CONNECTIONISM, DISCIPLINARY IDENTITY AND CONTINUITY

by

Adam Harris Serchuk

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Science and Technology Studies

APPROVED:

Peter Barker, Chairman

Gary Downey

David Lux

Robert Paterson

April, 1989

Blacksburg, Virginia

CONNECTIONISM, DISCIPLINARY IDENTITY AND CONTINUITY

by

Adam Harris Serchuk

Committee Chairman: Peter Barker
Science and Technology Studies

(ABSTRACT)

Connectionism, a new technique for modeling cognitive processes, has been presented by its supporters as a revolutionary advance that will soon replace conventional artificial intelligence (AI) research based on the serial computer. In this thesis, I identify three 'gambits' with which critics attempt to undermine connectionist claims, and show that use of these gambits depends on the status of the respondent's own discipline. I argue that in cases where the respondent's discipline has an accepted identity, for example biology and psychology, they take contradictory stances on the issue of continuity between their discipline and connectionism. By contrast, responses from supporters of AI, which has an uncertain status, insist on a continuous relationship between connectionism and AI. To account for this, I suggest that claims made by both supporters and critics of connectionism, which those actors would regard as purely cognitive, are tacitly structured by Kuhn's model of scientific change. As certain claims which the actors would describe as purely cognitive can be accounted for by the presence in common scholarly parlance of a particular

philosophical model of scientific change, I conclude that in the confrontation between connectionism and conventional AI there exists a complex relationship between social and cognitive processes.

ACKNOWLEDGEMENTS

My list of acknowledgements is extensive, as I received much assistance on this project. First, I thank the chairman of my thesis committee, Dr. Peter Barker, who set connectionism in my path, and then wisely sat back and waited for me to stumble over a research topic. Dr. Barker constantly encouraged me to refine my ideas, and his insistence made this a better thesis. I also thank my thesis committee as whole; Drs. Barker, Gary Downey, David Lux and Robert Paterson showed a surprising amount of faith in letting me struggle with the material on my own, while providing good advice on the frequent occasions when I needed it. I must also individually thank Dr. Paterson, who, as Director of the Center for the Study of Science in Society, authorized support for my attendance at the Emory Cognition Project Workshop on Connectionism in July of 1988, where I received valuable exposure to my topic. In addition, and Dr. Steve Fuller have been extremely helpful, answering my many questions and directing me to important sources. Finally, I thank my colleagues, and . Any insight that I bring to this thesis has been laboriously hammered out in two years of collaboration with and , and I look forward to many more years of fertile hand-waving with them.

Adam Serchuk
April 3, 1989

TABLE OF CONTENTS

ABSTRACT ii

ACKNOWLEDGEMENTS iv

INTRODUCTION 1

CHAPTER ONE: CONNECTIONISM AND AI 3
- I. Conventional Artificial Intelligence Research 4
  - i) Intelligence as problem-solving 4
  - ii) 'Rudimentary' mental abilities 8
  - iii) The appeal of connectionism 10
- II. A Comparison of Technologies 11
- III. A Representative Network 15

CHAPTER TWO: THE RECEPTION OF CONNECTIONISM 20
- I. Responses to Connectionism 21
  - i) Accounts aimed at non-experts 21
  - ii) Philosophers and connectionism 25
  - iii) Biologists and connectionism 28
  - iv) Psychologists and connectionism 31
  - v) Connectionism and conventional AI 32
- II. Discourse Concerning the Verb-tense Model 38
- III. Continuity and Group Identity 49

CHAPTER THREE: AN STS ANALYSIS 55
- I. Two Axes of Continuity 57
  - i) The role of data structures 57
  - ii) The role of the modeler or programmer 63
- II. Continuity and Critical Distance 68
- III. Continuity as a Pervasive Issue in Insider's Accounts 71

CONCLUSION: SIGNIFICANCE AND FUTURE RESEARCH 76

REFERENCES 81

VITA 88

INTRODUCTION

I initially became interested in artificial intelligence (AI) research because it seemed to be an arena in which scientists' epistemic and ontological views influence their research. As I studied the field, I found myself skeptical that AI programs provide accurate models of cognition. My reasons for doing so led directly back to my initial point of departure; I felt--as I still feel--that many of the implicit assumptions concerning the nature of the world and of human cognition built into the AI research program are mistaken.

On first exposure to connectionism, I, no doubt like many others, had hopes that it would provide alternatives to what I saw as the conceptual fallacies of AI. While I reserved judgment on the pervasive claim that connectionism is a 'revolutionary paradigm,' I began my investigation primed to accept connectionism as valuable. My enthusiasm was short-lived. I have come to believe that connectionism and AI are open to similar criticisms. Nevertheless, I find nothing in connectionism that damns it forever. I see no reason--at least, no intellectual reason--why conventional AI, connectionism, cognitive psychology and neuroscience may not symbiotically produce valuable models of cognition in the future.

