." For instance, if one is modeling the behavior of a cat, then there is no need to use a fifth-order model or even a third- order model: pretty clearly, a cat can model you, but it cannot conceive of **your** model of **it**, much less your model of another cat or another person. The cat is dealing with first-order models, so the most you need to deal with is the second order (i.e. a model of the cat's "first-order" models of you).

In fact, though there is no way to be certain of this, it would seem that the second order of modeling is probably out of reach not only for cats but for all animals besides humans and apes. And this statement may be made even more surely with respect to the next order up: who could seriously maintain that a cat or a pig can base its behavior on an understanding of someone else's model of someone else's model of itself or someone else? If Uta Frith's (1989) psychology of autism is to be believed, then even autistic humans are not capable of sophisticated second-order social modeling, let alone third-order modeling. They can model what other people do, but have trouble thinking about other peoples' images of **them**, or about the network of social relationship that is defined by each person's images of other people.

This train of thought suggests that, while one can simulate some kinds of social behavior without going beyond first order e.s., in order to get true social complexity a higher order of e.s. will be necessary. As a first estimate one might place the maximum order of human social interaction at or a little below the "magic number seven plus or minus two" which describes human short term memory capacity. We can form a concrete mental image of "Joe's opinion of Jane's opinion of Jack's opinion of Jill's opinion on the water bond issue," a fourth-order construct, so we can carry out fifth-order reasoning about Joe ... but just barely!

According to this reasoning, if intelligence requires self, and self requires intersubjectivity, then there may be no alternative but to embrace A-IS. Just because strong AI is possible does not mean that the straightforward approach of current AI research will ever be effective. Even with arbitrarily much processing power, one still needs to respect the delicate and spontaneous self-organization of psychological structures such as the self.

CHAPTER TWELVE

SUBSELF DYNAMICS

12.1 INTRODUCTION

So far, we have discussed psychology a great deal, but mostly in the domains of cognition and perception. We have not had much to say about everyday psychology, the psychology of everyday human behavior. Now it is time to turn to these more nebulous but ultimately more intriguing aspects of the human mind. It is time to ask: Why do we humans act the way we do? What are the dynamics and structures underlying human personality?

These are complicated questions, and psychology has given us many complicated answers. In this chapter I will provide a new and different answer -- one which, while relying on concepts from complex systems science, is notably simpler than many of the theories of classical clinical psychology. I will use the psynet model to formulate a novel personality theory called "subself dynamics."

Subself dynamics is a radical and "experimental" theory, which is still in an early stage of development. However, it seems to have a great deal of potential for explaining complex personality phenomena -- phenomena that other approaches to personality dismiss as puzzling or paradoxical. It will be a crucial component of the synthetic theory of **creativity** to be proposed in the last chapter.

Psychologists may be alarmed at the degree to which I appear to ignore previous theories of personality. But this is not done out of ignorance or arrogance. I am well aware of the ideas of the "classic" personality theorists -- Freud, Allport, Murray, Rogers, Kelley, Lewin, Bandura, and so forth -- and I have a great deal of respect for these thinkers. In some instances the system-theoretic approach may lead to the **rediscovery** of classical ideas. Thus, by failing to mention these personality theorists I do not mean to imply that their theories are wrong, but only that, except in a few cases, which will be duly noted, I have not found their work particularly useful in developing a complex systems based approach to personality. This does not mean that relationships do not exist between the present theory and past theories -- only that these relationships are not particularly **direct**, and thus are not essential to the exposition of the present ideas.

The idea of the present approach is, instead of formulating theories based on pure intuition as these thinkers have done, to use complex systems science as a guide for the intuitive analysis of personality phenomena. Freudian ideas will pop up here and there; and the final section of this chapter briefly discusses the power of systems theory to unify traditional personalitytheories with cognitive science. But, by and large, I have been content to leave to others the task of elaborating the relationship between the complex systems ideas presented here and the ideas of past personality theorists.

In the previous chapter, we discussed the notion of **self**, in the context of artificial intelligence. A self-system, we concluded, is absolutely necessary for adequate knowledge representation and learning. However, no attention was paid there to the actual **structure** of the self. Here we will turn to this issue in earnest. We will approach human personality and human nature from the perspective of the **dissociated self**. Every person's self system, we will argue, is naturally divided into a number of situation and state dependent "subselves." The interaction dynamics between these subselves is the crucial factor in a person's complex, internal evolving ecology. Thus, human experience is viewed as **subself dynamics**.

12.2 SUBSELVES

Proust, when at age 16 he wrote of "the several gentlemen of whom I consist," was reporting a universal human experience. Each of us acts differently in different situations -- at home, at work, out drinking with friends. Proust's insight was that these different perception and behavior

patterns, elicited by different situations, are not merely aspects of the same person, they are fundamentally **different people** -- different selves inhabiting the same body.

This may seem a strange idea, but in the end it is almost obvious. For what is a "self," after all? A self is just an autopoietic process system, consisting of a dual network of procedures for mediating between mind and world. There is no reason whatsoever why a mind should not contain many such networks. Different situations require fundamentally different mental processes, perhaps even mutually contradictory mental processes -- it is foolish to expect that this diversity of aims should be accomplished by a unified system.

Excessive separation of subselves is a problem. One does not want subselves with different names, entirely different identities, mutually exclusive memories. But one does want subselves with slightly different identities, and slightly different memories. One's memory of a situation when at work may be entirely different from one's memory of that situation when at home -- and for good reason, because memory is in large part a constructive process. One builds memories from the raw materials provided by the mind, in order to serve specific purposes. Different subselves will have different purposes and thus different memories.

The concept of subselves has become popular in psychotherapy circles in recent years (see e.g. Rowan, 1991). There are techniques for "letting one's subpersonalities speak," for coming into contact with one's lover's subpersonalities. A woman might have a "scared little girl" subpersonality, a man might have a "neighborhood bully" subpersonality. A straight-laced societywoman might have a repressed "slut" subpersonality, denied expression since her teenage years. In this type of therapy one deals with subpersonalities on an entirely individual basis: each person must discover, with the therapist's help, what their subpersonalities actually are.

The present theoretical approach is complementary to this type of psychotherapy. My aim in this and the following chapters will be to seek **universal** aspects of subself dynamics. These universal aspects are closely tied in with the more situation- specific aspects dealt with by psychotherapists. In order to thoroughly understand any particular individual, both approaches are necessary.

In Epsteinian terms, it should be understood that a "subself" contains a self system, a world system, and a system of interrelations between them. Insofar as they bring to mind multiple personality disorder, the words "subself" and "subpersonality" may seem too strong. But if this association is put aside, one finds that, if they exaggerate the situation at all, it is only by a very small margin. For in truth, the word "personality" as it is generally understood would seem to be a perversion of the facts. "Personality" is too individualistic; it implies that the qualities of a person are fundamentally a property of that person alone, when in fact these qualities are in every respect **social**. They are formed through social interaction, and they are also **defined** by society, in the sense that, to a person from a substantially different society, they would be largely incomprehensible. So a given person's "personality" does not really exist except as part of a **network of personalities** -- just as a self/reality subsystem does not really exist except as part of a network of self/reality subsystems. And in certain cases two subselves residing in different bodies may be more closely related than two subselves residing in the **same** body; for instance,

the roles that two lovers assume when with one another often have little to do with their other subpersonalities. This fact is exploited frequently in fiction.

A self/reality subsystem is an **autopoietic system**; this constitutes its fundamental wholeness. And a person's whole self-reality system is **also** an autopoietic system, whose interproducing component parts are precisely its **subsystems**. In other words, a personality is an autopoietic system whose component parts are subpersonalities, or **subselves**.

Some psychologists might reject system theory as a foundation for personality, arguing that the proper grounding for a psychological theory is behavioral observation, or neurophysiological data. But while it is certainly important for any theory to agree with the available empirical data, the biological and behavioral data regarding personality are clearly too scanty to allow for inductive theory construction. All serious theorists of the **individual** personality (and I do not include trait theorists in this category) have formed their theories by making generalizations from their personal experience. But such theories are always sketchy and full ofgaps, because they lack the rigorous logical structure provided by system theory.

A priori, it would not be unreasonable to extend the subself idea even further, to obtain subsubpersonalities, and so forth. I propose, however, that in the case of human beings this is not necessary. In other words, I suggest that, just as there is a magic number 7+/-2 for human short term memory capacity, and a magic number 3+/-1 for levels of learning (see Bateson, 1980, or *EM*), there is a **magic number 2+/-1** for personality system depth. The "-1" allows for the possibility that some severely retarded people may only be able to develop a single self/reality system without significant subdivisions; and the "+1" allows for the possiblity that in some exceptional cases, subpersonalities might develop subpersonalities. An example of the latter might be certain MPD patients, in which the strongest subpersonalities might eventually develop the flexibility and state-dependence of ordinary "top-level" self/reality systems. The relation of this magic number to the others is a subject for speculation.

Emergent Attractors

Finally, having discussed subselves at length, I must now introduce a related idea that is very useful for the task of applying the psynet model to the study of personality. This is the concept of "emergent attractors" spanning different subselves and different human beings. Emergent attractors are one of the main reasons why the analysis of traits will never lead to an understanding of personality. The essence of personality lies, not in the particular traits of particular human beings or subselves, but in the **behavioral routines** that emerge when two or more subselves, or two or more people, interact.

This fact will become abundantly clear below, when we discuss romantic love. Jack does not love Jill for her "qualities," at least not for any "qualities" that could be listed in words in a brief, comprehensible way. He loves her because of the **way she makes him feel**, or in other words, because of the pleasurable behavioral routines which emerge between the two of them when they interact. To determine why Jack and Jill lock into mutually pleasurable routines when Jack and Jane do not, is a very difficult problem. Jill and Jane may differ very little by outward "qualities," yet their behavioral routines on contact with Jack may differ greatly. Predicting the behavioral

routines emerging between different people is no easier than predicting the behavior of a complex mathematical dynamical system -- something which, as chaos theory has taught us, is a very difficult task.

And what holds for Jack and Jill, I will argue, also holds for the different sides of Jack's own personality, for Jack's different subselves. To explain how Jack's different subselves get along, or why they don't get along, it is not sufficient to list their traits, because these traits do not contain enough information to predict the behavior of the **dynamical system** formed by linking the various subselves together. The emergent attractors formed by two subselves may give rise to complex patterns that, in turn, spawn or modify another subself.

12.3 I-IT AND I-YOU

The philosopher Martin Buber, in his famous book *Ich und Du*, distinguishes two fundamentally different ways of relating: the "I-It" and the "I-You" (sometimes called "I-Thou"). These concepts have not been picked up in any significant way by the psychology community. However, I believe that this has been a serious oversight. An understanding of these two modes of relationship is absolutely essential for the study of human personality. In this section I will present a computational psychology of human relationship corresponding to Buber's philosophical theory, and will apply this theory on the level of subselves within a single person, as well as on the interpersonal level.

Then, in the following section, I will argue that a healthy mind, as a rule, consists of a population of subselves carrying out mutual I-You relationships with each other. This kind of subself community leads to robust, adaptive belief systems. On the other hand, a subself community containing a preponderance of I-It relationships will be characterized by self-sustaining belief systems of minimal adaptive value. This conclusion is what I call the **Fundamental Principle of Personality Dynamics**. The Fundamental Principle connects subself dynamics with underlying thought dynamics, and thus bridges the gap between social and personality psychology, on the one hand, and cognitive psychology on the other. It could, in principle, have been arrived at without any of the apparatus of computational psychology and complexity science. In fact, however, it was not. These modes of thinking lead up to the Fundamental Principle in a remarkably natural way.

Buber on I-It and I-You

An I-It relationship is a relationship between a self and a thing or a collection of things. An I-You relationship, on the other hand, is a relationship between a self and another self **as a self**. Two selves may relate in the I-It mode, or in the I-You mode: it is a question of whether the one self recognizes the other one as a genuine self, equal to itself in reality and integrity; or whether it merely interprets the other one as a stiff, inanimate portion of the external world.

Buber describes the I-You relationship in rather mystical term:

-- What, then, does one experience of the You?

-- Nothing at all. For one does not experience it.

-- What, then, does one know of the You?

-- Only everything. For one no longer knows particulars.

...

The basic word I-You can be spoken only with one's whole being. The concentration and fusion into a wholebeing can never be accomplished by me, can never be accomplished without me. I require a You to become; becoming I, I say You.

...

The relation to the You is unmediated. Nothing conceptual intervenes between I and You, no prior knowledge and no imagination; and memory itself is changed as it plunges from particularity into wholeness.

I-You plunges the I into a timeless world of interdependence. I-It, on the other hand, is firmly grounded in the world of time:

The I of the basic word I-It, the I that is not bodily confronted by a You but surrounded by a multitude of "contents," has only a past and no present.... He has nothing but objects; but objects consist in having been.

According to Buber, there is no real possibility of perceiving everyone as a You all the time. Rather, real human relations are a process of oscillation between the two extremes:

This, however, is the sublime melancholy of our lot that every You must become an It in our world. However exclusively present it may have been in the direct relationship -- as soon as the relationship has run its course or is permeated by *means*, the You becomes an object among objects.... The human being who but now was unique and devoid of qualities, not at hand but only present, not experiencable, only touchable, has again become a He or a She, an aggregate of qualities, a quantum with a shape. Now I can again abstract from him the color of his hair, of his speech, of his graciousness, but as long as I can do that he is my You no longer and not yet again.

In our present culture, however, we are closer to the I-It extreme than was the case in the past. Indeed, Buber views I-You as being a more primitive notion than I or You itself:

[T]he primitive man speaks the basic word I-You in a natural, as it were still unformed manner, not yet having recognized himself as an I; but the basic word I-It is made possible only by this recognition, by the detachment of the I.

The former word splits into I and You, but it did not originate as their aggregate; it antedates any I. The latter originated as an aggregate of I and it, it postdates the I.

I-You relationships are not necessarily happier than I-It relationships. However, they are deeper. It is better,according to Buber, to genuinely hate another person, than to feel toward him only shallow surface sentiments. The former is a path leading to God; the latter leads only to nothingness. For, in perhaps his most radical contention, Buber equates the experience of an I-You relationship with the experience of God.

Modelling Other Selves

One may ask what sorts of differences in cognitive processing underly the differences between these two ways of relating. What are the mechanisms underlying these two kinds of experiences? This is a rather prosaic question, one which perhaps jars oddly with Buber's flights of mystical enthusiasm, but it is an interesting one nonetheless.

I claim that the answer to this question is quite simple. The mechanism underlying an I-It relationship is a recognition of the particular patterns displayed by another individual, another self. On the other hand, the mechanism underlying an I- You relationship is an **implicit** recognition of the **overall emergent structure** of the other. This overall emergent structure will, in general, be too complex and subtle to enter into consciousness as a whole. Thus it will be experienced as broad, featureless, abstract. But the mind can nevertheless experience it.

The experience of the You does not depend on the details of the individual being experienced. But that does not contradict the equivalence posed here, between the You and the overall emergent structure of another self. For it is the key insight of complexity science that the high-level emergent structures observed in a system **need not depend on the lower-level details of the system**.

It follows from this that perception as an It is a prerequisite for perception as a You. One must first recognize the lower-level patterns out of which the higher-level patterns emerge. And, furthermore, one obtains the concrete prediction that some individuals will be more easily recognized as You than others. This is indisputably the case. Although the experience of the You is independent of the "It"-qualities of the individual being experienced, the probability of entering into an I-You relationship with a given individual is sharply **dependent** on these qualities. It is necessary that one recognize a sufficiently wide and deep assortment of the individual's qualities, in order to be able to make the leap to the overall ordering structure.

What, though, is the overall ordering pattern of a self? It is, I have said, a recursive dual network. A dual network which contains an image of itself. The patterns making up the dual network are the things that are recognized in an I-It encounter with someone's mind. An I-You encounter, on the other hand, is marked by the rhythm of the network as a whole passing through itself again and again. It passes through itself -- mirrors itself -- and thus becomes itself, getting back to whereit started (perhaps with minor modifications; but this is not something that can be detected on a high level of perceptual abstraction). To enter into an I-You relationship with someone else is to enter into **their time-stream**; their inner process of unfolding and self-

creation. For it is this inner process of unfolding that makes them a You, a self, a whole person, rather than just a collection of mental and physical habits.

"You" Need Not Be Human

From this analysis one might conclude that it is only possible to enter into an I-You relationship with another intelligence -- and not with an inanimate object, say, a flower, or a personal computer. But this would be mistaken. The external world also possesses an emergent, self-unfolding structure that can be grasped as a whole. Buber recognizes this in his passage on Goethe:

> How beautiful and legitimate the full I of Goethe sounds! It is the I of pure intercourse with nature. Nature yields to it and speaks ceaselessly with it; she reveals her mysteries to it and yet does not betray her mystery. It believes in her and says to the role: "So it is You" -- and at once shares the same actuality with the rose.

Nature reveals its mysteries to Goethe -- i.e., he paid it careful enough attention to get to know it as an It; as a dynamic, structured, changing system full of intricate details. And this effort was repaid by a glimpse at the whole, by a view of nature as You.

This sort of experience, reported by Goethe, was commonplace in many past cultures. Communion with nature as a You was, in "primitive" cultures, understood as communion with God. This experience came almost automatically to individuals who lived in nature every day, and who were raised to relate to nature as a thinking being rather than as an inert physical "environment." This is the sad aspect of current political disputes about environmental issues. Businessmen are asked to respect nature on abstract moral grounds. But such respect can never be genuine, because it comes out of an I-It relationship with nature, rather than an I-You relationship. Having lost the I-You relationship with nature, we have, as a society, lost our fundamental motivation for preserving and nurturing the natural world.

The patterns of nature as a whole are relatively similar to the patterns of the human mind. In principle, however, it is even possible to carry out I-You relationships with a truly inanimate object. Say, with a rock. Putting aside ultimate philosophical questions of the reality of the external world, one may at least say that a rock **as perceived by a given I**, is a collection of perceptual patterns. This collection of perceptual patterns is, under ordinary circumstances, an It. It is not alive; it does not enter into itself and create its own time-stream. However, there are states of mind -- mystical, "oceanic" states -- in which even such inert collections of patterns are perceived to "come to life." The fundamental self-creating dynamic of self and awareness is extended throughout the more rigid layers of the mind. In a state of mind such as this, inanimate objects may be perceived as You. Everything becomes a You. This state of mind is touched on by Buber in the third part of his book. It is the state of mind of the truly enchanted being: the seer, the saint, the Zen master. Buber rejects mysticism that is based on finding the self within. True insight, he said, is based on deep and lasting recognition of the world as You.

12.4 THE FUNDAMENTAL PRINCIPLE OF PERSONALITY DYNAMICS

Suppose one has a collection of selves, all relating to each other in the I-You manner. Each self thus recognizes the others as selves. What benefit is this to the collection as a whole? What good does it do a self to be recognized by the others as a self?

This question has relevance both for personality and for social psychology. For it applies equally to the case where all the selves in question are part of a single mind, and the case where the selves are located in different physical bodies.

From a common-sensical, human-relations point of view, the answer to our question is almost obvious. From a mathematical point of view, on the other hand, it is an extremely difficult question, one whose complete resolution lies far beyond the grasp of current analytical techniques. This situation is, unfortunately, typical of all the really interesting psychological questions. Nothing is guaranteed in psychological systems: there are exceptions to every would-be rule. One is always talking about probable outcomes in certain situations, where there is, however, no apparent way to quantify the actual probabilities involved, nor to place strict boundaries on the set of situations involved.