Although it would be dishonest to pretend that I have no opinion on the validity of connectionist models, the

following analysis does not attempt to evaluate connectionism. Instead, my thesis addresses an issue unexpectedly encountered in my research. The cognitive claims of connectionists seem to depend not only on assumptions about the nature of the world, as one might predict, but also on a tacit model of scientific change. My analysis attempts to account for the actors' positions by bringing these assumptions to light. My aims are three-fold: to show that actors' positions depend on an underlying notion of scientific change, and to explicate that tacit model; to detail the disciplinary context in which the connectionism debate is played out; and to show that my account of the connectionist debate provides understanding absent in the actors' own accounts. I conclude that in this case scientists' cognitive claims have been structured by their involvement in a struggle to establish credibility, and by a tacit notion of scientific change.

CHAPTER ONE
Connectionism and AI

Introduction

In Section I of Chapter One, I present some of the major features of conventional artificial intelligence and connectionist research. In Section II, I contrast the technologies of connectionism and conventional AI. The features characteristic of connectionist models are further clarified in Section III by a description of a representative network, Rumelhart and McClelland's verb-tense acquisition model (1986a). This section provides background for Section II of Chapter Two, where, as part of an analysis of the debate over connectionism, I discuss Rumelhart and McClelland's presentation of the model as well as Pinker and Prince's (1988) critique of it.

While Chapter One focuses on the relationship between AI and connectionism, I do not claim to present a history of AI; I intend the chapter as background for the analysis which follows. Nor do I assert that AI provides the only relevant historical or conceptual background to connectionism. As Chapter Two illustrates, connectionism is the nexus of a variety of research traditions. However, as I will argue, the relationship between AI and connectionism contains elements not found in the relationship between connectionism and, say, psychology; this justifies using AI as a backdrop against which to view connectionism.

I. Conventional Artificial Intelligence Research

i) Intelligence as problem solving: Historical accounts of artificial intelligence research suppose the field to have opened with the observation that computers are not simply powerful adding machines. Boden (1977) offers a typical version of the enabling insight: "Computers do not crunch numbers; they manipulate symbols." The first generation of artificial intelligence researchers (e.g., Ernst and Newell 1969) defined natural intelligence as problem solving, and asserted that the mind thinks by performing the operations of Boolean logic on symbols standing for elements of the external world. This perspective has come to be known as the symbol- processing paradigm of cognition, or 'computationalism.'

Concomitant to the assumption that intelligence is a form of symbol-processing came the corollary that the mind solves problems by following a small set of powerful and general rules, or 'heuristics,'[1] for manipulating symbols. Early researchers defined the task of AI as identification of the heuristics supposed to be the basis of natural intelligence, and the implementation of these heuristics in computer programs.

---

[1]A heuristic usually provides a sensible and efficient way to proceed in a given task, although it guarantees neither that the user will arrive at the optimal solution, nor even that a solution will be found at all.

A program typical of this period (Waldrop 1984, Dreyfus 1972) is the General Problem Solver, or GPS, created by Herbert Simon and Allen Newell of Carnegie Mellon University and J. Clifford Shaw of the Rand Corporation. The authors of GPS defined problems as a progression from an initial state to a goal state; to reach the goal state, each problem was broken down into a series of sub-goals. The initial report of GPS offered the following example of the type of reasoning used by the program:

> I want to take my son to nursery school. What's the difference between what I have and what I want? One of distance. What changes distance? An automobile. My automobile won't work. What's needed? A battery... and so on (Newell, Shaw and Simon 1960).

According to Newell (Ernst and Newell 1969), the principal aim of the GPS project was generality; the authors of the program believed that by using the type of reasoning illustrated in this passage, a computer could solve a wide array of problems.

Although initially praised as an important advance, modern supporters and critics of AI consider GPS a blind alley. Dreyfus (1972), a critic, phrases the problem as one of background information. To solve a problem such as the one outlined above, the program would have to know about commerce, automotive mechanics, finance and the capacities

of the human body. In addition to a set of heuristics powerful enough to solve all problems, one would need to encode the knowledge of how humans relate to the world possessed by the average person. Similarly, Boden (1988), a supporter of AI, describes the problem as concerning "generality" itself: "The most radical failing of the GPS approach was the assumption... that all problems can be represented by a state-space, and that all solutions consist of a search in a state-space." In general, contemporary writers relegate GPS to the background of current research.

By the mid-1960s, AI seemed to be at an impasse, which researchers identified as the size and complexity of the world. They did not admit that the problem of coding sufficient information made AI impossible. They felt, rather, that it would be impractical to tackle it just yet. In the meantime, perhaps a way could be found to cut up the world into smaller chunks.