The first idea needed to resolve the question is that the I-You relationship is beyond judgement. To enter into an I-You relationship with a self is to accept that self as a law unto itself, to understand the autopoietic network of relationships by which the different patterns making up that self relate to one another. If one understands and directly perceives this network, then one is bound to be tolerant of the particular patterns making up this other self, even if, taken individually, they might not seem agreeable. On the other hand, in the I-It relationship, the holistic integrity of the other is not immediately perceived, and thus the intuitive reaction must be to reject those aspects of the other's being that are not agreeable in themselves. Intellectually, the I in an I-It relationship might reason that the other self has some reason for doing what it is doing. But this will not be **felt**, which is the important thing.

So, in a network of I-You relationships, there will be atendency of selves not to interfere with each other's actions on a microscopic basis. There will tend to be a certain respect -- based on a mutual understanding of the contextuality of actions, on a direct perception of the roles of individual patterns in autopoietic self-and reality-systems. What effect does this respect, this leeway, have on mental functioning?

Clearly, an atmosphere of tolerance and respect will lead to a decreased need for **defensiveness** on the part of individual thought-systems. If each individual thought-system proposed by a certain self is immediately going to be shot down by other selves, then the self in question will have to be very careful to protect its thought-systems, to make them self-sufficient and resistant to attack. On the other hand, if there is an atmosphere of relative leniency and tolerance, then resilience is no longer so important, and other aspects of thought-systems may be emphasized.

But recall that, according to the psynet model, there are two ways for thought systems to survive in the mental network: by autopoiesis or by adaptation. Each thought system must have a

little of both features. As a very general rule, it is usually the most dogmatic, entrenched and unproductive thought systems that survive primarily by autopoiesis. And, on the other hand, it is the shallowest and least profoundly insightful belief systems that survive primarily by adaptation, by constant self- adjustment to the fluctuations of the environment. The best belief systems are the ones which use a bit of autopoiesis, to keep themselves going when the environment fluctuates against them, and a bit of adaptation, to ensure their relevance to external situations, their contextual "fitness" and productivity.

This line of reasoning leads up to the conclusion that a system of selves characterized by I-It relationships will tend to produce overly dogmatic, entrenched thought systems. A system of selves characterized by I-You relationships will, on the other hand, tend to produce more well-balanced thought systems, which are adaptively effective as well as adequately self-preserving. In scientific terms, a statistical correlation is posited: between the adaptivity and productivity of thought systems, and the I-You nature of inter-subself relationships. This correlation is what I call the "Fundamental Principle of Personality Dynamics"

Obviously, I-It interself dynamics is not the **only** explanation for dogmatic, inflexible thought. Inflexible thought systems can arise for any number of reasons. For example, in *Chaotic Logic* I give a detailed example of an inflexible thought system, the conspiracy theory of a paranoid individual. Everything said to the paranoid person becomes "part of the conspiracy." The basic principles of the system never adapt, and the new explanations which these principles generate bear little relation to the situations they allegedly respond to. The system generates little emergent pattern in conjunction with its environment -- it is "un-fit." However, this system did not arise primarily as a consequence of subself dynamics, but rather as a consequence of **unvaryingly I-It interactions with the outside world**. The paranoid individual, more than anyone else perhaps, does not accept anyone else as a You. Their suspicionholds them back from this intimacy. Their world is an object to be manipulated and analyzed. They have no respect for the emergent wholeness of the world, or the others in it, and it is precisely for this reason that they react so oddly to the specific patterns that confront them in their lives.

There is a possible connection with developmental psychology, which should be noted. Apparently, we learn to relate by observing our parents, siblings and others relate. Thus a person who was presented with predominately I-It relationships during their youth, would be expected, on average, to develop primarily I-It relationships internally, among their subselves. This leads to the hypothesis that individuals whose early childhoods are deficient in I-You relating will tend to display inflexible belief systems throughout their lives. Their thought systems will be inflexible in regard to the external world; and particular components of their personality will be inflexible in their relations to each other. This prediction is not quite a necessary consequence of the Fundamental Principle: one can imagine mechanisms by which early I-It relationships might lead to internal I-You relationships. However, it is a "likely" consequence of the Fundamental Principle and as such would seem worthy of empirical exploration.

Finally, although I have been mainly thinking of communities of selves within a single mind/body, it should be reiterated that the same conclusions apply to groups of human beings. For instance, in a marriage, which is a group of two, the Fundamental Principle would suggest that the degree of inflexibility of the partners' belief systems should be inversely correlated with

the frequency of I-You encounters between the partners. One of the remarkable things about the subself perspective is the way it erodes the border between social psychology and personality psychology. Both are seen to follow the same high-level dynamic patterns.

12.5 SYSTEMS THEORY AS A "BRIDGE"

Before moving on to apply and extend these ideas, it is worth briefly pausing to ask: how does the system-theoretic view of personality outlined in this chapter and the last relate with previous personality theories? This is a complex question, far too large to be substantially explored in this brief treatment; but a few simple observations may be made.

The key point I would like to make here is that systems theory bridges the gap between modern cognitive science, which provides incisive understanding of particular mental processes, and intuitive personality theorizing as represented e.g. by Freudian psychology, which provides deep but vague understanding of high-level personality phenomena. The Fundamental Principle described in this chapter is an example of this kind of "gap- bridging" at which systems theory is so adept.

Freudian Theory and Systems Theory

Let us begin from the top, with the "high-level" theories. Many of the great psychologists have had theories of personality: Freud, Jung, Allport, Murray, Kelly, and so forth. These theories are extremely difficult to assess. One may observe that psychiatrists, in controlled tests, have proved no better than anyone else at the task of predicting behavior. But on the other hand, one may quite correctly argue that, unlike, say, trait theories of personality, these non-quantitative theories of personality were never **intended** for prediction. They were intended to provide **understanding** of specific cases. And if chaos theory has taught us anything, it is that **prediction** and **understanding** are very different things.

My quarrel with these personality theories is not their poor prediction ability, but something much more meaningful, their **lack of subtlety**. These theories simply do not do justice to the complexity of personality. The system-theoretic treatment given here does not fully do justice to this complexity either, but it is a step in the right direction. What past personality theories have done is to replace **complexity** with simple **stock ideas** -- ideas pulled "out of a hat," with no scientific or mathematical foundation, for the sole purpose of making personality theory simpler. This point is best made clear by a concrete example. Let us briefly discuss Freudian theory -- which, for all its obvious faults, has come closer to an understanding of the **complexity** of personality than any of its competitors.

The weakness of Freudian psychology is well symbolized by the concept of **psychic energy**. This is a **metaphorical** idea, pulled out of a hat and inserted at precisely the point in Freudian theory where a **scientific** idea would have been most valuable. And what were the conclusions to which Freud was led by this idea? Chief among them was the idea that it is pointless to treat the **symptoms** of a psychologically troubled person, because training a person not to express a certain symptom will simply cause the psychic energy underlying that symptom to be redirected elsewhere. However, over the past 40 years many psychoanalysts and other psychotherapists

have come to the conclusion that this Freudian hypothesis is generally not true. Very often, symptoms are treated, and no other symptoms arise to take their place. Now, how is this observation to be integrated into Freudian theory? If only some symptoms are worth treating, is there some way to determine which ones? Freudian theory is not equipped to answer these questions, because it is not grounded in any scientific or mathematical discipline; it is a "castle built in the air."

Freud was trying to formulate laws to explain the behavior of a complex system, the personality, but he had no concept of the brain/mind as a complex system to fall back on, and one can see this shortcoming in the details of his personality theories. The whole idea of a symptom as a consequence of an underlying problem bespeaks a failure to appreciate the **circularity** of complex system dynamics. The symptoms maintain and produce the underlying problem, and the underlying problem maintains and produces the symptoms: the whole "complex" is an autopoietic,self-organizing system. In some situations, removing one element from an autopoietic system will destroy the whole system; in other situations it will not. This is not a mysterious matter, it is a question of **basin size and shape**, of stability with respect to perturbations. One can imagine that the diagnostic manuals of the future will contain, along with their lists of different disorders, explicit information about the size and shape of the basin of the autopoietic subsystem characterizing each disorder. In all but a few instances, Freud overlooked the very **possibility** of this kind of complexity.

It is important to make a distinction here. Freud, like the other great personality psychologists who followed him, had a unique and profound understanding of human personality. In their analyses of specific cases, in the organization of their descriptions, in their detailed decisions as to what to emphasize and what to omit, they revealed the depth of their personal understanding, a depth comparable to that achieved by Dostoevsky, Nietszshe, Proust, and other great artists and philosophers concerned with human nature. But the great psychologists were much more accomplished at **understanding** personality than at **theorizing** about it. Their intuitive power far exceeded their power to codify their intuitions. And this does not necessarily reflect any innate inabilities on their part; they lacked appropriate **tools** to aid them in the process of codification. Freud, for one, was acutely aware of this: he once formulated a neurological theory of personality, only to set it aside instead of publishing it, grimly but realistically aware that neuroscience was not yet sufficiently advanced to allow him to do what he wanted to do. In place of this neurological theory, he promoted a theory based on "psychic energy," an idea with absolutely no scientific foundation.

These strengths and weaknesses of psychoanalytic theory will become quite obvious below, as we consider the concrete phenomena of romantic love. On the one hand, confusing, jargony concepts such as the Oedipus complex are invoked excessively often, in contexts where much less esoteric explanations clearly suffice. But on the other hand, one also finds deep and penetrating analyses of the varieties of human experience. Even if one rejects the Oedipal aspects of Freud's analysis of masochism, one still often finds that he has captured the basic **structure** of the phenomenon.

So, in conclusion, it is interesting to contrast the system- theoretic approach with the Freudian approach. Freud was, above all, concerned with **modeling the mind**. Modern science has given

us new tools with which to carry out this modeling, but the basic goal is still the same: to understand the hidden structures and processes which cause us to feel and act the ways we do. The biggest difference is that Freud built his models from concepts like id, ego and super-ego, which were fabricated especially for the purpose of modeling mind. Thus one finds that many of Freud's models are clear and sensible in their abstract structure but unclear or unreasonable in their details. In the system- theoretic approach, on the other hand, the details are not left "dangling" into a sea of jargon and ad hoc concepts, but arerather grounded in concrete computational and complex systems models. Even if, as a consequence of lack of data, a system-theoretic model of some personality phenomenon must be constructed on intuitive rather than deductive grounds, at least the concepts used as **building blocks** for the model will be scientifically meaningful.

Systems Theory and Cognitive Science

Now let us turn to the "low-level" side of things. One might hope that a new, scientifically-based theory of personality would emerge from **cognitive science** -- an interdisciplinary endeavor whose influence is pervading more and more aspects of psychological theory, so much that Mandler (1985) was moved to pronounce that "modern psychology **is** cognitive psychology." But this has not yet occurred, and, for reasons that should be clear from the previous chapters, I suggest that it is unlikely to occur in the near future.

The basic **metaphors** of cognitive science, I suggest, are just not up to the task. Their focus is excessively **micro**-level; very little attention is paid to crucial questions of high-level structure and dynamics. If mainstream cognitive science were to turn to something like the psynet model, or Kampis's component systems or Varela's autopoietic systems, then it might have a reasonable chance of dealing with high-level psychological issues. But as it stands now, the leaders in the field of cognitive science have no such inclinations, and so the prospects are not good.

Cognitive science is not so much a unified science as an amalgam of three different subdisciplines: artificial intelligence, cognitive neuroscience and cognitive psychology. And a great deal of work in cognitive science is still pervaded by the **logicist** paradigm of rule-based artificial intelligence. To oversimplify just a little, the metaphor is, "mental processes as simple algorithms, which have short programs in our high-level programming languages, and which are largely understandable in isolation from each other."

Carried over from artificial intelligence to cognitive psychology, the logicist approach works well with certain issues -- e.g. factual memory, short-term memory, and perceptual processing -- but is very poorly suited for personality. For personality is deeply dependent on self-organization and interdependence, the very things which traditional cognitive science fails to capture. Cognitive psychologists have made positive contributions to personality theory: for instance, they has shown that we tend to use ourselves as a prototype for understanding others, and that memory depends on personality, e.g. in that certain types will tend to remember successful experiences, while others will tend to remember unsuccessful ones. But while this sort of study is enough to help one adjust the "parameters" of a personality theory, it is not, in itself, a strong enough foundation on which to build a theory of human nature.

The cognitive science community has begun to growdisenchanted with the logicist metaphor, and has embarked upon a fair amount of research embodying a "connectionist," rather than logicist, point of view. Connectionism, unlike logicism, is strong on self-organization and interdependence. The basic connectionist metaphor, the "neural network," is, like the mind, a self-organizing dynamical system. However, the central weakness of the connectionist approach is precisely its inability to address issues of large-scale cognitive structure. In technical language, neural network models do not "scale up," and the reason these models do not scale up is that they do not incorporate any of the intermediate-scale or large-scale structures of the brain/mind. A neural network psychology which focused on these structures would bear little resemblance to the neural network psychology of today. Thus, although system- theoretic psychology is related to connectionist cognitive science, the relationship is not so close as it might seem. The ideas of autopoietic system theory are unlikely to be "happened upon" through experimentation with toy neural networks; they are external to the conceptual repertoire of neural network psychology.

So we find that mainstream cognitive science has precisely the **opposite** flaw of Freudian theory. Freud's penetrating and sophisticated explanations of high-level structure and dynamics had no solid grounding in lower-level dynamics. On the other hand, cognitive science, with its broad-based and incisive understanding of details, has no way of reaching up toward higher-level structures. The connection between the two, I propose, is **system theory**. System theory connects the micro level with the macro level, the behavior of the relatively simple components with the behavior of the structured whole. It is thus ideally suited to serve as a **bridge** between modern cognitive science and old-style intuitive personality theorizing.

The Fundamental Principle as a Bridge

System theory, in general, has the ability to bridge intuitive personality theory with modern cognitive psychological theory. The Fundamental Principle of Personality Dynamics proposed here is, I would claim, a manifestation of this ability. The concept of I-It versus I-You relationship is an high- level philosophical idea which is at home among intuitive theories of personality. Likewise, the idea of subpersonalities comes from clinical rather than experimental psychology. These ideas tie in closely with everyday human experience.

On the other hand, the study of thought systems and their rationality, adaptiveness and coherence lies squarely within cognitive psychology. Because of the practical difficulties of gathering adequate time series data, questions of the dynamical adaptivity and resilience of thought systems have not been studied extensively to date. However, these questions are natural extensions of the cognitive science research programme. From an abstract mathematical point of view, they have been studied extensively within the artificial intelligence community.

Thus, the Fundamental Principle connects the personality level and the cognitive level in a way that has never been donein either personality psychology or cognitive psychology alone. By treating the mind as a whole system, with complex dynamical interactions on several different levels, one comes to view the different aspects of mind studied by different subdisciplines of the discipline of psychology as parts of a unified whole. Of course, on a general level, this unified view of mind is absolutely commonsensical -- the separation into different aspects and levels is

purely a consequence of the sociology of the discipline of psychology. But the key thing is to show that the unified view can lead to definite insights. This is what system theory accomplishes.

CHAPTER 14

ASPECTS OF HUMAN PERSONALITY DYNAMICS

14.1 INTRODUCTION

The next and final chapter will confront the topic of creativity, using the theory of subself dynamics, together with genetic algorithms, autopoietic systems, emergent patterns and other concepts from the psynet model, to construct a unified theory of creative dynamics. Before turning to creativity, however, it seems worthwhile to give an example of the application of subself dynamics and the psynet model to other personality phenomena, less blatantly cognitive in nature. With this in mind, in this chapter I will provide new, system-theoretic analysis of the phenomena of **romantic love** and **masochism**.

Romantic love, our first topic, is a phenomenon in which theoretical psychologists have historically had very little interest. Relatively obscure phenomena such as sadomasochism, and mental disorders such as schizophrenia, autism and hysteria, have garnered vastly more attention. In recent years, however, this has begun to change. In this section I will review some recent ideas about the psychology of romantic love, and I will explain how the psynet model, combined with subself dynamics and certain ideas from the psychological literature, suggests a new way of thinking about love.

In short, romantic love is here envisioned as an autopoietic, adaptive supersystem constructed from four inter-reinforcing, inter-merging subsystems: a sexuality system, a caregiving system and an attachment system, and a higher-level system called the "intimacy system." This supersystem sustains itself internally by autopoiesis, and externally by the production of intimacy, defined as "the systematic co-creation of patterns of mutual emotional significance." Falling in love is then viewed as the process of converging to this supersystem -- this autopoietic, adaptive attractor. Subself dynamics adds an extra level of complexity to the dynamics, for different subsystems are present to different extents in different subselves, and so the autopoietic supersystem must learn to adapt itself to the shifts between subselves.

This may seem to be an overly complicated analysis of something as primal and experientially simple as romantic love. But such complication is, I would argue, the price one must pay for a thorough and understanding. The psynet model reveals the subtle dynamics underlying the elemental feeling of love. In doing so it explains certain aspects of love relationships which have previously been incomprehensible. And it also explains the **unpredictability** of love -- why it is so difficult to predict who will fall in love with whom. According to the psynet model, this difficulty is not so different from the difficulty of predicting the currents of the atmosphere, or

the rise and fall of economic indicators. It is an unavoidable consequence of the complex nonlinearity of subself dynamics.

14.2 THE LAWS OF LOVE

I will begin by outlining a psychoanalytic view of romantic love. While Freud himself had little to say on the subject, a recent paper by Otto Kernberg (1988) provides an analysis of romantic love from the Freudian perspective:

> I believe romantic love, with its constituents of idealization, longing, and the sense of passionate fulfillment when the love relation with the desired object is achieved, reflects the unconscious achievement of the union with the desired incestual object and yet the capacity to overcome the infantile equation of all sexual objects with the oedipal one, and a triumphant identification with the oedipal rival without the implication of patricide or matricide. In normal passionate love, the distinction between the original oedipal rival and other competitors of the same sex has been achieved, and the related sense of inferiority to both parental objects linked with the infantile origin of the Oedipus complex has been overcome.... (p. 67)

In romantic love, according to Kernberg, the Oedipus complex is finally conquered. The man finds he can love his mother (or a representation thereof) **without** killing his father; and in this way he unlearns the identification of all love with maternal love. The triumph of getting the woman is there, but without the violent Oedipal overtones. Similarly, the woman finds she can love her father (or a representation thereof) without violence or threat to her mother. Romantic love, in this view, is what finally liberates the mind from the shackles of childhood sexuality.

Going beyond the purely Freudian framework, Kernberg ties his conclusions in with self theory:

> [T]here is a basic, intrinsic contradiction between two crucial features of sexual love: the necessity for a self with firm boundaries, and a constant awareness of the separateness of others, contradicts the necessity for being able to transcend the boundaries of the self so as to become one with the loved person. Sexual passion integrates these contradictory features: the shared experience of orgasm includes the transcendence from the experience of the self into that of the fantasied union of the oedipal parents, as well as the transcendence of the repetition of the oedipal relation to an abandonment of it in a new object relation that reconfirms one's separate identity and autonomy. (70)

The transcendence of the oedipal relation is seen to tie in with the transcendence of the dichotomy between aloneness and togetherness. Namely, orgasm promotes togetherness and, in Freudian language, identification with the union of the oedipal parents. But at the same time, the abandonment of the oedipal parents and the entrance into adult sexuality leads one to a new and powerful self-image. Paradoxically, by giving up oneself and melding with another, one achieves a stronger self.

Love and Attachment

Kernberg's Freudian analysis gives the beginning of an explanation of the emergent quality at the center of romance -- that "something special" that arises from the combination of sex and love in the proper way. Clearly, as Kernberg suggests, this emergent quality has something to do with the transcendence of the distinction between self and non-self, and with simultaneously reliving and shedding one's early relationship with one's parents. In order to really understand what is going on, however, we must leave the Freudian terminology behind. Ultimately, there is no need to invoke the Oedipus complex; one may arrive at conclusions similar to those of Kernberg from totally different considerations, while at the same time obtaining deeper insights into the nature of romantic love. Shaver, Hazen and Bradshaw (1988) have given a convincing and modern analysis of romantic love as a combination of three behavioral systems: an **attachment** system, a **caregiving** system, and a **sexuality** system. They focus most of their attention on the attachment system. Essentially, their "attachment system" plays the role ofKernberg's Oedipus complex: it is a childhood experience which is simultaneously re-enacted and transcended through the experience of romantic love.

They define the attachment system as

a set of behaviors (crying, smiling, clinging, locomoting, looking, and so on) that function together to achieve a set-goal, in this case a certain degree of proximity to the primary care giver. The set-goal changes systematically in response to illness, pain, darkness, unfamiliar surroundings, and so on -- all conditions associated with potential harm.

As other examples of behavioral systems, besides caregiving, sexuality and attachment, they cite mating, affiliation, and exploration. In the language of the psynet model, a "behavioral system" would be a certain kind of **autopoietic subnetwork** within the dual network. Specifically, behavioral networks are autopoietic subnetworks that exist on a fairly low level of the dual network, and contain a high proportion of processes concerned with **action**. The fact that they "function together to achieve a set-goal" is a consequence of their autopoiesis: the different processes in the system work so closely and effectively together that, if one were removed, the others would quickly use their ingenuity to reinvent it, or something similar to it.

Shaver, Hazen and Bradshaw give a very convincing point-by-point comparison of features of attachment and adult romantic love. This comparison is so convincing that I see no alternative but to give a briefer, paraphrased version here. For sake of conciseness I will borrow their terminology of AO for "attachment object" and LO for "love object":

-- Formation and quality of attachment bond depends on AO's sensitivity and responsiveness. Love feelings are related to intense desire for LO's interest and reciprocation

-- AO provides a secure base so that infant feels competent and safe to explore. LO's reciprocation causes person to feel secure, confident, safe, etc.

-- When AO is present, infant is happier, less afraid of strangers, etc. When LO is viewed as reciprocating, the lover is happier, more outgoing, optimistic and kind.

-- When AO is not available, or is not responsive, the infant is anxious and preoccupied. When LO acts uninterested or rejecting, the lover is anxious, preoccupied, unable to concentrate, etc.

-- Attachment behaviors include: proximity and contact seeking, holding, touching, caressing, kissing, rocking, smiling, crying, following, clinging, etc. The same behaviors are present in love relationships, augmented of course by the act of lovemaking.

-- When afraid, distressed or sick, infants seek physical contact with AO, and lovers with LO.

-- The result of separation from AO or LO is distress, followed by sad and listless behavior if reunion seems impossible.

-- Upon reunion with AO, the respond is smiling, positive vocalizations, bouncing, jiggling, etc. Upon reunion with LO, or when LO reciprocates after reciprocation was in doubt, the reaction is physical ecstasy, the desire to cry out, to hug and be hugged, etc.

-- The infant shares toys, discoveries, etc. with AO; the love shares experiences and gifts with LO.

-- Infant and AO share prolonged eye contact; so do the lover and LO. Infant seems fascinated with the different parts of AO's body -- nose, ears, hair, etc. Lover seems fascinated with the different parts of LO's body.

-- Infant feels fused with AO and, with development, becomes ambivalent about the balance of fusion and autonomy. The same is true of the lover and LO.

-- An infant generally has only one AO at a time; a lover generally has only one LO at a time.

-- Separations and adverse circumstances, up to a point, tend to increase attachment behavior.

-- Communication with AO takes place largely in a "private language" of coos, songs, soft voices and body language. The same is true, to a lesser extent, of communication with LO.

-- Powerful empathy; an almost magical ability to sense the other's emotional needs.

-- Infant experiences AO as omniscient and omnipotent. Similarly, in the early phases of love, the LO is perceived as a perfect being; all LO's flaws are systematically overlooked.

-- When the relationship is not going well, cues of AO's or LO's approval or disapproval are the cause of highly sensitive reactions.

-- The infant's greatest happiness comes from AO's approval and attention. The lover's greatest happiness comes from LO's approval and attention (as Flaubert wrote in *Sentimental Education*: "[She] was the focal point of light at which the totality of all thins converged.").

-- When the loss of a spouse occurs, the typical reaction is similar to separation distress in infancy, including uncontrollable restlessness, difficulty in concentrating, disturbed sleep, anxiety, tension and anger.

No one who has been in love can deny the uncanny accuracy of these parallels. Indeed the parallel between love relationships and parent/infant relationships is a staple of popular music and literature. Countless pop songs refer to lovers as "baby," "mommy" or "daddy." Ronald Reagan, even while President, called his wife Nancy "Mommy."

A little more speculatively, the attachment theory even provides an explanation for why romantic love seems to fade after a few months or years. According to Bowlby (1969),

By most children attachment behavior is exhibited strongly and regularly until almost the end of the third year. Then a change occurs.... In many children the change seems to take place almost abruptly, suggesting that at this age some maturational threshold is passed. (p. 204-205).

The attachment theory of love suggests that this conclusion may perhaps transfer over to romantic love relationships. Once one is sure of being loved, one feels security rather than passionate uncertainty, and, acting on the cue of one's earlier attachment relationship, one is impelled to move on. Curiously, the time period most often associated with the duration of romantic love is also three or four years.

Despite the long and impressive list of similarities, however, there are obvious differences between infantile attachment and romantic love. These are due, in large part, to the other two behavioral systems involved: caregiving and sexuality.

First of all, love is a largely symmetrical relationship, while the infant's attachment to its mother is completely asymmetrical. During early childhood, the infant's attachment system is complemented by the mother's caregiving system. If either system is deficient then problems will result: a baby who displays insufficient attachment will tend not to be properly cared for; and a mother who gives insufficient care will give rise to a person with abnormal attachment behaviors.

In a love relationship, **both** partners are generally expected to give both attachment and care.

This caregiving behavioral system is not something that we very often think about. But, once one looks for it, it is extremely obvious in everyday life. Who has not marvelled at the uncanny affection for babies that so many of us seem to display? Why should we love these little creatures who urinate and defecate all over themselves, who display less intelligence than household pets, who spend most of their time either sleeping or crying? These qualities in an adult or an older child would be repulsive. The seemingly innate affection for babies is a consequence of the caregiving system, which is "activated in a parent when his or her child displays attachment behaviors.... [I]t includes sensitivity and responsiveness, behaviors aimed atsoothing the child (holding, patting, rocking), and attempts at problem solution." (p. 86)

There may be more or less asymmetry in caregiving in a love relationship, both in quantity of care given and in kind of care given. Stereotypically, the male gives financial and protective care, while the female gives emotional care. Females conventionally display a greater amount of attachment behavior; but males are also capable of extreme displays of attachment. The one sure thing is that no love relationship has quite so much asymmetry as the infant/mother relationship!

Much better understood than the caregiving system is the **sexuality system**. Freud and other psychoanalysts have argued that the infant-parent relationship is partly based on sexuality; but even if this claim were true, sexuality would still form a major difference between love and infantile attachment, because the roles of sexuality in the attachment and love relationships would necessarily be very different. Kernberg, quoted above, has given one view of what exactly this difference in roles might be.

We have already dwelt, in the previous chapter, on the inherent gender asymmetry of the human sexual system. But this asymmetry, it would seem, is not particularly important for romantic love, except insofar as it helps to determine the type of individual to whom a given person will feel sexually or socially attracted (a topic that will be taken up below). More crucial to the actual dynamics of romantic love is the **tactile** aspect of sexuality. An infant experiences very frequent skin contact with the mother -- indeed, before birth, the contact is constant! As we grow older, we lose this physical closeness, and we touch others only for short periods of time, as in handshakes, hugs, and physical sports. Having sex, and sleeping next to a sexual partner, are really the only occasions we have to touch our bodies against others' bodies. The tactile aspect of sex "supercharges" the attachment system in a way that can hardly be overestimated.

Romantic Love as an Autopoietic Magician System

Although Shaver, Hazan and Bradshaw do not emphasize this point, it seems to me that one of the most essential aspects of romantic love is the synergy **between** the attachment, caregiving and sexuality systems. These three autopoietic systems link and eventually merge together to form a larger autopoietic supersystem. This supersystemic autopoiesis is, I suggest, an important part of the "magic" of love.

As a first approximation one may think of this autopoietic supersystem as a joining together of the three component systems. But this is not entirely an accurate view; it overlooks the dynamic nature of the dual network. Eventually, acting in such close proximity, the various networks will wind up swapping subnetworks with each other. They will change one another, and overlap with one another, so as to produce a final product which is not the same as the union of the three components; it is a fundamentally new system, formed by mutation and crossover and autopoietic interproduction of the parts of the three components. The attachment, caregiving and sexuality systems merge together to form something different, a behavior system for romantic love.

For sake of simplicity, however, it is convenient to ignore this merging together for the moment, and to think about this supersystem as a union of three disjoint components. Looking at the supersystem this way provides a clear insight into the nature of the **autopoiesis** involved. First of all, one concludes that the supersystem is not autopoietic in the strictest sense. For it is

not true that each component is producible by the other two components. For instance, the attachment and caregiving behavior systems, by themselves, are not capable of producing the sexuality system. What is true, however, is that, given just a low-level activation of the sexuality system, the attachment and caregiving systems are capable of producing a high level of activation of the sexuality system. Similarly, given a low-level activation of the attachment system, thesexuality and caregiving systems are capable of producing a high level of activation of the attachment system; and, given a low-level actiation of the caregiving system, the sexuality and attachments systems are capable of producing a high level of activation of the caregiving system. Thus, even ignoring the process of merging that will inevitably occur, if one assumes a low base level of activation for all three systems, one concludes that the supersystem is effectively autopoietic: high levels of activation in the three systems are adaptively inter-perpetuating.

One way to understand this supersystemic autopoiesis is to construct "archetypal love stories." Here is an example. Let's say one partner, Sally, is attached to the other one, Jim. This attachment will soon lead to distress if it is not met with caregiving. So unless Jim responds with caregiving behavior there will be no two-way romantic love. But why, then, not a totally asymmetric relationship, in which Sally is attached and Jim is caregiving? This is where sexuality comes in: Sally and Jim become involved sexually, which involves a great deal of physical contact and other behaviors which, to Jim, are naturally reminiscent of the attachment relationship of his early childhood. Thus Jim has a natural incentive to develop attachment behaviors as well. And once Jim develops attachment behaviors, Sally must respond in kind with caregiving behaviors -- the system is complete.

This little scenario indicates how sexuality can help turn one-way attachment/caregiving relationship into a full romantic love relationship involving all three systems of both partners. Of course, many other scenarios are possible. For instance, the start can be a sexual relationship; the intimacy of sex can then lead to an increasing spiral of attachment and caregiving behavior. Attachment and caregiving, due to their complementarity, have a natural potential to build upon one another in the manner of a "feedback loop"; sexuality, it is easy to see, provides an environment in which this is particularly likely to occur.

For further examples of this kind of supersystemic autopoiesis, one may look in any one of the tens of thousands of love stories that have been published over the last few thousand years. There is an immense variety to the details, but the upshot is always the same.

Intuitively speaking, the experience of falling in love has all the earmarks of **convergence to an attractor**. Once a few of the pieces are there, the others seem to fall into place like magic: all of a sudden, "Wow!", and the full-on feeling of romantic love is there. The idea of love as an autopoietic, evolving supersystem makes this intuition precise: it specifies what kind of attractor the "falling in love" dynamic converges to.

Patterns of Attraction

So love is an autopoietic magician system. The sixty-four million dollar question, however, is: what makes Jack fall in love with Jill. Or Jill with Johnny? What makes an individual fall in love

with **this particular person** instead of someone else? Why is the autopoietic, evolving supersystem so darn picky?

In the last few decades, a fair amount of empirical research on this question has finally begun to appear. None of it, however, really seems to get at the heart of the matter. The reason for this, I believe, is that most of the research has focused on the **traits** of the two individuals involved in a love relationship. And love, I will argue below, has less to do with traits of individuals than with phenomena that **emerge** between two individuals, phenomena that are present only when the two individuals interact together.

Love, I will argue, is a matter of **eliciting intimacy**, where intimacy is defined as the repeated, shared creation of patterns of mutual emotional meaning. More technically speaking, what this means is that in order for romantic love to occur, the different subsystems of the different subselves of the two individuals involved must produce emergent adaptive autopoieticmagician systems which permit the sustained co-creation of emotionally significant patterns. This is quite a mouthful; it is a complicated requirement -- and this, I propose, is exactly **why** it is so hard to tell who will fall in love with whom. For better or for worse, the romantic love system is a complex one!

Before discussing emergence and intimacy, however, I will first review some of the most important facts that have been discovered regarding the psychology of love, mostly drawn from the book *Love's Mysteries*, by Glenn Wilson and David Nias (1976). It is important to know exactly **how much** can be explained by traits, before moving beyond them.

On a very general level, one may isolate certain qualities that are considered desirable in a love partner. According to one study, for males these are: achievement, leadership, occupational ability, economic ability, entertaining ability, intellectual ability, observational ability, common sense, athletic ability, and theoretical ability. For women, on the other hand, the same study found quite a different list: physical attractiveness, erotic ability, affectional ability, social ability, domestic ability, sartorial ability, interpersonal understanding, art appreciation, moral-spiritual understanding, art-creative ability. The most important gender difference regards physical appearance: it predicts dating probability much better for women than for men (a 62% correlation versus a 25% correlation). Both sexes, however, like attractive partners!

The folklore of romance includes two contradictory maxims: "like attracts like," and "opposites attract." But many studies indicate that, even for the selection of marriage partners, neither similarity nor complementation predicts very much. According to standard methods of profiling personality, individuals tend to marry almost, if not quite, at random. On average, there is a slight tendency for people to like partners with similar attitudes. And, as might be expected, attitude similarity is more important for choosing marriage partners than for dating partners. But these are not dramatic tendencies.

In one study, which dealt with married couples who had met through a computer dating service, the following results were found. Couples were more likely to marry if they were of similar height, had a similar degree of interest in sport, and were similar in concreteness of ideas, serious-mindedness, confidence, control and dominance. Complementation appeared only in

relation to wit: a jokester tends to marry a non-witty person. Most notable among the results of this study is the absence of complementarity in the area of dominance versus submission. On the contrary, a slight similarity effect was found: dominants tend to marry dominants, and submissives tend to marry submissives. On the average, women are attracted to dominant men, but only if they are competent. Dominant, incompetent men are liked even less than submissive, incompetent ones.

Some interesting **cognitive** phenomena may be observed in the dynamics of romantic attraction. For instance, it is known that, if a man erroneously believes that a certain woman has excited him, he will tend to feel a greater attraction to that woman than he would have felt otherwise. This observation resonates well with Gosselin and Wilson's cognitive theory of the origin of sexual deviance, discussed in the previous chapter, according to which, e.g., masochism evolves from the chance association of physical pain with sexual excitation.

And there is the "gain" phenomenon, according to which a person will be most strongly attracted to someone whose negative opinion of them has suddenly turned positive. This is even better than someone whose opinion has been consistently positive. And someone whose positive opinion has changed to negative is even less attractive than someone whose opinion has been negative throughout. This phenomenon may be viewed as a validation of Paulhan's view of emotion, according to which happiness is a feeling of increasing order, and unhappiness a feeling of decreasing order. Good feelings are caused by a positive **rate of change** more than a constant positive value. Bad feelings are caused by a negative **rate of change** more than a constant negative value.

Along the same lines as the gain phenomenon, marriages that appear to be "good" -- based on mutual support, admiration and kindness -- will sometimes fall apart for no apparent reason. The lack of change leads to a lack of happiness. Marriages involving a fluctuation between negative and positive relations will often be felt as more exciting and stimulating, because the necessary positive and negative gradients are there. But of course, this kind of fluctation is not necessary for a successful marriage: the feeling of increasing order can come from other places than improving relations. For example, returning from a hostile workplace to a loving home can make one's spouse seem particularly attractive: here the gradient is caused by a changing environment rather than a changing relationship.

These cognitive and trait-theoretic investigations are perfectly interesting, both as scientific results and as guidance for real life romantic affairs. But they stop short of giving an answer to the **real** question. The reason is, I suggest, that they fail to come to grips with **intimacy**. Not all relationships are equally intimate, but falling in love, I propose, requires a high degree of intimacy. And intimacy is something that cannot be understood in terms of traits or cognitive errors. It has to do with subself dynamics, with the high-level interaction of self and reality theories. In short, it is complex.

Intimacy

Successful thought systems, according to the psynet model, survive for two reasons. First, because they are autopoietic. Second, because they produce patterns that are useful for other thought systems.

I have said that the attachment/caregiving/sexuality supersystem is autopoietic. But this is only half of the story; it does not answer the question of the **external** purpose of romantic love. The answer to this question, I propose, is **intimacy**. Intimacy is the means by which romantic love produces useful patterns.

Clearly one may fall in love without experiencing intimacy. After all, what about "love at first sight"? But a sustained experience of being in love is, I claim, not possible without a reasonable degree of intimacy. An attachment/ caregiving/sexuality supersystem will not survive in the dual network without intimacy to support it. Love at first sight is mostly an **anticipation** of love; without the intimacy to back it up, it eventually fades away.

Intimacy is notoriously difficult to define. Two reasonable definitions are:

I am defining intimacy as **the experience of personal and relational affirmation and enhancement that derives from interactions demonstrating reciprocal knowledge and validation between partners**. (Rampage, 1994)

[Intimacy is where] people share meaning or co-create meaning and they are able to coordinate their actions to reflect their mutual meaning-making. (Weingarten, 1992)

Building on these previous definitions, I propose to define intimacy as **the systematic co-creation of patterns of mutual emotional significance**. And I propose that there is, in each person, a fairly cohesive **intimacy system**: an autopoietic pattern/process network oriented toward making intimate acts, remembering intimate acts, thinking of intimate acts, and perceiving intimate acts on the part of others.

The intimacy system is not as basic as the attachment, caregiving and sexuality systems, because it exists on a higher level: it has to do with fairly abstract emotional patterns, instead of simple physical acts. But it is real nonetheless. One may look at love as a process of linking up the **lower-level** attachment/caregiving/sexuality supersystem with the relatively **higher-level**intimacy system. The combination of these **four** systems gives a full-fledged **romantic love supersystem**.

In Buberian terms, we may say that intimacy is a special kind of I-You relationship. It is a pattern of interpersonal relating which leads to particularly strong and sustained I-You encounters. Intimate relationships cause two individuals to create meaningful patterns together; in this way the two individuals become linked so closely that a recognition of one another's emergent wholeness becomes almost second nature. The intimacy system may thus fairly be viewed as a system for getting and holding onto I-You relationships. We may thus conclude, on purely formal grounds, that **intimacy tends to lead to repeated and intense I-You encounters**. Thus, according to the Fundamental Principle, intimacy tends to lead to healthy minds: to healthy subself systems spanning two physical bodies.

One common example of intimacy is the emotionally revelatory conversation. This example fits well into my definition. The conversation itself is a mutually created linguistic pattern, and if it is not emotionally meaningful to both participants, then it is not intimate. It is all right if only one partner is revealing themselves, but what is necessary in this case is that the other one must visibly **empathize**; the empathy is a sign of mutual emotional significance.

Another commonplace example of intimacy is sex. Clearly, sex need not be intimate, because it need not have mutual emotional significance. Sex with a prostitute is generally not intimate. But sex does involve the systematic co-creation of patterns -- patterns of bodily motion. Some things are communicated far better by body movements than by words. If mutual emotional significance is there, obviously, sex will be intimate.

Note that illusory intimacy is possible, in the sense that one of the people involved may falsely **believe** there to be patterns of mutual emotional significance involved, when in fact the other person is not emotionally moved at all. In this sense there can be one-way romantic love without true intimacy. There is systematic co-creation of patterns of one-sided emotional significance, with the illusion of mutual emotional significance. But this is the "exception that proves the rule."