A representative of the new approach was Terry Winograd's SHRDLU,[2] written in the early 1970s while Winograd was a graduate student at MIT. The program represents a 'micro-world,' a table-top stacked with a variety of blocks. The user of the program enters questions about the position of the blocks, or requests the program to move the blocks. The program performs in a competent and

---

[2]Named by Winograd, according to Dreyfus and Dreyfus (1986), after a meaningless phrase from Mad Magazine.

apparently intelligent fashion, making the judgments about what is possible in the world of blocks that a human being would make. On this basis, fans of the program claimed that it was genuinely intelligent. They assumed, furthermore, that micro-worlds could be combined into a facsimile of human intelligence, which functions, they thought, in a patchwork of distinct domains. This synthesis proved impossible; one cannot, it appeared, simply bind micro-worlds together at the edges to produce an intelligent computer program. Like GPS, the micro-world approach to AI was eventually judged a dead-end.

After abandoning the micro-world approach, AI took two conceptual paths; articulation of data structures, and construction of expert systems. The former path investigates the best way to store information about the world. Appropriate data structures were thought to require (at least) accessibility, abstraction, completeness and adaptability. Accessibility requires that knowledge, once stored, is findable. A system capable of abstraction would have to keep track of items not only as individuals, but also as members of class, to make analogy possible. Because future knowledge might make unsuspected aspects of a situation important, information should be stored even if not immediately relevant, leading to the requirement of completeness. Finally, perhaps the method of representation itself should be adaptable, changing its form in response to

changing use. In general, the search for data structures presumes a division of the natural world into distinct categories, and supposes that the key to constructing an intelligent program is representation of the categories thought to exist in the world.

A second approach to the knowledge problem is the expert system, a computer program meant to reproduce the decisions made by human expert; an example is the 'geologist' program, **Prospector**, described by Duda, Gashnig and Hart (1982). This program, when given geological data, gives advice on the likely locations of mineral deposits. It seems that expert tasks to which we have accorded high cultural status, such as predicting the presence of mineral ore or diagnosing disease, appear more tractable to AI methods than tasks that we consider rather ordinary, such as using language or deciding how to take one's son to nursery school. Expert systems provide a sort of poor man's answer to the problem of knowledge representation. By judicious choice of domain, AI researchers have found that something approximating intelligence can be constructed, because certain domains are almost micro-worlds. While there are theoretical objections to the claim that expert systems truly duplicate the performance of human beings (Dreyfus and Dreyfus 1986), they can in some cases come close.

ii) 'Rudimentary' mental abilities: The equation of intelligence with problem-solving is a major theme in AI. However, sub-fields of AI attend to less conscious mental abilities, such as vision, pattern recognition, and natural speech processing. It has proven difficult to reproduce these 'rudimentary' abilities of the human mind. Although programs that see or understand speech exist, they work well only in constrained, well-defined domains. For example, Tenenbaum's computer vision program (Tenenbaum, Garvey, Weyl and Wolf 1974) can pick out a black telephone, among other objects, in a two-dimensional image of a cluttered desk, but it uses special features of the subject matter to make the identification, rather than a general vision procedure. Hurlburt and Poggio (1988) call this the "expert system approach to vision."

Another way to present the limitations of conventional AI concerns the use of rules. A computer program is essentially a list of rules, statements of the form 'IF x is the case, THEN do y.' Perhaps domains in which AI has been less successful are less amenable to being captured by such a list of rules. In this view, there is a direct correspondence between the successes of the AI venture and the extent to which the domain in question can be separated from its background. Without such a separation, one cannot describe completely all of the 'UNLESS' clauses that must be added to the rules of

procedure (Dreyfus and Dreyfus 1986, Barker 1989). In sum, one might argue that artificial intelligence will not be successful in certain highly contextual domains because the language of computer intelligence, rules, cannot completely describe the complex relationship between foreground and background in these domains.

iii) The appeal of connectionism: The failure of AI adequately to represent 'rudimentary' cognitive tasks such as language use and pattern recognition is used to justify what Greenco (1987) calls the "cognitive counter-revolution," variously known as connectionism, parallel distributed processing (PDP), and neural network modeling. Supporters of connectionism argue that conventional AI rests on a false concept of mind, and that the reliance of conventional AI on the serial computer is therefore misplaced. Connectionism is based on a new kind of computer, the 'connection machine,'[3] and proposes new theories and models of cognition.

---

[3]Most connectionist research does not use the 'connection machine,' but simulates networks on large serial computers. This might be taken as proof that connectionism does not differ from conventional AI, as it does nothing that a conventional computer cannot do. Such a position mistakes the nature of what connectionist models claim to provide. Connectionist networks display no behavior that cannot be reproduced by a conventional program, given sufficient computational power, but the behavior produced is not the only test of a cognitive model's validity. Rather, the issue is whether the model produces that behavior in a biologically and psychologically plausible manner.

Connectionist research shows promise of success at representing tasks such as pattern recognition and the sifting of relevant data from background noise. Connectionism has been less able to reproduce sequential rule-based tasks. Debate between connectionism and conventional AI often centers on which type of skill subsumes the other. For example, proponents of conventional AI (e.g., Fodor and Pylyshyn 1988) argue that if they can show that connectionism reduces to a rule-based computational strategy, then connectionism would lose its status as a revolutionary model of cognition, being instead merely an interesting technology. On the other hand, it would be a startling victory were connectionism to reproduce the successes of conventional AI, or solve any of AI's outstanding problems.

Connectionism, then, can be partially explained as a response to research which preceded it. It is attractive partially because its successes, while so far modest, have come in precisely those areas where AI research has been least successful. This theme will reappear in Chapter Two.

II. A Comparison of Technologies

Conventional artificial intelligence research rests on the computational assumption that the mind and the computer are the same type of entity, namely symbol-processors. AI is therefore tightly bound to the serial

computer. AI views the serial computer as more than simply a convenient technology or a blank page; AI researchers find the computer strongly appropriate for their purposes because they suppose that the structure of the computer mirrors that of the mind.

The serial computer can be identified by two characteristics, its hierarchical construction and its use of rules. Concerning the first of these, the computer is componential, with different levels of control. At the top level is a central processing unit (CPU) directing the activity of the other components of the computer according to the instructions in the program. There is, for example, a separate unit for storage of information. When the program so directs, the CPU can send a piece of information to the memory, where it remains until needed. Second, the activity of a serial computer is governed by a list of rules, called the program. The program directs the CPU to carry out procedures under certain conditions. As noted above, these rules, at their most basic, are statements of the form 'If X is the case, then do Y,' and specify appropriate responses to situations that the program may encounter.

The connection machine differs from the serial computer in four ways. Instead of one complicated central processing unit, the connection machine has many simple units, each attached to many others in a network. Instead

of a hierarchical construction, with one unit controlling the others, no one unit in the network directs the activity of the others, nor does any unit 'know' what is happening to units other than the few to which it is wired. Instead of compartmentalized entities doing a specific jobs, the connection machine contains unspecialized units, all of which perform the same operation. Lastly, instead of a long list of rules, the units of the connection machine are directed by only one rule.

The units of the connection machine are simple switches with two states, on or off. When a switch is on, it sends a small excitatory impulse to its neighbors; when it is off, it sends a small inhibitory impulse. When a unit receives enough excitatory impulses from the surrounding units, it turns on, and begins to excite the surrounding units--some of which will be the same units that helped to turn it on.

When the operator of the network manually stimulates a small number of the units, a pattern of activity results. The wave of excitatory impulses flowing out from the switched-on units meets the waves of inhibitory impulses from the switched-off units; the interfering 'ripples' set the whole network into fluctuating activity. In most cases, a stable pattern is eventually reached, with some units on, and others remaining off. The relationship between the units turned on manually by the human operator at the

beginning of the process and the ones left on at the end is analogous to the relationship between the input and the output of a conventional computer. A specific initial configuration will result in most cases in a unique final configuration.

Pairs of units may have varying connection strengths. That is, the state of a given unit may influence certain neighboring units more than others. This property allows the operator to structure the activity of the network. Suppose the operator wants a specific configuration of manually turned-on units to lead to a specific configuration of turned-on nodes at the end of a cycle of activity. She or he can start the network moving with arbitrary connection strengths between the units, and then compare the final configuration to the desired configuration. The strengths may then be adjusted according to some algorithm, and the cycle begun again, to see if the outcome is any closer to the desired result. This process, called 'back propagation,' if repeated many times, can lead to a network that produces the desired result every time the initial units are stimulated.

The memory of the network does not exist in a separate compartment, but in the connection strengths themselves. It is the network as a whole that 'remembers' how to produce a desired output. More important, the same network, with the same set of connection strengths, can hold

many pairs of initial-desired configurations in this 'memory.' Furthermore, because the information is distributed over the entire network, with no individual unit standing for any identifiable bit of the information, the network can accept damaged input and produce a reasonably accurate answer. In contrast, a serial computer requires clean input; feeding the computer imperfect data generally results in a failure of the program.

Connectionist networks are relevant to cognitive research because the units can be given any label. We can, for example, designate two groups of twenty-six units each to stand for the letters of the alphabet. We can manually turn on members of the first group, calling that a 'word,' adjust the connection strengths, and begin another cycle until we get a certain configuration in the second group, calling that too a 'word.' The promise of connectionism is the possibility of producing intelligent behavior in the domains of language and pattern recognition where conventional AI has been less successful. Section III describes a network that addresses a particular aspect of human language acquisition.