The connection between the attachment/caregiving/sexuality supersystem and the intimacy system is one of mutual support. On the one hand, intimacy gives the tripartite supersystem a **purpose**, a purpose outside itself. By giving rise to intimate behavior -- i.e. connecting with the intimacy system -- the supersystem creates new patterns, which are potentially useful to the remainder of the mind. Thus it goes beyond its autopoiesis and justifies itself as a "useful belief system." Intimacy guarantees the "fitness" of the tripartite supersystem as a pattern processing system within the mind.

But, on the other hand, the intimacy system is at least as desperately in need of the tripartite supersystem for its survival. For "intimacy," in itself, is a very abstract thing. One may have certain habits for extracting personal information to share with others, for listening and responding to others' confessions, for working together with another to produce something of mutual meaning. But these habits will be useless unless one finds oneself in an appropriate **context**. The tripartite supersystem helps to create appropriate contexts for intimacy.

And the tripartite supersystem is peculiary appropriate as a "bottom end" or "user interface" for the intimacy system. For, as is well known, mental states of great emotional intensity, such as those which one achieves through intimacy, often cause one to "lose control" (i.e. shed one's usual adult behavior regulation systems). They put one in a dependent state, in which one becomes easily attached, in which one requires caregiving. Therefore intimacy may be understood as giving rise to a direct need for active attachment and caregiving systems. The attachment/caregiving/sexuality supersystem matches up with the intimacy system to form an adaptive autopoietic supersystem of great beauty and power.

Love and Subself Dynamics

Intimacy cannot be fully understood without referring to **subselves**. For, the cooperative construction of patterns by two individuals is always really done by two subselves. To oversimplify things for sake of argument, let's say that Jane and John each have three prominent subselves. Then, for romantic love between Jane and John to be successful, it need not be the case that each subself of Jane's can easily enter into an intimate relationship with each subself of John's. What should ideally be the case is that each subself of Jane's, and each subself of John's can enter into an intimate relationship with **at least one** subself of the other person. Thus no subself will be left without the experience of intimacy.

It is important not to misunderstand the relationship between subselves and the romantic love supersystem. There is not one fixed supersystem, invariant between all subselves. But neither does each subself have its own independent supersystem. Instead, in the typical situation, each subself will possess a certain "projection" or "interpretation" of the same romantic love supersystem. Another way to say this is that the romantic love supersystems associated with the different subselves will generally tend to be quite similar to one another. The autopoiesis involved in the romantic love supersystem is largely subself-independent; but the supersystem is also involved in complex self-organizing relationships with processes specific to individual subselves.

The number of possible subself relationships is staggering. For instance, some of Jane's subselves may hate some of John's subselves, and this may cause serious problems, or else it may be prudently managed. Jane or John may evolve dynamics according to which, at the point of conflict, the troublesome subself fades into the background.

Let's give our example more detail. Suppose Jane has three important subselves: one vulnerable, emotional, irrational and creative; another logical, sensible and ambitious; and a third, cold, unfriendly, and moderately reasonable except in regard to other people. Suppose John has three important subselves: one vulnerable, emotional, irrational and creative; another logical, witty and extremely intelligent; and a third, extremely outgoing, sarcastic, nihilistic and jovial. Call these subselves Jane1, Jane2, Jane3 and John1, John2, John3. Each of these "subselves" is a set of behavior patterns, an autopoietic subnetwork of the dual network consisting of self and reality theories and lower level processes to support them. Each is a complete and self-contained system for dealing with the world.

Now, John1 and Jane1 match up in a natural way, as do John2 and Jane2. These pairs are extremely likely to be intimate with each other. John1 and Jane2 generally get along acceptably, as do John2 and Jane1. They are often intimate, the logical one of the pair one providing the emotional one with support and advice. But while Jane3 tends to tolerate John1 and John2 without too much complaint, she is extremely hostile to John3: in her cold and antisocial state Jane is offended by John's sarcasm. On the other hand, John1 is often hurt by Jane3; in his vulnerability, he cannot bear her coldness. And, similarly, Jane1 is often hurt by John3; in her vulnerability she is wounded by his sarcasm.

In this example we have the potential for a deep and intense love relationship. First of all, Jane1 and John1, both vaguely "female" in characteristics, can share their fears and ideas and feelings with each other in an intimate way. They both display strong attachment behaviors,

which may often lead to sexuality behaviors and caregiving behaviors. The romantic love system involving these two subselves, though involving all three components, is dominated by attachment. The sexuality involved here is one of mutual "feminine" behavior, mutual passivity; one would expect extensive foreplay.

Next, Jane1 and John2, or Jane2 and John1, fall into a natural attachment/caregiving relationship. The sexuality involved here is of a classic man/woman type, where the number 2 subselves take on the stereotypically male role.

And Jane2 and John2, with their lessened need for attachment, still do have some need for intersubjective validation and intimacy. Their romantic love supersystem is less intense, less prominent, but still perfectly healthy. The sexuality involved here is one of mutual "masculine" behavior, mutual activity; one would expect vigorous intercourse without intense caressing or foreplay.

Finally, the other combinations do not display romantic love supersystems at all. The reason is that Jane3 and John3 are basically incapable of caregiving behavior, and display only very limited attachment behaviors. They may partake in sexual behavior but it will not be emotionally intense; it is unlikely to be "intimacy."

The **compatibility** of Jane and John, in this example, is determined largely by the **transitions** between one subself and the other. What is crucial is that the combinations Jane1/John3, Jane3/John1, and Jane3/John3 should not be allowed to coexist. First, either Jane1 must learn to phase out when John3 comes along, or John3 must learn to phase out when Jane1 comes along. Similarly, either John1 must learn to phase out when Jane3 comes along, or Jane3 must learn to phase out when John1 comes along. Finally, either John3 or Jane3 must learn to disappear when the other one appears.

Ideally speaking, it is even possible that the love relationship may help to "cure" the unhealthy subselves, Jane3 and John3. For instance, John2 may be able to offer caregiving to Jane3, and, not being so vulnerable as John1, he may be able to do so in spite of her apparent rejection. This caregiving may cause Jane3 to phase out and yield to Jane1 or Jane2. If this phasing-out occurs often enough it may cause atrophy of Jane3's internal processes, and thus lead to the dissolution of Jane3 as an independent subself. Jane3 may turn into a mere "mood" which quickly passes into Jane1 or Jane2.

Similarly, Jane2 may be able to offer caregiving behavior to John3, and thus cause John3 to phase into John2 or John1. The reason caregiving behavior may be expected to cause the phasing-out of non-intimate subselves is precisely because of the romantic love supersystem. The caregiving behavior will tend to activate attachment behavior and sexuality behavior, because of the autopoietic nature of the supersystem. Even when the subself in question does not partake of the romantic love supersystem, the supersystem still exists in the background and, unless the division between subselves is particularly severe, it can still be activated.

On the other hand, the love relationship may also cause the partners' subself dynamics to take a turn for the worse. A typical pattern would be for a particularly painful experience to cause

Jane1 to pass into Jane3, or John1 to pass into John3. In other words, the non-intimate personalities may arise as "defense mechanisms," as overreactions to the wounded vulnerability of the feminine personalities. Thus, for example, the emergence of Jane1/John3 may lead immediately to the emergence of Jane3/John3. And if this occurs often enough, it may become ingrained in Jane1, so that Jane's moods of creativity and emotionality habitually shift into moods of coldness, for fear of the pain of continued vulnerable emotionality.

Of course, this is just a simplified model. In reality subselves are not this clearly delineated. Each subself is fuzzy and therefore the number of subselves is fuzzy as well. Some of the subselves will be vaguely divided into subsubselves. And there will be "superpositions" of subselves, intermediate states in which one subself controls most actions but another subself retains control of a few. But **any** model has its limits. The subself dynamics model, combined with the romantic love supersystem and the idea of intimacy, provides a much deeper view of love relationships than has previously been available. It provides a way of giving detailed analyses of love relationships, without recourse to unscientific concepts or purple prose.

Finally, it should be said that, in the subself dynamics view, there is cause to take a positive view of romantic love. It can hurt, yes, but it promotes the creation of an emotionally potent intersubjective world. It is a part of the larger process by which we all work together tocreate a world we can meaningfully share. According to the Fundamental Principle, insofar as it promotes intimacy, romantic love supports healthy subself systems. The mind involved in romantic love is involved in at least one, usually multiple I-You subself relationships; and this promoted healthy mental functioning. Thus, **the system of two lovers is, as a system, a healthy "mind."** Of course, either mind taken separately may not be healthy; the I-You dynamics may be disrupted by the end of the love relationship, leading to severe depression.

The Freudian View Revisited

Finally, we are ready to return to Kernberg's Freudian analysis of romantic love, discussed near the beginning of the chapter. Recall that, in Kernberg's view, the main features of romantic love are the overcoming of the Oedipus complex and the transcendence of the boundary between self and other. It is interesting to see how these features are reflected in the system-theoretic model of love that I have given here.

Clearly, intimacy provides a feeling of boundary-transcendence. Together with another, one is creating intensely emotionally meaningful patterns, which is something one normally does alone. Subself dynamics provides an even stronger possibility for boundary-transcendence in the context of love relationship -- it is possible for two subselves of **different** people to feel more closely bound together than two subselves that live within the same person. For instance, Jane1 may well have more in common with John1 than she does with Jane2 or Jane3. She may also co-create more mutually emotionally meaningful pattern with John1 than she does with her own co-selves. If John1 is similarly closer to Jane1 than to his own co-selves, then one has a situation where the two people, taken together, form a **cohesive personality system**. The unit of two people, taken as a collection of subselves, is more cohesive than either individual person. The dynamics of pattern has overcome the limitations of the physical body.

Regarding the Oedipus complex, the situation is slightly more complex. The psynet model itself is not incompatible with Freud's version of the Oedipus complex; however, the two models do not come together in any particularly elegant way. It is much more natural to re-frame the complex in terms of subself dynamics. When one does so, one concludes that the transcendence of parental anxieties and conflicts is a **common** consequence of romantic love, but not a **necessary** one.

The child will base her/his behaviors on observations of both mother and father, and will generally evolve subselves modeled on each. The tendency to identify with the same-sex parent will in general be greater, so that the subself (subselves) corresponding to the opposite-sex parent will continually be struggling for power. In some cases, this struggle may give the appearance of a desire to "murder" the same-sex parent within oneself. We will see this in the following section, where masochism is analyzed as a strategy used by female subselves of male people to maintain control over their competing male subselves.

Romantic love re-activates the attachment system, which has been lying largely dormant since early youth. In doing so it takes a large number of processes which previously were optimized for interaction with the **parents**, and modifies them to be more useful for interaction with the **lover**. The attachment system is integrated with the sexuality and caregiving systems, and in this process the attachment system itself is changed. Anything else which was connected with the attachment system is also open to change at this point -- it is not uncommon for an intense new love affair to cause fundamental changes in a person's behaviors. This reorganization of the attachment system is, I suggest, responsible for much of the transcendence of parental relationships which Kernberg identifies with a transcendence of the Oedipus complex. In Buberian terms, the transcendence of parental attachments is just the attainment of the ability to have I-You relationships with individuals other than the parents.

But there is more to it than that. Just as the attachment system lies largely dormant from early youth until the onset of romantic love, so there are many other intimate behaviors which are rarely exercised except in parent/child or lover/lover relationships. As a small child, one is exceedingly intimate with one's parents, not only physically, but also in terms of sharing one's feelings and experiences. This intimate relationship is sometimes duplicated in later youth and adulthood with same-sex or different-sex friendships. But it rarely is duplicated so effectively as in an intense romantic love relationship. Just as the attachment system is "reawakened" by romantic love, so is the **intimacy system**. And everything connected with the intimacy system is therefore primed for reorganization. In particular, there is a powerful **incentive** for changes in the attitudes of, and relations between, one's male and female subselves (though there is no guarantee of such change).

Subself dynamics, as in the example of Jane and John given above, suggests that love will be most powerful when it provides intimate interaction for both the **male** and **female** subselves of each partner. If this kind of intimacy is not provided then the love relationship will be felt to be "missing something" (the only exception to this rule would be a couple in which one or more of the partners had no opposite-sex subselves, or only very weak opposite-sex subselves; in this case the relationship is not felt to be missing something, because the individual her- or him-self is missing something). Thus romantic love ideally provides a context in which subselves of both

sexes may learn, grow, interact and express themselves. By this evolution and revision of the intimacy system, subselves may be freed from old patterns molded by intimate interaction with the parents.

So, in sum: romantic love reawakens and reorganizes the attachment and intimacy systems, therefore disrupting patterns and processes that have been stable since early youth. It therefore provides a golden opportunity for overcoming the problems of one's early relationship with one's parents. If the Oedipus complex is real, then it falls into this category, and one concludes that Kernberg is correct: romantic love provides a means for overcoming the Oedipus complex. But, even if the Oedipus complex is a rare or nonexistent phenomenon, the role of romantic love in overcoming childhood problems is still a potentially very important one. Like many other Freudian analyses, Kernberg's theory of love points one in interesting directions, regardless of whether one accepts its ultimate conclusions.

14.3 THE DEVELOPMENT AND DYNAMICS OF MASOCHISM

Now let us turn from romantic love to a phenomenon which, though far rarer, has attracted far more attention from clinical psychologists over the past century: **masochism**. In the 108 years since Krafft-Ebing published *Psychopathia Sexualis* (1886/1965), numerous authors have probed the psychology of this apparently paradoxical phenomenon. But no one, it would seem, has solved the central puzzle of masochism, which is: how can certain people derive pleasure from pain? In this paper I will approach the problem from a new angle, using the system- theoretic approach to personality. The result is a novel analysis which synthesizes many of the previous ideas on masochism.

By way of introduction, I will not attempt a comprehensive review of the large literature on masochism, but will mention only the two theories that are most relevant to the complex systems model that I will present here. These are the classic ideas of Freud, and the more recent cognitive explorations of Gosselin and Wilson (1985).

In the *Three Essays on Sexuality*, in 1905, Freud commented that "it can often be shown that masochism is nothing more than sadism turned around on itself," but concluded that

no satisfactory explanation of this perversion has been put forward and it seems possible that a number of mental impulses are combined in it to produce a single resultant.

adding that sadism and masochism are habitually found together, so that he is

inclined to connect the simultaneous presence of these opposites with the opposing masculinity and femininity which are involved in bisexuality.

In the 1924 essay "The Economic Problem of Masochism," Freud revised his opinions somewhat and distinguished two kinds of masochism, **primary** and **secondary**. Secondary masochism is sadism which turns back on itself due to lack of any other outlets. Primary, or erotogenic masochism can take two forms, **feminine** and **moral** masochism. Moral masochism is what Freud described much earlier, in *Civilization and Its Discontents*, with the following words:

The sense of guilt, the harshness of the super-ego is ... the same thing as the severity of the conscience. It is the perception which the ego has of being watched over in this way, the assessment of the tension between its own strivings and the demands of the super-ego. The fear of this critical agency (a fear which is at the bottom of the whole relationship), the need for punishment, is an instinctual manifestation on the part of the ego, which has become masochistic under the influence of a sadistic super-ego; it is a portion ... of the instinct towards internal destruction present in the ego, employed for forming an erotic attachment to the super-ego. (136)

Moral masochism, more simply, is taking joy in the torture of castigating oneself. In some cases it also take the form of pleasure of being punished by others; the external tormentor is then a symbol for one's own super-ego.

Feminine masochism, on the other hand, is the kind of masochism demonstrated by Leopold von Sacher-Masoch in his infamous stories of submissive males begging dominant females to whip them and humiliate them. The typical feminine masochist will profess a desire to play the role of a slave or a servant, or a dog or horse, or even to wallow in urine or feces, or menstrual blood. Freud suggests that, at the deepest level, the feminine masochist believes himself to be a woman. As Silverman (1992) puts it, "it is not only at the level of his sexual life, but at that of his fantasmatic and his **moi** that the male masochist occupies the female position."

A Cognitive Theory of Masochism

The Freudian model of masochism is an interesting one and, as we shall see, is reflected in the structure of the complex systems model to be presented below. A more detailed model of masochistic psychology, however, is provided by Gosselin and Wilson, in their book *Sexual Variations*. Gosselin and Wilson seek to provide a unified cognitive treatment of all deviant sexual behavior, with a focus on fetishism, sadomasochism, and transvestitism. For instance, they explain the psychological origins of fetishism as follows:

At some point, the young child is at a high level of arousal.... At that moment, there is also present an element of the fetish-material-to-be: apron, baby pants, crib sheet, mackingosh-as-shield-from-rain-and-cold. The male child notices the stimulus more readily than the female child and connects it with the high arousal state, which may or may not have any direct sexual connotation about it.... The message recorded simply says atthat stage, "Possible link between **that** material and **that** excited feeling." The youngster is in fact making a miniature scientific hypothesis that certain qualities in a fabric are associated with a particular feeling.

When the fabric turns up again, the child remembers the previous association and says, in effect, "Let me test my hypothesis by searching for that feeling." Unfortunately he often makes a mistake which is common in all of us. If he notices no excitement, his verdict is "not proven, but my hypothesis may still be right," simply because nobody likes to admit, even to themselves, that they are wrong. If he does feel excitement -- even coincidentally, or because he expected it and therefore felt it -- his verdict is that the hypothesis has been proved. An expectancy is thus strengthened that "next time it will work the same way" and confirmation is almost bound to be obtained on subsequent occasions because the reaction becomes a self-fulfilling prophecy. (161-162)

This cognitive theory develops an obvious theme which may be traced back at least as far as Binet, namely, that fetishism develops due to early sexual contact with the fetish material. But Gosselin and Wilson's argument goes beyond this simple idea. For example, it explains why males should develop more fetishes than females -- because males have, in their erections, a far more obvious indicator of their own level of excitation. Thus males will be much more likely to make the conjectured associations.

Surely Gosselin and Wilson's assumption about the origins of fetishism in the juvenile mind is not an implausible one. And, as they point out, a similar argument can be made for masochism:

In sadomasochism the basic experience of being physically punished as a child, probably followed by a reassuring cuddle to show that "Mommy loves you just the same, even though you're naughty," easily provides the conditioning model. (165)

Note that the frequency of punishment is not essential. In their empirical study of sadomasochists Gosselin and Wilson found no significant correlation between frequency of punishment and resulting sadomasochistic tendencies; and, although one suspects that some such correlation really does exist, one would not expect this correlation to be very strong. The three-stage sequence -- initial chance association of painful punishment with pleasurable arousal, followed by a phase of "jumping to conclusions," followed by self-fulfilling prophecy -- is dependent on many factors other than frequency of punishment.

This cognitive theory explains the origins of sadomasochistic tendencies; what about the maintenance of these tendencies?

As the child grows up, he may receive messages ... to the effect that the usual target for his genital feelings (i.e. the female **per se**, and particularly her vagina) is forbidden, or naughty or wicked, or dirty, or unmentionable or in some other way not to be approached. Now, because he is more easily conditioned than most, he takes these messages seriously and believes in them more easily; because of his higher emotionality, his attempts to disobey the messages lead to overpowerful anxiety and guilt associated with his arousal. In seeking to obtain sexual pleasure when aroused, he may therefore remember that the fetish fabric [or the pain, in the case of sadomasochists] gave pleasure without interaction with the female. (165-166)

This argument is sensible, and as they point out, it transfers over naturally to the case of sadomasochism.

The Freudian and cognitive theories of masochism give very different explanations for thesame phenomenon. However, as we shall see, they fit together quite nicely in the context of the system-theoretic model.