III. A Representative Network

Rumelhart and McClelland's (1986a) verb-tense acquisition model displays features characteristic of connectionist research and provides an important example of

connectionist attempts to represent language. This network, which transforms the base (present-tense) forms of English verbs into the correct past-tense form, addresses a well-documented phenomenon in language acquisition behavior (e.g., Berko 1958, Kuczaj 1977, Bybee and Slobin 1982). As Rumelhart and McClelland describe, past studies have identified three stages through which children pass in learning the past-tense. In Stage One, the child knows a small number of regular and irregular verbs, and uses both types correctly. In Stage Two, the child learns a large body of additional verbs, most of which are regular. At this point, the child often incorrectly regularizes irregular verbs, as in 'camed;' this behavior includes irregular verbs hitherto used correctly. The child also attributes regular forms to spurious words, as in 'glumped.' Rumelhart and McClelland report that researchers (e.g., Berko 1958) have taken these phenomena as evidence that children first learn rules of verb transformation, and subsequently the exceptions to those rules. Finally, in Stage Three, the child speaks naturally, using the past tenses of both regular and irregular verbs correctly. According to Rumelhart and McClelland, a satisfactory cognitive model of language acquisition should exhibit this continuum of behavior.

The verb-tense model is primarily a pattern associator consisting of a two-layer network of nodes. One

layer consists of a pool of 'input units,' the other a pool of 'output units.' Each input unit connects to each output unit, and vice-versa. The input units correspond to the letters making up the present-tense form of a word; these units are manually excited at the beginning of each cycle by the operator. The output units correspond to the past-tense form; at the end of each cycle, the operator compares these to the correct past-tense form. The units may be either on (excited) or off (inhibited). To determine whether a given output unit will turn on, one simply adds the values of the impulses (with inhibitory impulses having a negative impact) received from the relevant input units.

Initially, the influence of each input unit on each output unit is equivalent. After the first cycle, the result at the output layer is generally rather far from the desired result. The operator then adjusts the connection strengths, reinforcing some links and desensitizing others. The input units are stimulated again, and the network runs through another cycle of activity. The connections are re-adjusted, and the procedure repeats, perhaps for hundreds of cycles. When the network has 'learned' to associate the present- and past-tenses, it is given another pair on which to train. As explained above, a single network, with set connection strengths, can hold many pairs of words in its distributed 'memory.'

A central aspect of Rumelhart and McClelland's model is its representation scheme. To symbolize letters, Rumelhart and McClelland use 'Wickelfeatures,' adapted from Wickelgren (1969). In this system, a string of symbols corresponding to letters represents each word. To indicate the order of the symbols, each unit is coded with its predecessor and successor in the string. For example, the three symbols making up the word 'bat' are {#}B{a}, {b}A{t}, {a}T{#}, with the '#' representing the border of a word. Each triplet is referred to as a Wickelfeature; in the verb-acquisition model, each input and output node represents a unique Wickelfeature. In order to pare down the huge number of possible permutations, Rumelhart and McClelland introduce a number of restrictions, such as disallowing those Wickelfeatures that cannot occur in English.[4]

While the success of the model will be addressed in Chapter Two through the accounts of its authors and their critics, my purpose here has been to give the flavor of connectionist research, with regard both to its general aims and the basis of the technology. The principal differences

---

[4]This description simplifies some aspects of the model. For example, Wickelfeatures actually represent not letters, but phonological features of letters. These phonemes, however, are encoded contextually in the manner related. In addition, the model contains a 'Boltzmann device,' which introduces a certain amount of randomness into each cycle. This feature ensures that the network will be nudged out of a non-optimum but stable configuration (i.e., a local maximum). Although these and other important aspects of the network have been glossed in the interest of clarity, their absence does not impair the accuracy of my account.

between connectionism and conventional AI concern the nature of the tasks claimed as successes. Conventional AI is supposed to be appropriate for tasks that can be represented as rule-based procedures, while connectionism is seen as appropriate for tasks concerning the identification of relevant information from a complicated context. This difference can also be seen in the technologies used. Conventional AI uses the serial computer, which functions on the basis of a list of rules, while connectionist models consist of an undifferentiated network of simple nodes, and contain no representation of rules. In Chapter Two, it will be seen that actors' claims about the merits of the two models depend on factors other than the details of the technology.

CHAPTER TWO
The Reception of Connectionism

Introduction

In Chapter Two, I survey the reception of connectionism by other academic disciplines. The first section of the chapter presents responses to connectionism in terms of source, intended audience and the rhetorical strategies used. I find that responses to connectionism use three basic strategies to refute connectionist claims. The first, which I call the 'historical gambit,' appeals to history to cast connectionism as a subset of the respondent's discipline, lacking a distinct identity of its own. The second strategy, which I call the 'correspondence gambit,' consists of arguments that connectionist models do not, in fact, correspond to the natural world, as represented by previously published empirical research. Finally, what I call the 'mere-technology gambit' argues that connectionist architecture is equivalent to serial architecture at the relevant level of abstraction, and that connectionist research provides only an interesting device, rather than a new theory of cognition.