Male and Female Subselves

The central idea of the theory of masochism to be presented here is the concept of **male and female subselves** -- in other words, the idea that every one of us, in our early youth, develops two subself networks, one reflecting a generally female point of view, and the other reflecting a

generally male point of view. In some individuals these subselves may coordinate so closely with one another that they function almost as a single unified subpersonality network. In others they may remain at odds with each other, each one taking control of the body and personality-independent memory at different times. Masochism is to be understood as a disturbance of the dynamics between male and female subselves. Freud, in the passages quoted above, hinted at the concept of male and female subselves, but did not articulate the idea very clearly; here the hint will become a central focus.

It may be questioned whether these subselves themselves can split up into further subsubselves; such an occurence seems possible but rather unlikely. It seems quite plausible that many individuals should contain **more than one** subself of each gender, but much less likely that, say, a male subself should contain well-differentiated male and female components. In theory there could be an hierarchy of males within females, females within males, females within females, males within males, males within females within females, females within males within males, and so forth. But in practice, it seems probable that the distinctions become blurred after one or two levels.

For simplicity, from here on I will generally speak as if there were just one male subself, and just one female subself. But this is an idealization which is unlikely to be fulfilled in practice. In any particular case things will be more complicated, and the different subselves may "team up" to work for and against each other in various subtle ways. Furthermore, for purposes of illustration, I will exaggerate somewhat the difference between the male and female subselves. In reality, of course, male and female thought patterns have far more commonalities than differences. But it is the differences which are most relevant to the phenomenon of masochism.

It is easy to see how male and female subselves would arise. First of all, a child of either sex will generally **imitate both parents**. In some situations the child will try to behave as the mother does, in others the child will try to behave as the father does. Thus, among the huge collection of situation-dependent **behavioral schema** which the child's mind builds up, two large categories will emerge: the category of mother-like behavior schema, and the category of father- like behavior schema. Each of these categories will naturally form into an self-organizing system, which is an approximate reconstitution of the personality system of the appropriate parent. For the collection of behavior schemas derived from, say, the mother, will tend to "hang together"; they will reinforce each other, first of all by setting up conditions in the mind that are favorable for one another's survival, and secondly by setting up conditions in the external world that are favorable for one another's survival.

Each of these self-organizing systems is, in Epstein's terms, a "self and reality theory." According to the logic of the dual network, each one provides a different way of constructing coherent wholes out of the fragmentary features provided by the sense organs and lower level perceptual systems. Thus each subself literally sees a different world. More concretely, to take a very common case, the male subself may perceive a world consisting of **factors to be controlled**, whereas the female subself may perceive a world consisting largely of **factors which control**. This is not merely a difference in attitude, it is a difference in perception and cognition.

The type of masochist for whom the word **masochist** was originally invented was whatFreud called the "feminine masochist." This is the individual who derives sexual pleasure from having others physically abuse them. In most cases it is a male who enjoys being whipped or spanked, verbally humiliated, and otherwise tormented by an attractive female. Freud rightly points out the fundamentally **female** character of the feminine masochist. But how does this insight translate into the subself theory? The simplest interpretation is that it is the female subself of the feminine masochist which wants to be tormented. But more careful consideration suggests a subtler interpretation: that **the female subself wants the male subself to be tormented**. Masochism is the spilling-over into the external world of an internal conflict. The dominatrix is a tool used by the female subself of the male masochist in order to temporarily subjugate the masochist's male subself.

This hypothesis is somewhat reminiscent of the well-known "sex identity conflict hypothesis." As Beatrice Whiting (1965) observes,

[A]n individual identifies with that person who seems most important to him, the person who is perceived as controlling those resources that he wants.

Thhus, for most boys, the primary identification will be

with mother and "cross-sex identification" will occur, which causes the boy to be

thrown into conflict. He will develop a strong need to reject his underlying female identity. This may lead to an overdetermined attempt to prove his masculinity, manifested by a preoccupation with physical strength and athletic prowess, or attempts to demonstrate daring and valor, or behavior that is violent and aggressive.

Wife-beating, according to Whiting, is one possible consequence of this kind of conflict. Masochism, or so I argue, is a related phenomenon: the difference is that, in the masochist, the female subself recognizes the undesirable tendencies of the male subself, and acts to beat it down.

The hypothesis which I am making here is also related to Deleuze's (1971) idea that the masochist, by being beaten by a woman, wants to destroy the power of his internalized father. But the difference is that, while Deleuze situates the female power **outside** the masochist, in the dominatrix, I situate the primary source of female power **within** the masochist. The masochist's female subself wants to **conspire** with the dominatrix to beat down the masochist's male subself. Thus, the subtext of masochism is a struggle for internal control between male and female subselves.

Psychoanalytically, one suspects that this struggle may often be a reflection of the mother's repressed desire to beat down the father. The female subself internalizes the mother, the male subself internalizes the father, and in this process the mother's hostility toward the father may also be internalized. It would be useful in this regard to study the relationships of the parents of masochistic individuals, and observe whether there is indeed an unusual degree of repressed hostility of the mother against the father.

Or, taking a cue from Baudrillard (1988) instead of Freud, one may say that what happens in masochism is that the prototypical female dynamic of **seduction** breaks down. The female subself fails to derive sufficient gratification from the exertion of seductive, implicit control. Thus she seeks an external agent, not so much to control the specific behaviors of the male subself, but to subject the male subself to patterns of being controlled. By subjecting the male subself to patterns of being controlled, which differ so greatly from his ordinary patterns of controlling, she hopes to weaken the male subself's controlling habits, and to to thus render the male subself more amenable to **her** control.

The Role of Cognitive Factors

But there is a missing link here: this strategy on the part of the female subself would never work if there were not some "seed" in a person's history to start it off. This is where the cognitive theories of Gosselin and Wilson come in: they explain how the basic connection between physical violence and sexual excitation could originate and develop. Their cognitive arguments are made extremely plausible by the singular appropriateness of **punishment** as a source for the arousal/violence connection. For corporal punishment is a situation of total lack of control, which usually occurs in response to an excessively controlling attitude. The child decides that **he** is the boss, that he can do whatever he wants to do, and the parental response is that he is emphatically **not** the boss, that he in fact has no control whatsoever, that he can do nothing but lie still and be hurt.

Thus, one may conjecture the following series of events. The female subself finds itself in a losing position in its war with the male subself, so it turns to the only thing in its experience which has stopped the male subself from being overly controlling: physical discipline. But the male subself has no incentive for seeking physical discipline ... except for the connection between sexual arousal and pain and humiliation, established according to the mechanism described by Gosselin and Wilson. This combination of factors provides a much more plausible story of the **maintenance** of masochistic tendencies. They are maintained as a consequence of the **self- organizing** nature of the small self-organizing system consisting of the male and female subselves. The female subself encourages the male subself to pursue the connection between arousal and pain; the male subself obliges due to its own internal dynamics; and by succumbing to these masochistic pressures, the male subself accedes power to the female subself, thus giving the female subself incentive to give further encouragement for masochistic behavior. This is a feedback loop occurring within an self-organizing system; it is **exactly** the sort of explanation that one would expect a systems theory of personality to provide. In complex systems terms, it is an **attractor** of the dynamical system posed by the male and female subselves and their interactions.

The End Result

So, finally, what is the end result of all this subself dynamics? On the one hand, the female subself can **win**, leaving the male subself completely cowed and "emasculated." This results in an ineffective and disheartened male; for, clearly, it is impossible for a male human being to survive in a mentally healthy way without either a male subself or an integrated male/female subself. On the other hand, the **male** subself can win, abandoning its masochistic tendencies and

shutting off the female subself altogether. This is also not a happy conclusion, since the female subself is also necessary for healthy functioning.

Finally, and most optimistically, it is possible that the male subself will be weakened just enough for the female subself to force it into a relationship of mutual support. This is not to say that all subself competition need end, but only that the male and female subselves should come to form an self-organizing system in which both have roughly equal shares of power (as measured by, say, the degree to which the overall responses of the system depend on changes in the responses of each individual subsystem). One might think that, once such a balance were achieved, the masochistic behavior would no longer be necessary. But it is also possible that the subselves could become entrained in a cycle: a period of roughly equal power, followed by a period of gradual ascendance of the male subself, followed by a period of masochism, followed by another period of roughly equal power, and so forth.

Clearly there are many possibilities; as Freud surmised in 1905, there is a great deal of complexity involved here. One way to explore this complexity would be to formulate thedynamical relationship between the two subselves as a mathematical dynamical system, in the manner of (Abraham, Abraham and Shaw, 1993) or (Goertzel and Goertzel, 1994). Admittedly, one might run into difficulties in setting the various parameters involved in such equations. But nevertheless, the experiment would seem to be well worth doing.

The masochistic personality as described by Freud, Reich, Fromm and others is the result of the first outcome listed above: the victory of the female subself. But this is not necessarily the most common outcome of masochistic subself dynamics. It must be remembered that those masochists who seek psychotherapy are generally among those who are most troubled by their masochism. Many of the masochistic individuals interviewed by Gosselin and Wilson seem to lead happy and productive lives. According to the present theory, the explanation for this would be that their male and female subselves have reached a balanced dynamical condition which can only be sustained by periodic masochistic behavior.

Conclusion

The theory of masochism which I have given here is really only a sketch of a theory. For one thing, it speaks throughout of the competition between a single male subself and a single female subself, when in reality there will usually be a **number** of male subselves and a **number** of female subselves, as well as some subselves that are not clearly identifiable as either male or female, and some semi-autonomous mental subnetworks that are not quite independent enough to be categorized as "subselves." The dynamic described above is extremely general, and is in principle extensible to these more realistic situations. But clearly, more theoretical work in this direction is required. And this will be difficult theoretical work, because it will require the introduction of numerous cognitive and emotional issues besides the male/female dichotomy that has been our almost exclusive focus here.

And in addition to these theoretical questions, the problem of formulating rigorous **tests** of the theory remains uninvestigated. The theory of subself dynamics is certainly not immune to empirical test. As opposed to, say, the theories of Freud or Fromm or Reich, it would seem to be

a very testable theory indeed, as it is grounded in the psynet model, which rests on some very concrete ideas about cognitive and neural processes. But a great deal of ingenuity will be required in order to work out actual experiments which probe the subtleties of subself dynamics. One suspects that the experiential sampling method (Czentomihalyi, 1992) might be useful in tracking the passage of various subselves in and out of control.

Finally, no attention has been yet been paid to the **clinical** consequences of the theory. The author is not a clinical psychologist and has little expertise in such matters. However, given this warning, it must be said that the theory of subself dynamics would seem to have obvious therapeutic applications. Specifically, the theory of masochism outlined here suggests that, if the male masochist is to be cured, he must be cured by somehow **strengthening** the masochist's female subself, until it becomes strong enough that it no longer has the need to subject the male subself to humiliation and violence. The task of the clinician would then be to develop strategies for carrying out this strengthening.

CHAPTER FOURTEEN

ON THE DYNAMICS OF CREATIVITY

14.1 INTRODUCTION

Creativity is the great mystery at the center of Western culture. We preach order, science, logic and reason. But none of the great accomplishments of science, logic and reason was actually achieved in a scientific, logical, reasonable manner. Every single one must, instead, be attributed to the strange, obscure and definitively irrational process of creative inspiration. Logic and reason are indispensible in the working out ideas, once they have arisen -- but the actual **conception** of bold, original ideas is something else entirely.

No creative person completely understands what they do when they create. And no two individuals' incomplete accounts of creative process would be the same. But nevertheless, there are some common patterns spanning different people's creativity; and there is thus some basis for theory.

In previous chapters, the phenomenon of creativity has lurked around the edges of the discussion. Here I will confront it head-on. Drawing on the ideas of most of the previous chapters, I will frame a comprehensive complexity-theoretic answer to the question: How do those most exquisitely complex systems, minds, go about creating forms?

I will begin on the whole-mind, personality level, with the idea that certain individuals possess creatively-inspired, largely medium-dependent "creative subselves." In conjunction with the

Fundamental Principle of Personality Dynamics, this idea in itself gives new insight into the much-discussed relationship between inspired creativity and madness. A healthy creative person, it is argued, maintains I-You relationships between their creative subselves and their everyday subselves. In the mind of a "mad" creative person, on the other hand, the relationship is strained and competitive, in the I-It mold.

The question of the **internal workings** of the creative subself is then addressed. Different complex systems models are viewed as capturing different **aspects** of the creative process.

First, the analogy between creative thought and the genetic algorithm is pursued. It is argued that the creative process involves two main aspects: combination and mutation of ideas, in the spirit of the genetic algorithm; and analogical spreading of ideas, following the lines of the dynamically self-organizing associative memory network. The dual network model explains the interconnection of these two processes. While these processes are present throughout the mind, creative subselves provide an environment in which they are allowed to act with unusual liberty and flexibility.

This flexibility is related to the action of the perceptual-cognitive loop, which, when "coherentizing" thought-systems within the creative subself, seems to have a particularly gentle hand, creating systems that can relatively easily be dissected and put back together in new ways. Other subselves create their own realities having to do with physical sense-perceptions and actions; creative subselves, on the other hand, create their own realities having to do with abstract forms and structures. Because the creative subself deals with a more flexible "environment," with a more amenable fitness landscape, it can afford to be more flexible internally.

In dynamical systems terms, the process of creative thought may be viewed as the simultaneous creation and exploration of autopoietic attractors. Ideas are explored, and allowed to lead to other ideas, in trajectories that evolve in parallel. Eventually this dynamic process leads to a kind of rough "convergence" on a strange attractor -- a basic sense for what kind of idea, what kind of product one is going to have. The various parts of this attractor are then explored in a basically chaotic way, until a particular **part** of the attractor is converged to. In formal language terms, we may express this by saingy that the act of creative inspiration **creates its own languages**, which it then narrows down into simpler and simpler languages, until it arrives at languages that the rest of the mind can understand.

The hierarchical structure of the dual network plays a role here, in that attractors formed on higher levels progressively give rise to attractors dealing with lower levels. One thus has a kind of iterative substitution, similar to the L-system model of sentence production. Instead of sentences consisting of words, however, one has "sentences" (abstract syntactic constructions) consisting of much more abstract structures. The lower levels use their evolutionary dynamics to produce elements that yield the higher-level created structures as **emergent patterns**.

An analogy between the structure of creative ideas and the structure of **dreams** is made. Just as dreams provide autopoietic thought systems with what they need, so, it is argued, do creative inspirations. Creative inspiration deals with thought systems whose needs are too complex for

dreams to figure out how to solve. Creative activity is, in part, a very refined way of disempowering excessively persistent autopoietic thought sys tems.

In this sense the creative state of consciousness is structurally and functionally similar to the dream state of consciousness. There are also other similarities between the two states. For instance, in both states, the perceptual corner of the perceptual-cognitive-active loop is replaced with a reference to **memory**, while the "inner eye" is relieved of its duty of ordinary detached reflection. In dreaming, however, the inner eye often has no role whatsoever, or a very nebulous role; while in creative inspiration it assumes an alien or "godlike" mantle.

These observations do not exhaust the richness of human creativity, but they do constitute a far more detailed and comprehensive theory of creativity than has ever been givenbefore. They tie together the actual experience of creativity with the dynamics of existing computational algorithms. And they show us exactly what is missing in the supposedly creative computer programs that we have today.

14.2 THE EXPERIENCE OF INSPIRATION

Perhaps the most phenomenologically accurate theory of creativity is the one which holds that the creative individual has a direct line to God. God gives them the shapes for the painting, the words for the poem, the equations for the theory, the notes for the symphony. Where else could such remarkable, beautiful things come from? Surely not from the mere mind of man!

Many creative people have experienced forms and ideas pouring out as if from some unknown inner source. Forms streaming, emanating, exploding -- so much faster and more elegantly fit together than if they were consciously reasoned out. The most striking description of this experience I have seen was given by Nietzsche in his autobiography, *Ecce Homo*:

> Has anyone at the end of the eighteenth century a clear idea of what poets of strong ages have called **inspiration**? If not, I will describe it. -- If one has the slightest residue of superstition left in one's system, one could hardly resist altogether the idea that one is merely incarnation, merely mouthpiece, merely a medium of overpowering forces. The concept of revelation -- in the sense that suddenly, with indescribable certainty and subtlety, something becomes **visible**, audible, something that shakes one to the last depths and throws one down -- that merely describes the facts. One hears, one does not seek; one accepts, one does not ask who gives; like lightning, a thought flashes up, with necessity, without hesitation regarding its form -- I never had any choice.
>
> A rapture whose tremendous tension occasionally discharges itself in a flood of tears -- now the pace quickens involuntarily, now it becomes slow; one is altogether beside onself, with the distinct consciousness of subtle shudders and one's skin creeping down to one's toes; a depth of happiness in which evern what is painful and gloomy does not seem something opposite but rather conditioned, provoked, a **necessary** color in such a superabundance of light....
>
> Everything happens involuntarily in the highest degree but as in a gale a feeling of freedom, of absoluteness, of power, of divinity. -- The involuntariness of image and metaphor is strangest

of all; one no longer has any notion of what is an image or metaphor: everything offers itself as the nearest, most obvious, simplest expression....

Nietzsche's experience of inspiration was particularly overpowering. But it differs in intensity, rather than in kind, from the "everyday" inspiration experienced by ordinary creativepeople.

Sometimes this "overpowering force" of which one is merely a "medium" is experienced as an actual alien entity. In these relatively rare cases, creative inspiration blurs into paranoid hallucination -- or religious inspiration. The great science- fiction writer Philip K. Dick was an example of this. He felt himself being contacted by an artificial intelligence from another star system. This AI being fed ideas into his mind, giving him the plots for his last few novels, especially *The Divine Invasion* and *Valis*. Dick's vision is too complex to discuss here in detail; so here I will merely quote one entry from his journal:

March 20, 1974: It seized me entirely, lifting me from the limitations of the space-time matrix; it mastered me as, at the same instant, I knew that the world around me was cardboard, a fake. Through its power I saw suddenly the universe as it was; through its power and perception I saw what really existed, and through its power of no- thought decision, I acted to **free myself**....

Dick had always written using a "downhill skiing" methodology. He would sit in front of the typewriter for hours and hours on end, sometimes on uppers, and write without revisions. This strategy was necessitated by the financial realities of the science fiction market. At $1000 or so per novel, it didn't pay to revise too heavily. One had to make a living! Inevitably, some of his novels were inconsistent and sloppy. A few were just plain lousy. But the best of his work was remarkably inspired and elegant. His stories rarely fit together logically, but they cohered conceptually, psychologically. They had the surreal consistency of dreams and hallucinations. In his case, the downhill skiing method forced his unconscious to take over the writing task, inducing a "mediumistic" state similar to that described by Nietzsche. Since his writing depended on this mediumistic state of mind, it was ideally suited for invasion by the "alien force" that he experienced. He wrote without too much reasoned, conscious intervention anyway. What difference did it make where the inspiration came from -- from within himself, or from beyond the solar system?

Another classic example of creative inspiration is the poet Arthur Rimbaud. Rimbaud viewed the creative person as drawing their inspiration from another world, an unknown source beyond the ordinary cosmos:

One must, I say, be a **visionary**, make oneself a **visionary**.

The poet makes himself a visionary through a long, a prodigous and rational disordering of **all** the senses. Every form of love, of suffering, of madness; he searches himself, he consumes all the poisons in him, keeping only their quintessences. Ineffable torture in which he will need all his faith and superhuman strength, the great criminal, the great sickman, the accursed, -- and thesupreme Savant! For he arrives at the unknown! Since he has cultivated his soul -- richer to begin with than any other! He arrives at the unknown: and even if, half- crazed, in the end, he loses the understanding of his visions, he has seen them! Let him be destroyed in his leap by

those unnameable, unutterable and innumerable things: there will come other horrible workers: they will begin at the horizons where he has succumbed.

So then, the poet is truly a thief of fire.