In the second section of Chapter Two, I examine in detail two representative pieces of the connectionism debate, Rumelhart and McClelland's (1986a) account of their 'past-tense acquisition model,' discussed in Chapter One, and Pinker and Prince's (1988) critique of that model. I

highlight the ways in which Rumelhart and McClelland seek to establish the validity of their research, and show also Pinker and Prince's use of the gambits described in Section I.

I conclude this chapter by arguing in Section III that responses to the connectionist challenge by biologists and psychologists differ from responses by conventional artificial intelligence researchers and its supporters. While the two sets of responses to connectionism are similar in that they attack its identity, they differ in the approach they take to the issue of continuity. Biologists and psychologists use continuity in an unclear and perhaps contradictory manner; AI researchers, on the other hand, claim that connectionism is in every way continuous with their discipline. I suggest that this difference can be partially accounted for by considering the disciplinary status of the respondent. I conclude that continuity is a principle actors' category in the connectionism debate, and that the debate provides circumstantial evidence that AI itself has not established a universally accepted group identity.

I. Responses to Connectionism

i) Accounts aimed at non-experts: In the past two years, connectionism has been discussed in The New York Times Book Review (Greenco 1987), Omni (Larsen 1986) and

Business Week (Port 1986). It has merited an entire issue of the Institute of Electrical and Electronics Engineers (IEEE) journal, Transactions on Computers. An article on connectionism by Paul Smolensky (1988) along with peer commentary dominates an issue of Behavioral and Brain Sciences. A supplementary issue of The Southern Journal of Philosophy contains papers from the 1987 Spindel Conference, the subject of which was "Connectionism and the Philosophy of Mind." And the Winter 1988 issue of Daedelus: Journal of the American Academy of Arts and Sciences, devoted to artificial intelligence research in general, gives discussion of connectionism a central position.

Articles on connectionism appearing in general news forums often have a dramatic tone. For example, a Business Week article titled 'Computers That Come Awfully Close to Thinking' (Port 1986) suggests that connectionist networks will solve the "bedeviling paradoxes" of conventional AI. While the author acknowledges a relationship between biology and connectionism, he casts the new field as a close relative of conventional artificial intelligence research, and one which may soon displace its previously dominant cousin.

Both the Business Week article and a similar piece in Omni describe connectionism historically, and depict it as a sort of phoenix, rising from the ashes of early humiliation. The Omni piece begins:

> Scientists had studied neural networks in the Fifties and Sixties, when a simple network, the perceptron, drew researchers the way the Beatles drew screaming girls. But the perceptron couldn't deliver on its promise. In 1969, Marvin Minsky, MIT's artificial intelligence czar, and co-author Seymour Papert, wrote an elegant, almost gleeful mathematical critique of the perceptron and killed off most of the research. A few scientists continued work on modeling the brain's neural ensembles, but many others followed Minsky and company into what is now known as traditional artificial intelligence (Larsen 1986).

Both articles tell the history of connectionism as a classic success story. After an early stage of popularity, the account goes, connectionism was squelched by Minsky and Papert. "Nevertheless," according to the Business Week article, "a band of about two dozen diehards... continued their work in relative obscurity and on shoestring budgets (Port 1986)." The Business Week article concludes with the image of a future in which, "with neural networks serving as eyes and ears, tomorrow's machines will not only be able to watch, listen and talk back, they'll also tolerate human foibles and idiosyncracies (Port 1986)."

More sophisticated audiences usually receive a more complicated story. The Daedelus issue contains pieces apparently intended by their authors to explain the nature of AI to the academic public, who might otherwise be misled. While the collection is aimed at an intelligent but non-specialized audience, the authors included are fairly expert.

According to the preface, one purpose of the issue is to explain "whether the AI endeavor has increased our understanding of cognition (Graubard 1988)." Significantly, eight of the fourteen articles included discuss connectionism when assessing AI. Yet, the volume offers no consensus on the value of connectionism. For example, Papert (1988) takes a dim view of what he calls "the new Prince Charming" and attributes the current popularity of connectionism to "a composite of cultural trends." Papert's position rests on a demarcation between cognitive and cultural factors; he finds acceptance of connectionism unwarranted because he can explain its appeal sociologically. A valid research field, he implies, should be accepted for purely cognitive reasons. Reeke and Edelman (1988) argue that computer intelligence research should pay closer attention to developments in neuroscience, and chide both the conventional AI community and connectionists for "looking sideways to biology." Cowan and Sharp (1988) consider connectionism a research topic within neuroscience,

and not essentially linked to computers. They are fairly positive in their assessment of its potential, and predict an eventual union between connectionism and AI. In general, whether or not the individual authors accept connectionism as valid, they describe it as a challenge to existing ways of thinking about cognition.

ii) Philosophers and connectionism: Many members of the philosophical community, particularly those who label themselves 'philosophers of mind,' also give connectionist theories much attention. Terence Horgan (1987), for instance, writes that: "Connectionism has rapidly become a major movement within cognitive science, and philosophers are naturally very interested." Philosophers generally play the role of informed outsiders to the connectionism debate; while interested in the implications of the research, their status as outsiders protects them against any real threat from connectionism. It may be, in fact, that many philosophers have been receptive to connectionist claims because they do not perceive it as a threat to their group identity.