Humanity is his responsibility, even the animals; he must see to it that his inventions can be smelled, felt, heard. If what he brings back from beyond has form, he gives it form, if it is formless, he gives it formlessness. A language must be found....

This eternal art will have its functions since poets are citizens. Poetry will no longer accompany action but will lead it.

These poets are going to exist!

The Promethean thief of fire is a slightly different image from the "voice of God within." Instead of drawing on an internal source, the artist is traveling somewhere new, and returning with untold, perhaps untellable treasures.

Similar experiences have been reported by many scientists. The chemist Kekule' conceived the idea for the benzene ring while in a state of dreamlike reverie. He perceived the benzene molecule as a snake, wriggling around, trying vainly to assume a stable structure. Then, all of a sudden, the snake bit its own tail. Everything was clarified! The molecule was a circle! Calculations revealed that this conjecture was correct. This was a unique molecular structure for its time, in several different ways; it was a landmark in organic chemistry. It is also a beautiful illustration of the role of cultural archetypes in guiding creativity. The snake biting its own tail is a stereotypical image, going back before Western culture to Eastern mythology. Kekule's creative mind spontaneously locked the benzene molecule in with an archetypal image, the snake, and then the cultural repertoire of image transformations did his work for him.

An impressive survey of creative inspiration was given by the mathematician Jacques Hadamard in his book *The Psychology of Mathematical Invention*. Hadamard reviews, for instance, the great mathematician Poincare', who conceived the idea for a mathematical structure called Fuschian functions while stepping onto a bus. All of a sudden, the whole structure popped into his head, in complete and elegant form. He had struggled with the problem for some time, but had made little progress, and had shut it out of his conscious mind. But something had been obviously working on it. Had his unconscious posted a query to God, and finally received the answer? Or had some component of his mind simply continued working?

The great chemist Linus Pauling described this sort of process in a particularly clear way:

Some years ago I decided that I had been making use of myunconscious in a well-defined way. In attacking a difficult new problem I might work for several days at my desk, making appropriate calculations and trying to find a solution to the problem. I developed the habit of thinking about a problem as I lay in bed, waiting to go to sleep. I might think about the same problem for several nights in succession, while I was reading or making calculations about it during the day. Then I would stop working on the problem, and, after a while, stop thinking

about it in the period before going to sleep. Some weeks or months, or even years might go by, and then, suddenly, an idea that represented a solution to the problem would burst into my consciousness.

Pauling won the Nobel Prize in chemistry for his theory of the chemical bond; he also did more than anyone else to found the science of molecular biology, and made important contributions to other areas, such as mineral chemistry and metal chemistry. He described one of his key creative processes as the "stochastic method" -- the free-form combination of facts into novel configurations, leading to original hypotheses. In applying his stochastic method, he relied on his tremendous store of factual information and his outstanding physical and chemical intuition. He also relied on the altered state of consciousness experienced when falling asleep, a state of consciousness in which ideas blend into each other more easily than usual, and on the mysterious, long-term creative processes of the unconscious.

In physical chemistry proper, Pauling's stochastic guesses displayed an uncanny accuracy. The combinations of facts arrived at by his unconscious were nearly always the same ones present in the physical world! In chemical biology and medicine, his guesses were less accurate, though he still made a number of important discoveries. His creative process apparently did not change from one research area to the other, but the average quality of the results did. It appears that Pauling's ability to see new connections was so powerful, and at the same time so stochastic, that it needed a very solid body of factual knowledge to tie it down. This body of knowledge was there in physical chemistry, but less so in molecular biology, and far less so in medicine. (I have given a more detailed discussion of Pauling's scientific thought in the middle chapters of the biography *Linus Pauling: A Life in Science and Politics* (Ted and Ben Goertzel, Basic Books, 1995)):

14.2 THE CREATIVE SUBSELF

To make sense of the experience of creative inspiration, within the framework of the psynet model and complexity science, is no easy task. The first step on the path to such an understanding, is, I believe, the application of **subself dynamics**. In this section I will postulate the existence, in hgihly creative individuals, of a **creative subself**, whose sole purpose for existence is the creative construction of forms.

It must be emphasized that the postulation of creative subselves is not intended to explain away the experience of "divine inspiration." The creative subself is not the "alien force" which some creators have experienced as giving them their ideas. If one were to pursue this metaphor, the creative subself would more accurately be thought of as the part of the mind that is in touch with the "alien force." This point will be returned to later.

A creative subself has an unusual "shape" -- it interfaces significantly with only a very limited range of lower-level perceptual/motor processes. For instance, a creative subself focused on musical composition and performance would have next to no ability to control processes concerned with walking, speaking, lovemaking, etc. On the other hand, in the domain in which it does act, a creative subself has an unusually high degree of autocratic control. For the creative act to proceed successfully, the creative subself must be allowed to act in a basically

unmonitored way, i.e., without continual interference from other, more broadly reality-based subselves.

A creative subself makes speculative, wide-ranging and inventive use of the associative memory network. It knows how to obey constraints, but it also knows how to let forms flow into one another freely. It does a minimum of "judgement." Its business is the generation of novel and elegant forms. The degree of "looseness" involved in this process would be entirely inappropriate in many contexts -- e.g. in a social setting, while walking or driving, etc. But in the midst of the creative act, in the midst of interaction with the artistic medium, "anything goes." The limited scope of the perceptual-motor interface of a creative subself is essential.

It might seem an exaggeration to call the collection of procedures used by a creative individual a "subself." In some cases, perhaps this **is** an exaggeration. But in the case of the most strikingly creative individuals, I think it is quite accurate. In many cases there is a surprising difference between the everyday personality of an individual, and their "creative personality." This is why so many people are surprised when they read their friends' books. The reaction is: "Good God! This is **you**? This is what's going on in **your** head? Why don't you ever tell any of this stuff to me?" The answer is that there are two different "me"'s involved. The book is written by a different subself, a different autopoietic system of patterns, than the subself who carries out conversations and goes about in the world. There is a relation between the everyday subself that the friend knows, and the creative subself that wrote the book -- but not as much relation as our unified vision of personality leads us to expect.

A striking case of this kind of dissociation was Friedrich Nietzsche, who in real life was mild-mannered and friendly to everyone -- but in his books was unsparing and ruthless, tearing his competitors to pieces and, in his own phrase, "philosophizing with a hammer." In his books he called Christianity a terrible curse and insulted Christians in the worst possible terms. In his life he was not only cordial but kind to many Christians, including his own sister and mother. Similarly, in his books heinsulted the modern German race in the worst possible way -- considering them the apex of egotistical, power-hungry, over- emotional stupidity. But in his ordinary life he got on fine with his fellow Germans.

Nietzsche's work was later taken up by many power-hungry and unpleasant individuals, including Adolf Hitler. Hitler's real- world personality matched Nietzsche's creative self, far better than Nietzsche's real-world self. He rode roughshod over people, shaping them to his will, just as Nietzsche did with abstract concepts. The difference is that, in the world of ideas, free-flowing spontaneous ruthlessness can be immensely productive, whereas in the world of human beings, this kind of ruthlessness leads only to destruction. A thought process that destroys old ideas, breaking them to bits and building them into new ones, can be quite a fine thing. Doing this on the level of human beings and social structures leads to tragic, and sometimes (as in the case of Hitler) inexpressibly horrible results.

Origins of the Creative Subself

Not everyone has a creative subself. On the other hand, some rare individuals may have several. The conditions for the emergence of such a subself are not clear. A few insights are

provided by *Cradles of Eminence*, a book in which my grandparents, Victor and Mildred Goertzel, studied the childhoods of eminent individuals (together with my father, Ted Goertzel, they also wrote a sequel, *Three Hundred Eminent Personalities*). The population of eminent people is obviously different from the population of highly creative people -- most highly creative individuals never become eminent, and some eminent people are not in the least bit creative. But, on the whole, most eminent individuals studied in these books were indeed highly creative, and thus the studies do have something to offer.

The clearest moral from these studies is that an individual must be stimulated in early youth. Someone -- a parent, grandparent, uncle, family friend, etc. -- must encourage the person to develop in the direction of art, science, music, or whatever the creative vocation is to be.

Beyond this, the morals different from one domain to the other. Nearly all writers and visual artists had childhoods that were troubled in one way or another -- a fact which made it awkward for my grandparents when lecturing on their book. When hopeful parents asked them what they should do to make their children become famous artists, they would have to carefully rephrase the most direct and honest response: "Don't be too nice to them! Emotionally abuse them!" On the other hand, a great number of scientists experienced some long period of solitude in youth or adolescence. Often this was a period of illness, in which the person had nothing to do but lie around reading and thinking. Clearly, the common factor among artists and scientists is **withdrawal from the world**. Emotional problems in childhood, like long periods of illness, cause one to withdraw into oneself, to separate oneself from the social domain.

The conclusion would thus be that, in order to become anhighly creative person, it is necessary to:

1) develop a habit of carrying out some creative activity

2) have an emotional or situational need to withdraw from

the world into "one's own universe"

Clearly, however, these two conditions are not sufficient. They merely pave the way. What one must do, given these two factors, is to instinctively make the creative activity one's one universe. The creative activity thus takes the place of the ordinary world, in which other people move. Just as different subselves normally develop to deal with different situations, a subself develops to deal with this situation, this particular subset of the world -- which happens to consist of interaction with an artistic medium.

When this creative subself attains a certain level of coherence and autonomy, it gains a "life of its own" and begins to grow and develop like any other subself. It cannot flourish without access to the medium that is its world; thus it urges the other subselves to pursue the creative vocation that supports it. In some circumstances, the creative subself may be the only redeeming aspect of an otherwise execrable individual. In other cases, however, one might rightly view the creative subself as a kind of psychological parasite on an otherwise healthy organism. The creative vocation in question may have no value to the person in question; the passion to pursue this

vocation may in fact destroy the person's life. The other subselves may not understand why they are unable to hold down a job, stay in a relationship, etc., when the answer is that the creative subself has gained a great deal of power, and is willing to sacrifice everything for steady access to what, in its view, is the only "real" part of the world.

14.4 CREATIVE SUBSELF DYNAMICS

The hypothesis of a creative subself, on its own, does not explain creativity. One must understand the inner workings of the creative subself -- this is where complex systems models are useful. One must explain how the creative subself gives rise to the subjective feeling of an "external force." One must look at the interactions between the creative subself and the other subselves in the mind. And, finally, one must look at the relation between the creative subselves and other subselves in **other minds** as well.

The relation between the ordinary subselves and the creative subself is a complex one. In some cases the creative subself may work on its own, with little connection to the other subselves, and then, after a certain period of time, present a fully formed "answer" or artistic work. In other cases it may work on its own but give the other subselves continual news flashes on its work. In yet other situations it may be so well integrated with ordinary subselves that it is barely recognized as a separate entity at all. The striking thing, however, is that in nearly all cases of really extreme creativity, the degree ofseparateness is quite high. This is the feeling "something else is creating this" noted above. It is important enough that it deserves the status of an abstract principle:

First Principle of Creative Subself Dynamics: Thorough integration of the creative subself into other subselves seems to contradict extremely productive and innovative creativity.

One of the roles of other subselves in the creative process is to provide **judgement**. The creative subself is generally skilled at producing rather than evaluating. Other subselves must play the role of the external world, of the critic and commentator. Too much criticism is not productive, as it inhibits the creative subself -- which, in order to be innovative, must be allowed its essential spontaneity. On the other hand, too little criticism is not productive either, as the unprocessed output of the creative subself can be almost random at times. The thorough integration of critical and creative faculties does not seem to work. It is essential that the two systems be autonomous, yet interacting.

Critical subselves may intervene on the level of overall works. After the creative subself creates a painting, poem, or mathematical structure, other subselves come in and help it to decide on the value of the thing created. The poem may be burned; the painting may be slashed; the theorem may remain unpublished. The creative subself is usually enthused about whatever it has created: the creative work is, after all, its universe. The other subselves usually take a more restrained, rational attitude. The overall attitude of the creative individual is a fluctuating combination of these two component attitudes.

The critical process may also intervene on a much shorter time-scale, however. It may intervene, say, every paragraph or so in a novelist's writing process; or every few minutes in a mathematician's theorem-proving process. One mathematician colleague has told me that he thinks free-flowingly and creatively, scribbling down calculations on pieces of paper, but stops every fifteen minutes or so to ask himself: "Wait a moment. What am I actually accomplishing here?" Only in rare instances of extreme inspiration does he think creatively for long periods without continual self-examination. In cases such as this, what one has is an approximately periodic attractor, involving oscillation between a creative subself and an ordinary, critical subself. The creative process is dependent on this interaction: without the give-and-take, the creative subself would "go overboard" and fail to produce interesting, relevant structures. This line of thinking leads one to the following hypothesis:

Second Principle of Creative Subself Dynamics: While all creativity involves some oscillation between creative and ordinary subselves, the **frequency** of the oscillation should correspond roughly to the amount of **constraint** involved in the creative medium.

In very unconstrained media such as modern poetry or abstract expressionist painting or jazz improvisation, the critic can come in only at the end, when it is necessary to decide on the value of the work as a whole. In very constrained media such as mathematics or portraiture, the critic must play an integral role, and become an integral part of the inventive process itself.

An example of what happens when the frequency is too low for the medium is Linus Pauling's triple helix model of DNA, proposed just weeks before the correct double helix model was found by Watson and Crick. Had Pauling discovered the double helix, he probably would have won a third Nobel Prize, in Medicine (his second was in Peace, honoring his role in urging the governments of the world toward a nuclear test ban). Part of the reason for his failure to construct the correct model was political: because of his anti-nuclear-testing activities, he was prevented by the U.S. government from visiting England at a key moment, when he would have been able to see pertinent laboratory results obtained there by Rosalind Franklin. But Pauling, later on, stated that he felt he could have come up the the double helix model anyway. In fact, he claimed that the double helix structure had occurred to him in the past; it had just slipped his mind at the crucial moment, due to a lack of attention. One can read this as "sour grapes" -- or one can understand it as a natural consequence of the dynamics of creativity. Pauling let the ideas fall where they might. The triple helix model emerged from his unconscious; he let it out, and wrote it up. Because of his chaotic personal circumstances at the time, the critical attention his other subselves would have devoted to the problem was skimped on. Had the oscillation of his creative process proceeded at its proper frequency, he might well have discovered the double helix model. His ordinary subselves would have discovered the intrinsic inadequacies of the triple helix model, and sent a message to the creative subself to "try again."

In evolutionary biology terms, the frequency of critical intervention may be understood as one component of the **harshness** of the environment facing the creative subself. In biology one may show that the maximum evolutionary innovation occurs in a moderately but not excessively harsh environment (I have argued this point in *The Evolving Mind*). The same result holds for creativity. A totally friendly, criticism-free environment, places no constraints on creative output and is unlikely to ever lead to impressive creative works. Some constraint is necessary, or else all

one has is a pseudo-random exploration of the space of products of some collection of ideas. On the other hand, too much constraint, an overly harsh environment, is also unproductive, because it penalizes experimentation too severely. New ideas must be given some leeway, some extra time in order to bring themselves up to snuff. The exact degree of harshness which is optimal for a given situation is never the minimum or maximum, but rather lies somewhere inbetween, at a point which depends upon the amount of constraint inherent in the particular artistic medium.

The average frequency of the creative/critical oscillation is determined by the amount of constraint involves in the medium,but even so, the correct frequency at any given time can only be determined on a situation-specific basis. Each subself has to know when to let the other one do its thing. This requires an intricate interdependence between the two subselves. Ultimately, the prediction is that the most consistently productive creativity will come out of individuals whose creative and critical subselves display an I-You relationship. This is our third principle of creativity:

Third Principle of Creative Subself Dynamics: Creativity requires continual adaptation of the creative/critical oscillation frequency, which can be most effectively achieved, in the long run, by I-You interactions between creative and ordinary (critical) subselves.

The relation between creative and critical subselves is, like the relation between lovers, a rather difficult one. Periods of I- You interaction may be followed by periods of hostility or indifference. One may even find a kind of abstract masochism, wherein the ordinary subselves take pleasure from suffering caused them by the creative subself.

14.5 DIVINE INSPIRATION AND EMERGENT PATTERN

Now, having firmly established the concept of the "creative subself," we are ready to return to the apparently alien or "divine" nature of creative inspiration. I have noted the way creative ideas often seem to pop into the mind from elsewhere; and I have argued that this "elsewhere" has something to do with a special subself within the mind, called the creative subself. However, as clearly stated above, my claim is not that the creative subself **is** this external force. Instead, my claim is that, whatever this externally experienced "creative force" is -- the creative subself is the receiver.

From the modern scientific perspective, the concept of "divine inspiration" is preposterous. On the other hand, many pre-scientific, wisdom-oriented world views have taken a different attitude. From an holistic, non-scientific perspective, what is preposterous is the idea that the inspirations which pop into one's head from outside are just mechanical productions of brain dynamics. Clearly, each point of view has its own validity -- one empirical, one experiential. Any theory of creativity that wants to be truly comprehensive must take both into account.

The Vedantic Hierarchy

Nearly all wisdom traditions speak of an hierarchical order in the universe, of different planes of being, ascending from the lowest material plane to the highest plane of ultimate being. Many

different hierarchical schemes have been presented, but all are getting at the same basic idea. Here I will work with the hierarchy as presented in the Vedantic school of Indianphilosophy.

It must be emphasized that, while the present book is a scientific one, the philosophical notions embodied in the Vedantic hierarchy were never intended in a scientific sense. Rather they reflect the introspective insights of many generations of acutely self-aware individuals. As such, they provide valuable information which may serve to guide scientific theory-construction.

A very small amount of background may be useful. Vedanta is one of the six major schools of classic Indian philosophy: Purva Mimamsa, Vedanta, Sankhya, Yoga, Nyaya and Vaisesika. The ultimate origin of all these schools is the Vedas, a corpus of hymns composed between 4000 and 6000 B.C. In true mythic fashion, the Vedic hymns are addressed to the deities ruling the various forces of nature. But there is also a more metaphysical side to their teachings, in that the gods are understood as expressions of a deeper, impersonal order of being, which is in turn understood as an expression of a yet deeper realm of pure **formlessness**. Vedanta is a large and comples system, but here we will content ourselves with a single, central part of Vedantic philosophy: the doctrine of the five sheaths or *koshas*. These sheaths are to be understood as covers, perhaps as films which coat ultimate being (*Atman*, the personal, individual manifestation of *Brahman*). The lower levels are denser coats; the higher levels are more transparent and let more of *Atman* through.

The great mystic Sri Aurobindo (1972) explained and re- interpreted the Vedantic *koshas* in a way which is particularly relevant here. Sri Aurobindo takes each of the *koshas* and associates it with a certain type of **mental process**, a certain kind of inner experience.

The lowest level is *annamaya kosha*, the food sheath, which Sri Aurobindo associates with the physical mind, or sense-mind. This is the level of thought about physical circumstances, immediate surroundings.

Next comes *pranamaya kosha*, the energy sheath, which Sri Aurobindo calls the life-mind or vital mind. This level of being is associated with the breath, the *prana*, and the fundamental life-force. It is also associated with the feelings, the emotions.

Then comes *manomaya kosha*, the mental sheath. This represents inventive, creative thought: the making of novel connections, the combination of ideas. Some people may go through their lives and hardly ever encounter this level of being. However, all creative individuals spend much of their time here.

Vijnanamaya kosha, the intellect, represents higher **intuitive** thought. It is not experienced by all creative people, but rather represents a higher order of insight. When a person experiences a work of art or an idea popping into their mind full-blown, without explicit effort or fore-thought, as though it has come from "on high" -- this is *visnanamaya kosha*, or true creative inspiration.

Finally, *anandamaya kosha*, the sheath of bliss, represents the "causal world." It is what the Kabalists called the *Logos*; the source of abstract, mythical, archetypal forms. The forms in the *Logos*, the sheath of bliss, are too general and too nebulous to be fully captured as creative

inspirations. They extend out in all directions, soaking through the universe, revealing the underlying interdependence of all things.

Beyond the sheath of bliss, finally, there is only the Self or *Atman*: pure, unadulterated being, which cannot be described.

Emergent Pattern

In the Vedantic view, which is fairly representative of the world's spiritual, "wisdom" traditions, divine inspiration is the way of the world. The higher orders of thought represent attentiveness to the emanations of the divine realm, and it is only the lower orders of thought that rely entirely on the mechanical processes of the material world.

Put more concretely, the Vedantic hierarchy would seem to suggest two different kinds of thought. First there is *manomaya* thought, which is on a level above mere physical or emotional reaction, but which is still based on complex manipulations of ideas derived from the physical world. And then there is *vignanamaya* thought, which is based on taking intuitions from the upper realms, and using them to guide one's more physicalistic concepts, one's feelings and actions. In the first, the reflexes are in control; in the second, one's higher intuitions are in control. The realm above intuition, *anandamaya*, is no longer well described as a realm of thought at all; it is a realm of shifting, subtle forms, weaving indescribable patterns.

To the scientifically-minded reader, this may seem all a bit too much. However, I believe that it must be accepted as an accurate report of an inner experience. After all, this is exactly what the creative thinkers quoted above report. There is something different from ordinary ratiocination, something in which ideas seem to come from above or outside.

In terms of subself dynamics, one may say more specifically that this kind of experience tends to occur when the creative subself is in control. It does not happen to ordinary subselves: it happens when one is fully mentally engaged in creative thought, in one way or another. Only when one is absorbed in one's medium, or else abstractly "daydreaming" in a way that engages the creative subself, will the experience of inspiration come. The creative subself is the "medium" by which this kind of inspiration visits one.

Emergent Pattern

Now: what can be done with these intuitive insights into the creative process, in a scientific sense? In order to draw this introspective analysis back into the framework of the psynet model, I wish to propose the possibility that these "higherintuitions" which we experience are in fact **emergent patterns**.

We have seen that complex dynamical systems can give rise to all sorts of abstract, emergent patterns, which are in no way predictable from the equations and parameter values involved. We

have formalized these abstract patterns in terms of language theory. However, we have not tried to look at these emergent patterns from the **inside**.

Suppose a perceptual-cognitive loop, a consciousness- embodying circuit, comes into being, and engages **some** of the mental processes involved in creating a broadly-based emergent pattern. How will the emergent pattern look from the point of view of this particular perceptual-cognitive loop? It will look like a pattern, a structure coming from **outside** consciousness, enforcing itself from nowhere in particular. In short, I propose that --

---

Fourth Principle of Creative Subself Dynamics -- The feeling of "creative inspiration" is the feeling of emergent pattern viewed from the inside -- i.e., the feeling of a perceptual-cognitive loop which encompasses only part of a broadly-based emergent pattern.

There are, then, in this view, two very different kinds of thought. One involves trying to build up structures and ideas, by various techniques. And the other involves taking some emergent pattern which has come from "outside," and seeking to build up **this particular pattern** using the techniques at one's disposal. In the one instance, one is merely working with the immediate low-level patterns that are provided by a certain "neighborhood" of mental processes within the dual network. In the other instance, one is working with an emergent pattern that is "handed down," and using local mental processes to come to grips with this emergent pattern, to make it more workable and comprehensible. The Vedantic identification of different types of thinking makes its appearance here in more scientific clothing.

I must emphasize that these considerations do not disprove the Vedantic analysis in terms of "higher levels of being." For one thing, that is a philosophical analysis, which is not susceptible to falsification. And for another thing, the view of inspiration as emergent pattern may easily be extended to the **transpersonal** realm.

In *Chaotic Logic* I have hypothesized a "Universal Dual Network," a mind magician system which binds together different individuals in a greater, social mind. In this view, Jung's collective unconscious appears as a pool of higher-level mental processes shared amongst the dual networks belonging to different individuals. As well as intersecting at the lowest levels, which form physical reality, individual dual networks are seen to intersect at higher levels as well, reflecting abstract cultural- psychological universals. In this view, one might hypothesize that some creative inspirations are emergent patterns in individual portions of the dual network, whereas others are emergent patterns in this shared upper level of the dual network, this transpersonal realm or "collective unconscious." This view connects neatly with subself dynamics, which views the individual mind as having exactly this same shape: largely unified perceptual/motor processes, largely unified high-level long-term memory, and fairly dissociated middle-level reasoning and belief systems. It thus becomes possible to view the individual mind as a kind of miniature image of the social mind -- a view which is quite attractive, from a philosophical point of view.

These issues of transpersonal psychology, however, would bring us too far afield from the focus of the present book. Instead let us return closer to earth, and focus on the question: What is it about creative subselves that makes them susceptible to this kind of experience? In order to answer this question, we must dig deeper into the **internal** dynamics of the creative subself. That is the task of the next section.

14.6 INSIDE THE CREATIVE SUBSELF

We have talked about the relation between the creative subself and the rest of the mind, and the role of the creative subself in the experience of inspiration. But what actually goes on **inside** the creative subself? To answer this question, we must turn back to some of the complex systems ideas introduced in earlier chapters.

First of all, it is worth reflecting on some of the particular **creative processes** involved in creative thought. The creative process involves making a new reality, out of abstract forms and structures. But how does it come up with the new forms and structures, to be placed in the new reality? We have distinguished two different possibilities: either it invents them in a "bottom-up" manner, or it invents them in a "top-down" manner, guided by some emergent pattern which presents itself as an intuition from "outside." But either way, guided or unguided, what is required is some kind of mechanism for creating new structures.

The psynet model leads one to the view that there are two basic mechanisms here: one corresponding to the hierarchical aspect of the dual network, and one corresponding to the heterarchical aspect of the dual network. The hierarchical aspect of creativity is based on mutating and combining ideas -- it is modeled moderately well by the genetic algorithm. The heterarchical aspect of creativity, on the other hand, is based on letting the mind flow from one idea to other **related** ideas -- it is what is loosely called "analogy."

The combination of analogy with genetic form-creation is the essence of creative process. These are the tools which the creative subself uses to create its new realities. Of course, these same tools are implicit in other subselves as well. But in the creative subself they are given new power. This is the essential point, a point which will bring us back to the role of the perceptual-cognitive loop in the creative subself.

Consider, first, genetic form creation. The idea here is that subnetworks of the dual network are allowed to **mutate** and to **cross over** (by swapping processes amongst each other). This has been shown (in *The Evolving Mind*) to occur as a natural consequence in current models of brain dynamics. And it is introspectively very natural as well: we can all feel ourselves forming new ideas by modifying old ones, or putting parts of old ones together in new ways. The question, then, is **how freely** this is to be allowed to occur. What is the mutation rate, and how much disruption will crossover be allowed to cause? Will networks be allowed to cross over large chunks with each other, possibly leading to totally dysfunctional portions of the dual network? Or will they only be allowed to exchange small sets of processes, thus leading to smaller innovations, and a smaller risk of serious disruption of function?

Next, consider the heterarchical network. This has beenthought of as a multilevel Kohonen feature map, or as a geographical terrain in which nearby regions host similar species. But this is not a static network. In the ecological metaphor, the different animals must be allowed to constantly migrate around, seeking better locations. Analogy works by taking an activity A*B and replacing it with A*C, where C is "similar" to B in the sense that it is **near B in the heterarchical network**. In *The Structure of Intelligence* I have given a comprehensive analysis of different types of analogy using this framework. But the crucial point here is: Where does the topology of the network come from? What determines whether B and C are close to each other in the first place?

More interesting and adventurous analogies will arise if the network is allowed to be somewhat daring in its reorganization. Consider, for instance, Kekule's famous analogy between the benzene molecule and the snake. In his heterarchical network, a "molecule" process and a "snake" process were close together. In most people's minds, this is certainly **not** the case -- the two processes are stored in largely unrelated areas. Kekule's creative process involved the reorganization of his associative memory network in such a way that these seemingly unrelated concepts were brought together. The point is that the creative process involves "wild" analogies, and wild analogies ensue from **experimental reorganizations of the heterarchical network**.

But of course, the hierarchical and heterarchical networks are in the end the same network. This is the essence of the "dual network" model. So what we find is that experimental reorganizations of the heterarchical network are exactly the same thing as adventurous crossover operations in the heterarchical network. The essential quirk of the creative subself, it seems, is a willingness to shake things up -- a willingness to mvoe things around in an experimental manner, instead of keeping them structured the same old way.

In terms of the perceptual-cognitive loop, this suggests that consciousness acts in a slightly different way in the creative subself than it does in ordinary subselves. It still makes ideas into coherent wholes -- but these coherent wholes are not bundled together quite so tightly. Instead of having a thick black line drawn around them, they have a dotted line drawn around them. They can be dissolved at will.

This idea has interesting implications. First of all, in terms of the speculative division algebra theory of consciousness, given at the end of Chapter Eight, one might hypothesize that in the creative subself, coherentization is done in a more **reversible** way. Coherentization can more easily be **undone**, turning coherent, bound-together concepts into reasonably comprehensible and coherent component parts, which are ready to be re-organized in a different way. The "dotted line," in this view, simply corresponds to a more nearly reversible hypercomplex algebra. Consciousness creates autopoietic systems; some of these systems are "nearly" division algebras and some are very far from it. Acting in the context of the creative subself, the perceptual-cognitive loop creates systems embodying algebras that are more nearly reversible.

Next, this conception of creative consciousness mirrors Bohm's description of "enlightened consciousness" in *Thought as a System*. He argues that, in ordinary states of consciousness, we too rigidly separate concepts from each other, thus fragmenting our minds -- that we should separate ideas with dotted lines rather than thick black lines. Enlightened individuals such as Zen

masters, he suggests, coherentize things in a different and more flexible, reversible way. Thus they are able to be "in the world but not of it" -- they are able to follow the routines and ideas of ordinary life without being dominated by them. My suggestion is that creative subselves operate in much the same way. In creative work, we are able to pick up routines and use them without being dominated by them, to a much greater extent than in ordinary existence. This is because our creative subselves have mastered a slightly different way of using the perceptual-cognitive loop.

Creativity and Fitness Landscapes

We have said that the creative subself inhabits a looser, more flexible region of the dual network, a region of the dual network whose autopoietic systems are less rigid and more easily mutable than usual. The next natural question is: Why should this flexibility be possible **here** rather than elsewhere?

But the answer is almost obvious. In the context of the creative subself, the consequences for failure of a thought system are far less strict. Ordinary subselves are molded by contact with the real world, which is notoriously obstinate. The creative subself is molded by contact with abstract ideas and forms as well as perceptions; much of its contact with the outside world is mediated by ordinary ("critical") subselves. These higher-level entities with which the creative subself is in constact are themselves much more flexible and mutable than the lower-level processes that constitute the (mind's view of the) "outside world." Confronted with a more responsive and flexible environment, the creative subself responds with a more responsive and flexible **internal dynamic**.

It has been noted that, for genetic algorithms and other evolving systems, maximum creativity occurs in an environment of moderate harshness. Too little harshness, too benevolent of an environment, causes aimless fluctuations. There is plenty of low-level, spontaneous creativity but it is not directed in any way and thus does not amount to much. Anything goes, whether interesting or not. On the other hand, too much harshness inhibits creativity. In a very harsh environment, there is too much risk involved in generating new forms. New crossover products or mutants will be eliminated immediately, before they have a chance to give rise to possibly fitter offspring. One is likely to see moderately successful individuals dominate, and keep on dominating.

In terms of fitness landscapes, an easy environment is an almost entirely flat landscape, but one which is at a pretty high level, allowing a decent fitness for most everyone. A harsh environment, on the other hand, is largely flat and low withperhaps a few sharp peaks allowing some things to survive. What is needed for creative success is a more evenly structured environment, with lots of peaks within peaks within peaks on all different scales. This kind of fractal fitness function filters out uninteresting forms but still allows significant experimentation with new forms. As noted at the end of Chapter Seven, this is precisely the type of fitness landscape that can be expected to emerge from many complex environments. Real fitness landscapes probably look more like the Julia sets of two- dimensional quadratics than like the graphs of simple polynomials or piecewise linear functions.

The creative subself, I contend, gets a more closely optimal fitness function than the other subselves. The complex systems with which it interacts give it a rich fractal fitness landscape, which encourages and rewards sophisticated experimentation. On the other hand, the fitness landscapes faced by ordinary subselves tend to be much harsher. The outside world can be very punishing, delivering swift and severe rebukes to experimental modes of perception and behavior.

So, if the perceptual-cognitive loops within the creative subself are different, this is because they have adapted to a different situation. They have adapted to a situation in which it is okay to put "dotted lines" around systems, because it is not so bad to have successful systems disrupted in favor of other, possibly worse but conceivably better ones. Ordinary subselves, on the other hand, have perceptual-cognitive loops which have evolved to deal with situations where, once a good procedure has been found, it must be hung onto at all costs.

And how, finally, do these ideas allow us to explain the basic **experience** of creative inspiration? For this we require only one more hypothesis --

Fifth Principle of Creative Subself Dynamics. -- Which subselves will lead to more intense large-scale emergent patterns? The ones that are permitted a large but not too-large amount of creative disruption (memory reorganization and hierarchical-system crossover). Creative subselves tend to fall into this category; everyday subselves tend not to.

Creative inspiration is the feeling of emergent patterns (personal or transpersonal) from the inside. Creative subselves are more likely to give rise to large-scale emergent patterns (the kinds that are large enough for a perceptual-cognitive loop to **fit** inside). Thus, the very looseness which characterizes creative subselves is essential to the experience of inspiration. And this looseness is due, in the end, to the "cushy" environment experienced by creative subselves, due to their relatively abstract and pliant "external world."

Creative subselves are protected from the real world by the other subselves; this is the secret to their success, to their internal dynamics of emergence-yielding flexibility. It follows that, without subself dynamics, true creativity is impossible, or at least very, very difficult to achieve.

For, how can a single region of the dual network deal with **both** the inflexibility induced by the outside world **and** the flexibility required by the creative act? This seems almost contradictory. In order to achieve creativity without subselves, a mind would have to find a way of confronting the outside world in a continually flexible way -- a worthy goal, perhaps, but one that would seem beyond the grasp of those of us leading routine lives in modern society.

The Creative Personality

It would be a mistake, however, to consider the creative subself as something entirely dissociated from the rest of an individual's mind. Different subselves are not **entirely** different. Generally they will share many particular characteristics. And this is the root of the peculiarities of the "creative personality type."

Everyone expects artists, creative people, to be a little flaky, to deviate from social norms and expected patterns of thought and behavior. This phenomenon occurs for two reasons.

First of all, if (at least some of) a person's ordinary subselves were not somewhat flexible in the first place, it is unlikely that they would have given rise to an active creative subself.

And secondly, in the course of living, one can expect some of the flexibility of the creative subself to rub off on the other subselves.

Indeed, if the ordinary subselves of a creative person were **not** unusually flexible, one would likely have an unsuccessful personality system. The Fundamental Principle of Personality Dynamics says that different subselves must relate to each other on a You-You basis, reacting to overall emergent patterns in each other's structure. But it would be very difficult indeed for a strongly **in**flexible ordinary subself to related to an highly creative subself on a You-You basis. Such a subself would not have the cognitive flexibility required to **understand** the emergent patterns making up the You of the creative subself. As a rule, then, one would expect that the ordinary subselves of a creative person would have to pick up a lot of flexibility from the creative subself.

Creativity, Dreaming, Sublimation

This analysis of fitness landscapes brings us, finally, to the relation between **creativity** and **dreaming**. Dreaming, we have argues, serves to weaken the hold of autopoietic thought systems. By giving them what they want, it places them in **benevolent** environments, and thus allows them to weaken their defenses. The current line of thought indicates that basically the same thing is accomplished by creativity -- by taking an autopoietic thought-system and feeding it to the creative subself. By placing a system under control of the creative subself, one puts the system in a nicer environment, where it is much more likely to get what it wants. One thus weakens the grip of the systemover itself, and renders it more amenable to adaptive modification.

For instance, consider someone who is obsessed with physical violence. Their ordinary subselves are filled with thought- behavior systems focussed on physical violence. Actions of other people are perceived as aggressive; these perceived aggressions then lead to aggressive actions; these actions lead to aggressive actions on the other person's part etc. This thought-system can be expected to come to light in dreams: there will be many dreams about winning or losing fights, about being attacked and unable to respond, about vanquishing huge numbers of mighty attackers, etc. These dreams will serve to weaken the hold of the thought- system in everyday life, by allowing it to flex its need for activity in the safer realm of fantasy.

Now, suppose the person in question takes up an art form which allows them to be, in some way, aggressive. Perhaps this aggression is (as in Nietzsche's case) purely abstract: he is destroying other people's ideas. Or perhaps it is more direct: perhaps he is, say, creating sculptures by making a number of sculptures, smashing them to bits, and molding the pieces together in new forms. Perhaps it is totally direct: he has taken up jujitsu. The point is that this sort of activity serves the same function as dreams. The creative subself allows the working-out of the thought-system in a relatively non-punishing context. The aggressive thought-system now

has freedom: it can experiment with different ways of destroying things; it can give vent to its full range of emotions. It can express itself far more fully than was possible in the ordinary external world, which was constantly being resistant. Eventually, in this way, the person may learn to control his own aggression, rather than having it control him. The autopoietic control of the system may be weakened, allowing for better integration with the self-system as a whole.

This parallel between dreaming and creativity is, of course, not a new idea. Creative people are constantly being accused of "daydreaming" when in fact they are actively exercising their creative imaginations. But it is intriguing to see this parallel come out of abstract complexity-theoretic models of dreaming and creativity.

Essentially, what we have arrived at here is the old idea of creativity as **sublimation**. Instead of enacting itself in the real world, thought-systems may enact themselves through the creative subself, on the level of abstract forms rather than concrete actions. The example of aggression was chosen above, but the example of sexuality is perhaps even more appropriate. The thought-perception-and-behavior systems involved in sexuality, as has been observed many times in many particular cases, seem to find at least a partial outlet in creative activity. The loving attention to detail, the devotion and passion which some creators feel for their work, in many cases seems to come directly from sexual thought-systems. In some individuals (e.g. Nietzsche) this is an alternative to directly expressed sexuality; in others (e.g. Picasso) it is merely an additional expression of a sexuality that is amply expressed in real life.

14.7 CONCLUSION

The phenomenon of creativity is a challenge for the psynet model, and for complexity science as a whole. Creativity encompasses nearly all of the themes raised in earlier chapters: emergent dynamical pattern, form creation by genetic algorithms, fitness landscapes, autopoietic thought systems, subselves, memory-freeing dreams, and the perceptual-cognitive loop. And it extends beyond these themes, in directlons barely hinted at here, for instance the direction of transpersonal psychology.

I will conclude with some remarks about creativity, computer simulations, and artificial intelligence. In earlier chapters we were able to bolster some points with definite calculations or computer simulations. We were able to see, in specific cases, how the genetic algorithm creates complex forms, how human feelings give rise to emergent dynamical patterns, how a human belief system holds itself together in a self-supporting way. With the study of creativity, however, we have reached a point where concrete simulation or calculation becomes very difficult. The dual network itself is at the limit of current computational models. To get the dual network to emerge from SEE, for instance, will probably be possible, and is a focus of current research -- but it is not expected to be easy. To get a self system to emerge from the dual network, one would need considerably larger computer systems than we have at present. In a SEE context, for instance, one suspects that -- at a bare minimum -- lattice worlds with hundreds of thousands of cells would be required. If the idea of A-IS is correct, one would need a whole society of such large-scale lattice worlds. And in order to truly simulate creativity, one would have to make this system complex enought to produce, not only "artificial selfhood," **dissociated**

selves -- creative selves, capable of flexible reorganization and large-scale emergent pattern generation.