A recent supplementary issue of The Southern Journal of Philosophy, subtitled "Connectionism and the Philosophy of Mind," contains papers from the 1987 Spindel Conference. As in the Daedelus collection, the authors included here differ in their evaluation of connectionism. Yet, they

concur in portraying it as a challenge aimed at AI. John Tienson (1987) typifies this perspective: "We should keep in mind that connectionism looks attractive in part precisely because it looks promising where (good old fashioned) AI has failed... If connectionism cannot solve these problems, it does not deserve to replace (good old fashioned) AI."[1] While the volume casts connectionism as a provocative unknown, AI itself is described as having had limited success. Horgan and Tienson's (1987) "Settling Into a New Paradigm" offers the most extreme version of the perceived antagonistic relationship between connectionism and AI. They use Kuhnian terminology, casting AI and connectionism as competing paradigms, with the latter as a response to a crisis occasioned by the operational failures of the former. Egan's (1988) commentary, published alongside Horgan and Tienson's paper, does not explicitly deny a Kuhnian crisis in AI, nor that connectionism responds to conventional AI, but only questions whether that response is truly revolutionary.

The appearance of Kuhn's model in this context is significant. Specifically, the use of Kuhnian terms indicates that the users accept, first, that conventional AI has accumulated unresolvable anomalies; second, that connectionism attempts to solve the failures of conventional

---

[1]Tienson uses John Hoagland's acronym, "GOFAI," which stands for "good old-fashioned artificial intelligence."

AI; and third, that connectionism and conventional AI are distinct, mutually exclusive ways of viewing cognition--one cannot endorse both at the same time. Most important, these authors frame their assessment of connectionism by asking whether or not it is a revolutionary successor to conventional AI.

Connectionism likewise receives attention in extended philosophical works. For example, Margaret Boden, in Computer Models of Mind (1988), is unenthusiastic about connectionism. While she concedes that a connectionist network might conceivably learn, she argues that connectionism could not provide theories of learning. Boden imagines a connectionist model that learns how to speak after running for five years, but points out that "psychologists already know that some five-year old connectionist systems, including some to be found in their own living rooms [i.e., children], can speak. The theoretical problem is to explain this." While Boden apparently accepts that connectionist networks are similar to brains, she evaluates connectionism as a research program on the basis of whether it can offer a theory of cognition; this, she suggests, is the only legitimate goal of cognitive research. Connectionist approaches to learning, Boden argues, attempt "painless" research.

Patricia Churchland, on the other hand, is more positive in her appraisal. In the concluding chapter of

Neurophilosophy (1986), entitled 'A Neurophilosophical Perspective,' she writes that connectionism is one of "the types of thing that I take to be a theory of how macro phenomena are produced by neuronal phenomena." Churchland, while she does not commit herself to connectionism, gives it a qualified statement of support.

Boden and Churchland disagree on the value of connectionism. Yet, Churchland's assertion that there is "within the AI community a growing dissatisfaction concerning the adequacy of sequential models to simulate the cognitive processes of creatures with brains (1986)" shares a premise with Boden's counter-attack on what she calls the "current renaissance of connectionism (1988)." Both writers treat connectionism as a divisive challenge, demanding that scholars state their position in the debate, and, with exceptions that will be addressed later, this position seems common in philosophers' responses.

iii) Biologists and connectionism: Responses to connectionism from philosophers take the form of commentaries from interested non-participants. In addition, as the papers of the Spindel Conference illustrate, philosophers tend to view connectionism solely as an alternative to conventional artificial intelligence research. Examination of the discourse surrounding connectionism shows a considerably more complex set of

disciplinary relationships. Most important, as Churchland implies by discussing connectionism in the context of her 'neurophilosophy,' connectionism has a curious relationship with biology, specifically with neuroscience.

Responses to connectionism from neuroscientists take two primary forms. On the one hand, neuroscientists often try to undercut current connectionist research by claiming the modeling of neural networks as an established branch of neuroscience. Cowan and Sharp (1988) maintain that the principles of connectionism have "been around almost since Ramon y Cajal first discovered neurons," and classify current research as "neoconnectionism." Cowan and Sharp give a synoptic history of connectionist concepts, casting the computer as a non-essential tool, rather than a major mover of past research. When they mention John von Neumann, they acknowledge his status as a developer of digital computers, but link his contribution to connectionism to his role of mathematician. In a sense, Cowan and Sharp first neutralize the images of computers that go with von Neumann's name by defining him as a member of a non-aligned field, and then co-opt him into their own camp by implying that in this case he acted as a neuroscientist.