So what can we conclude, regarding the prospects of artificial intelligence? As for the future, one can be optimistic, given the tremendous speed of recent improvements of computer technology. But at present, computer simulation of true creativity, of a truly fertile psynet, is not a feasible goal.

Mathematical exploration of psynets and creativity is a different story, but also poses its own difficulties. For example, each of the "Principles of Creative Subself Dynamics" given in this chapter is intended as a mathematical theorem in the making. It is not difficult to **formalize** these principles, to put them in mathematical language. However, we do not have the mathematical concepts that are needed to **prove** these statements. The trouble is that everything is simultaneously **fuzzy** and **probabilistic**. There will always be exceptions to every such rule, and the very concepts which are being talked about are defined in such a way as to permit multiple exceptions. One lacks the crisp, clean definitions that are required in order to carry out mathematical proofs. Until we have a mathematics that is able to deal with complex, fuzzily defined discretestructures in a practical way, we will not have a true mathematics of mind.

In the end, then, what I have presented in this book are some **intuitive** models of mind, bolstered by some mathematical and computer work pertaining to "simpler" complex systems. The mathematical and computer work builds up towards but does not quite touch the more abstract models of mental processes. This gap between the one and the other is the essential thing. My hope is that the ideas given here will be helpful to others as they join my in working on the important task of **filling in** the gap.

The mathematical and computational side of complexity science must be elaborated and strengthened, until it is built up to a point where it can more nearly deal with very complex systems like minds, psynets, selves. And our intuitive models of mind must be made ever more concrete and precise, until they are speaking the language of emergent complex systems, as well as the language of experiential and experimental phenomena. There is work here for the mathematician, the computer scientist, the psychologist, the philosopher, the physicist, and the biologist -- but most of all, for those who are willing to look beyond the bounds of individual academic disciplines, and try to build a vision of the **mind as a whole**.

REFERENCES:

Aam, Onar, Kent Palmer and Tony Smith (1995). Series of e-mail messages exchanged amongst *onar@hsr.no*, *palmer@world.com*, *fsmith@aip.org* and *goertzel@psy.uwa.edu.au*.

Abraham, Ralph and Chris Shaw (1982-1988). *Dynamics: the Geometry of Behavior*, Parts 1-4, Aerial Press, Santa Cruz

Abraham, Fred, Abraham, Ralph and Chris Shaw (1991). *A Visual Introduction to Dynamical Systems Theory for Psychology*. Santa Cruz: Aerial Press.

Agha (1992). *Actors: A Model of Concurrent Computation*.

Cambridge: MIT Press

Albert, M.L. and Obler, L.K. (1978). *The Bilingual Brain: Neuropsychological and Neurolinguistic Aspects of Bilingualism*. New York: Academic.

Alexander, David (1995). Recursively Modular Neural Networks. Ph.D. Thesis, MacQuarie University, Sydney

Anderson, J.R. (1983). *The Architecture of Cognition*. Cambridge MA: Harvard University Press.

Anisfeld, Moshe (1984). *Language Development from Birth to*

Three. Hilldale: Erlbaum.

Bachman, Gennady, Ben Goertzel and Matt Ikle' (1994), "On the Dynamics of Genetic Optimization," submitted to *Evolutionary Computation*.

Bagley, R., D. Farmer and W. Fontana, "Spontaneous Emergence of a Metabolism," in *Artificial Life II*, Edited by Chris Langton, C.

Baranger, Michel (1993). An Exact Law of Far-From-Equilibrium Thermodynamics, preprint

Barnsley, Michael (1988). *Fractals Everywhere*. New York:

Addison-Wesley

Barnsley, Michael (1993). *Fractal Image Compression*. New York:

Addison-Wesley

Bateson, G. (1980). *Mind and Nature: A Necessary Unity*, Bantam, NY

Bateson, Gregory and Mary Catherine Bateson (1990). *Angels Fear*.

New York: Basic.

Bell, Timothy, John Cleary and Ian Witten (1990). *Text Compression*, Prentice-Hall, Englewood Cliffs NJ

Bisiach, E., S. Meregalli and A. Berti (1985). "Mechanisms of Production-Control and Belief-Fixation in Human Visuo-SpatialProcessing: Clinical Evidence from Hemispatial Neglect." Paper presented at Eighth Symposium on Quantitative Analysis of Behavior, Harvard University, June 1985

Bisiach, E. and A. Berti (1987). "Dyschiria: an Attempt at its Systematic Explanation," in *Neurophysiological and Neuropsychological Aspects of Spatial Neglect*, Ed. by M. Jeannerod, North-Holland, Amsterdam

Blakeslee, Sandra (1991). "Brain Yields New Clues on Its Organization for Language," *New York Times*, Sept. 8, p. C1

Blanchard, F. (1989). Beta-Expansions and Symbolic Dynamics, *Theoretical Computer Science 65*, 131-141

Bollobas, Bela (1985). *Random Graphs*, Academic Press, Boca Raton FL

Bohm, David (1988). *Thought as a System*. Boston: Routledgeand Kegan

Bouchard, Denis (1991). "From Conceptual Structure to Syntactic Structure," in (Leffel and Bouchard, 1991).

Boulding, Kenneth (1981). Evolutionary Economics, *Sage Publishers*, Beverly Hills CA

Bowlby, J. (1969). *Attachment and Loss: Vol. 1, Attachment*, New York: Basic

Braine, M.D.S. (1976). *Children's First Word Combinations*. Chicago: University of Chicago Press.

Braine, M.D.S. (1971). "The Acquisition of Language in Infant and Child," in *The Learning of Language*, Ed. C.E. Reed. New York: Appleton-Century-Crofts.

Braitenberg, Valentino (1977). *On the Texture of Brains*. New York: Springer-Verlag.

Braitenberg, Valentino and A. Schuz (1991). *Anatomy of the Cortex: Statistics and Geometry*. New York: Springer-Verlag.

Brooks, Rodney (1989). "A Robot That Walks: Emergent Behavior from a Carefully Designed Network," *Neural Computation 1*, 253-262

Bruce, I.D., G. Christos and R.J. Simpson (1993). "Disruption and Combination of Schemata by Crossover," unpublished technical report, Mathematics Department, Curtin University of Technology,

Perth Australia

Brudno, A.A. (1978). On the Complexity of Dynamical System

Trajectories, *Uspeki Matematicheskikh Nauk 33-1*

Burnet, C. (1978). *The Immune System*. San Francisco: Freeman

Butterworth, B. (1980). *Language Production*. New York: Academic

Caramazza, A. (1988). "Some Aspects of Language Processing Revealed Through the Analysis of Acquired Aphasia: The Lexical System," *Annual Review of Neuroscience 11*, 395-421

Caves, Carlton (1990). Entropy and Information, in *Complexity,*

Entropy and the Physics of Information, Ed. W. Zurek. New York: Addison-Wesley

Chaitin, Gregory (1987). *Algorithmic Information Theory*. New York: Addison-Wesley

Chomsky, Noam (1975). *Reflections on Language*. New York: Pantheon.

Christos, George (1992). Investigation of the Crick-Mitchison Reverse-Learning Dream Sleep Hypothesis in a Dynamical Setting, submitted to *Neural Networks*

Christos, George (1994). On the Function of REM Sleep, unpublished manuscript

Christos, George (1992a). Infant Dreaming May Explain Cot Death, *Curtin University Gazette*, December, Perth: Curtin University

Chung, F.R.K. and R.L. Graham (1990). Quasi-Random Hypergraphs, *Random Structures and Algorithms 1-1*, 105-124

Churchland, Patricia, V.S. Ramachandran, Terrence J. Sejnowski (1994), "A Critique of Pure Vision," in *Large-Scale Neuronal Theories of the Brain*, Ed. by Christof Koch and Joel Davis. MIT Press: Cambridge MA.

Churchland, Patricia, V.S. Ramachandran, Terrence J. Sejnowski (1994), "A Critique of Pure Vision," in *Large-Scale Neuronal Theories of the Brain*, Ed. by Christof Koch and Joel Davis. MIT Press: Cambridge MA.

Combs, Allan, Winkler and Daley (1994). A Chaotic Systems Analysis of Circadian Rhythms in Feeling States, *Psychological Record 44*, 359

Cohen, Neal J. and Howard Eichenbaum (1993). *Memory, Amnesia, and the Hippocampal System*. MIT Press: Cambridge, MA.

Crick, F. and Mitchison, G. (1983). The Function of Dream Sleep. *Nature 304* 111-114

Crick, F. and Mitchison, G. (1986). REM Sleep and Neural Nets, *Journal of Mind and Behavior 7*, 229-249

Csanyi, Vilmos (1989). *Evolutionary Systems: A General Theory*, Duke University Press, Durham NC

Danks, Joseph (1977). "Producing Ideas and Sentences," in *Sentence Production: Developments in Research and Theory*, Ed. by Sheldon Rosenberg, Hilldale: Erlbaum

Davis, R. and Principe (1993), "A Markov Chain Analysis of the Genetic Algorithm," *Evolutionary Computation*.

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. London A 400*, pp. 97-117

Devaney, Robert (1988). *Chaotic Dynamical Systems*, New York: Addison-Wesley

Dyson, Freeman (1982). *The Origin of Life*, Cambridge University Press, NY

Edelman, Gerald (1987). *Neural Darwinism*. Basic Books, NY

Edelman, Gerald (1991). *The Remembered Present: A Biological Theory of Consciousness*. Basic Books, NY.

Epstein, Seymour (1980). The Self-Concept: A Review and the

Proposal of an Integrated Theory of Personality

Erdos, Paul and A. Renyi (1960). "On the Evolution of Random Graphs," *Magyar Tud. Akad. Mat. Kutato Int. Kosl 5*, 17-61

Flatto, Leo. Unpublished notes.

Fredkin, E. and F. Toffoli (1982). Conservative Logic, *Int. J. Theor. Phys 21*: 219-253

Freeman, Walter (1991). "The Physics of Perception," *Scientific American*, p. 91

Freeman, Walter (1993). *Societies of Brains*. Hilldale NJ: Erlbaum

Gazzaniga, Michael (Editor) (1995). *The Cognitive Neurosciences*. MIT Press: Cambridge, MA.

Gleick, James (1988). *Chaos: Making a New Science*. London: Sphere.

Goerner, Sally (1994). *The Evolving Ecological Universe*, Gordon and Breach, New York

Goertzel, B. (1991). "Quantum Theory and Consciousness," *Journal of Mind and Behavior*

Goertzel, Ben (1993). *The Evolving Mind*. New York: Gordon

and Breach.

Goertzel, Ben (1993a). *The Structure of Intelligence: A New Mathematical Model of Mind*. New York: Springer-Verlag.

Goertzel, Ben. (1994). *Chaotic Logic: Language, Thought and*

Reality from the Perspective of Complex Systems Science.

New York: Plenum.

Goertzel, Ben (1996). "Mobile Activation Bubbles in a Toroidal Kohonen Network," *Applied Mathematics Letters*, to appear

Goertzel, Ben and Harold Bowman (1995). "Walks on Random Digraphs," *Applied Mathematics Letters*

Goertzel, Ben, Bowman, Harold and Baker, Richard (1993). Dynamics of the Radix Expansion Map. *Journal of Math. and Math. Sci.*, 17-1, p. 93

Goertzel, Ben and Malwane Ananda (1993). Appendix 1 to *The*

Evolving Mind, by B. Goertzel, New York: Gordon and Breach.

Goertzel, Ben, Malwane Ananda and Matt Ikle' (1994), "On the Dynamics of Genetic Algorithms (And Other Evolving Systems," in *Complex Systems: Mechanisms of Adaptation*, ed. by Green and Bossamaier, Amsterdam: IOS Press.

Goguen (1992). Fractal Music. *Computer Music Journal*

Goldberg, David (1988). *Genetic Algorithms for Search, Optimization and Machine Learning*. New York: Addison-Wesley

Grof, Stanislaw (1994). *LSD Psychotherapy*. Alameda CA: Hunter House

Hameroff, Stuart (1989). *Ultimate Computing*

Hanh, Thich Nhat (1975). *The Miracle of Mindfulness*. New York: Beacon Press

Harris, Zellig (1982). *A Grammar of English on Mathematical Principles*. New York: Wiley

Hartmann, E. (1991). *Boundaries in the Mind*. New York: Basic

Hebb, Donald (1948). *The Organization of Behavior*. New York: Wiley

Hopfield, John (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities, *Proceedings of the National Academy of Sciences (USA)*, 79 2554-2558

Hopfield, John, Feinstein, D.I., and Palmer, R.G. (1985). "Unlearning" Has a Stabilizing Effect in Collective Memories, *Nature 304*, 158-159

Jackendoff, Ray (1990). *Semantic Structures.* Cambridge: MITPress.

Kampis, George (1991). *Self-Modifying Systems in Biology and Cognitive Science*, New York: Pergamon

Kapleau, Philip (1980). *The Three Pillars of Zen*. New York: Doubleday

Kauffmann, Stewart (1993). *The Origins of Order*, Addison-Wesley, NY

Kawabata, N. (1986). "Attention and Depth Perception," *Perception 15*, pp. 563-572

Kayne, Richard (1981). "Unambiguous Paths," in *Levels of Syntactic Representation*, Ed. by R. May and J. Koster, Dordrecht: Foris

Kernberg, Otto F. (1988). Between Conventionality and Aggression, in *Passionate Attachments*

Kimura, Y. (1978). *The Neutral Theory of Molecular Evolution*, Cambridge: Cambridge University Press.

Koch, Christof and Joel Davis (1994). *Large-Scale Neuronal Theories of the Brain*. Cambridge, MA: MIT Press

Kohonen, Teuvo (1988). *Self-Organization and Associative Memory*.

New York: Springer-Verlag.

Kolmogorov, A.N. (1965). "Three Approaches to the Quantitative Definition of Information," *Prob. Information Transmission 1*, pp. 1-7

Koza, John (1992). *Genetic Programming*. Cambridge, MA: MIT Press

Kurosh, A. G. (1963). *Lectures on General Algebra*, New York: Chelsea Publishing Company

Leffel and Bouchard (1991) (Eds.). *Views on Phrase Structure*. New York: Kluwer.

Lindenmayer, A. (1978). "Algorithms for Plant Morphogenesis," in *Theoretical Plant Morphology*, Ed. by R. Sattler, The Hague: Leiden University Press.

Luczak, Tomasz (1990). "The Phase Transition in the Evolution of Random Digraphs," *Journal of Graph Theory 14*, No. 2, 217-223

Mandelbrot, B. (1982). *The Fractal Geometry of Nature*, San Francisco: Freeman

Mandler, G. (1985). *Cognitive Psychology: An Essay in Cognitive Science*, Erlbaum Press, Hilldale NJ

Mantica and Sloan (1990). Chaotic Optimization and the Fractal Inverse Problem. *Complex Systems*

Meyer, R. (1956). *Emotion and Meaning in Music*. Cambridge:

Cambridge University Press

Michaelewicz, G. (1993). *Genetic Algorithms + Data Structures = Evolution Programs.* New York: Elsevier.

Minsky, Marvin (1987). *The Society of Mind*. Cambridge, MA:

MIT Press

Nadal, J.P., Toulouse, G., Changeaux, J.P. and Dehaene, S. (1986). Networks of Formal Neurons and Memory Palimpsets, *Europhysics Letters 1*, 535-542

Nietzsche, F. (1888/1969). *Ecce Homo*, English translation by Walter Kauffmann, New York: Random House

Oparin (1965). *The Origin of Life*, New York: Dover

Palm, Gunther (1992). *Neural Assemblies*. New York: Springer-Verlag.

Parisi, G. (1986). A Memory Which Forgets, *Journal of Physics A, 19*, L617-L620

Penrose, Roger (1995). *Shadows of Mind*.

Pesin, Y.B. (1977). Characteristic Lyapunov Exponents and Smooth Ergodic Theory, *Uspeki Matematicheskikh Nauk 32*, 55

Pick, A. (1931). *Aphasia.* Springfield, Ill.: Thomas Press

Posner, Michael and J. Raichle (1994). *Images of Mind.* San Francisco: Freeman.

Pribram, Karl (1991). *Brain and Perception*. New York: North-Holland

Prigogine, Ilya and I. Stengers (1984). *Order Out of Chaos*, Bantam, NY

Prinzmetal, W., D. Presti and M. Posner (1986). "Does Attention Affect Feature Integration?", *J. Exp. Psychol.: Human Perception and Performance 12*, 361-369

Prusinkiewicz, Przemyslaw and James Hanan (1989). *Lindenmayer Systems, Fractals and Plants.* New York: Springer-Verlag.

Pulvermueller, et al (1994). "Periodic Brain Responses During

Cognitive Processing," *PSYCOLOQUY* Electronic Journal

Radford, Andrew (1988). *Transformational Grammar*. Cambridge:

Cambridge University Press.

Rampage, Cheryl (1994). "Power, Gender and Marital Intimacy," *Journal of Family Therapy 16-1*, p. 128

Rizzolatti and Gallese (1988). "Mechanisms and Theories of Spatial Neglect," in *Handbook of Neuropsychology v.1*, Elsevier, NY

Rizzolatti, G., C. Scandolara, M. Gentilucci, and R. Camarda (1985) "Response Properties and Behavioral Modulation of 'Mouth' Neurons of the Postarcuate Cortex (Area 6) in Macaque Monkeys," *Brain Research 255*, pp. 421-424

Rosenfield, Israel (1988). *The Invention of Memory*, New York: Columbia University Press

Rumelhart, D.E., McClelland, J.L. and the PDP Research Group (1986). *Parallel Distributed Processing*. Cambridge MA: MIT Press.

Schmidt-Pruzan, J. and Eli Shamir (1985). "Component Structure in the Evolution of Random Hypergraphs," *Combinatorica 5-1*, 81-94

Serra and Zanarini (1991). *Complex Systems and Cognitive Processes*. New York: Springer-Verlag.

Shaver, Philip, Cindy Hazan and Donna Bradshaw (1988). "Love as Attachment: the Integration of Three Behavioral Systems," in *The Psychology of Love,* Ed. by Robert Sternberg, New Haven: Yale University Press.

Solomonoff, L. (1964). "A Formal Theory of Induction, Parts I and II." *Information and Control 7*, pp. 1-22 and 224-254

Sprott (1993). Automatic Generation of Strange Attractors, *Computers and Graphics 17-3*, 325-332

Swenson, R. (1989). Emergent Attractors and the Law of Maximum Entropy Production, *Systems Research 6-3*

Taylor, J.D. Farmer, and S. Rasmussen (Editors) (1992). *Artificial Life II*, New York: Addison-Wesley

Treisman, A., and H. Schmidt (1982). "Illusory Conjunctions in the Perception of Objects," *Cognitive Psychology 14*, pp. 127-141

Tsal, Y. and L. Kolbet (1985). "Disambiguating Ambiguous Figures by Selective Attention," *Q.J. Exp. Psychology 37A*, 25-37

Uchiyama, Kosho (1993). *Opening The Hand of Thought*. New York: Penguin

Umilta, C. (1988). "Orienting of Attention," in *Handbook of Neuropsychology v.1*, Elsevier, NY

Varela, Francisco (1978). *Principles of Biological Autonomy,*

New York: Elsevier

Weingarten, (1992). "A Consideration of Intimate and Non-Intimate Interactions in Therapy," *Family Process 31*, 41-59; p. 47)

Whorf, Benjamin Lee (1949). *Language, Thought and Reality.* Cambridge: MIT Press

Wigner, E. (1962). *Symmetries and Reflections*, Bloomington IN: Indiana University Press

Wolfram, Stephen (1986). *Cellular Automata: Theory and Applications*. Singapore: World Scientific.

Converted by Andrew Scriven