In general, this strategy attacks the identity of connectionism by an appeal to continuity. By claiming early sources of connectionist ideas such as von Neumann, D.O. Hebb, and McCulloch and Pitts as members of their own

discipline, Cowan and Sharp cast connectionism as a subset of their own neurological research program. This response pattern may be called the 'historical gambit;' by taking the voice of a historian, the respondent argues that connectionism merely continues established research and lacks a legitimate identity of its own.

Yet, the connectionist challenge gains strength from claims that the networks are grounded in neurological concepts. This leads to a second response pattern, which attacks the biological plausibility of connectionism. M.M. Segal's (1988) Science review of McClelland and Rumelhart's (1988) Explorations in Parallel Distributed Processing provides an example of this. Segal warns that: "non-neurophysiologists should be aware that several of the central assumptions of these models could turn out to be dead ends for neural network research." Following this warning, Segal's tone becomes much more blunt, and he claims that it has been "known for decades" that certain assumptions of the new research are wrong. While Segal calls the research represented by the book "a landmark," he implies that it is seriously flawed for not considering previously existing work from his own field. This strategy may be called the 'correspondence gambit;' it questions the correspondence of connectionist networks to the natural processes they attempt to model. In general, the

correspondence gambit attacks the connectionists for being unaware of empirical data from the author's home discipline.

Biologists' responses to connectionism, then, take two forms. In some cases, biologists argue that in modeling neural networks, connectionist do what some neuroscientists have done for years. The figures that connectionists claim in order to validate their work historically are pre-empted by the mainstream biological community. In other cases, however, neuroscientists question the biological plausibility of current connectionism. They attack the correspondence between connectionist networks and the mental processes they purport to represent. Frequently, both gambits appear in the same response.

iv) Psychologists and connectionism: Psychologists use the same gambits. Denise Dellarosa begins her (1988) response to Paul Smolensky's "On the Proper Treatment of Connectionism" with a historical gambit: "The appeal of connectionism has its roots in an idea that just won't die. It is an idea that was championed by Berkeley, Hume, William James, Ebbinghaus, and (in a different form) the entire behaviorist school of psychology." While Dellarosa sympathizes with connectionism, she places it in a tradition of psychological thought, and denies its popular status as a new idea.

An instance of the correspondence gambit in a psychological context appears in Keith Holyoak's (1987)

Science review of Rumelhart and McClelland's original Parallel Distributed Processing set. Holyoak, a psychologist, criticizes the PDP models for not matching the observed learning patterns of "organisms from rats to humans;" in doing so, he appeals to existing empirical research from his home discipline.

Considered in terms of continuity and discontinuity, the historical and correspondence gambits appear inconsistent. The historical gambit argues that connectionist concepts are already part of an established research program; this argument asserts continuity between connectionism and the respondent's discipline. On the other hand, the correspondence gambit tacitly accepts certain literature in the respondent's discipline as canonical; connectionism's failure to consider it indicates to the respondent that connectionism is distinct from his or her own field. This asserts discontinuity. Yet the two response patterns coincide in attacking the identity of connectionism. The underlying claim in responses from biologists and psychologists is that connectionism does not deserve the status of an independent discipline.

v) Connectionism and conventional AI: Published responses to connectionism from the conventional AI community are scarce. Numerous writers have compared the two endeavors, and many of these identify themselves as

supporters of conventional AI. Nevertheless, the majority do not identify themselves as computer scientists, or as AI researchers, but as philosophers or psychologists. Examples include Pinker and Prince (1988), Fodor (e.g., Fodor and Pylyshyn 1988; Pylyshyn is a computer scientist) and Boden (1988) who support conventional AI but not connectionism, and Tienson (1987) and Herbert Dreyfus (Dreyfus and Dreyfus 1988a and 1988b), who are skeptics concerning AI but support connectionism. Debate on connectionism within AI itself, and attempts to formally refute connectionist claims, are hard to locate. For example, Artificial Intelligence, a prestigious monthly journal with a lag between submission and publication of approximately one year published neither research articles nor critiques of connectionism in 1988, out of a total of fifty-four pieces. Nor was there a review of Parallel Distributed Processing, although the set appeared on the journal's 'Books Received' list. A survey over the same period of two other AI journals, Pattern Recognition, 'The Journal of the Pattern Recognition Society,' and Pattern Recognition Letters, 'An Official Publication of the International Society for Pattern Recognition,' yields the same result. While both journals feature research on the tasks that connectionism claims as successes, for example the recognition of written characters, neither journal published any material on connectionism in 1988, out of a total of sixty-four and

characters, neither journal published any material on connectionism in 1988, out of a total of sixty-four and ninety-one articles respectively. Again, both journals have a short lag-time, with Pattern Recognition Letters specifically devoted to short accounts of very recent research. If nothing else, the absence of connectionist material in these three journals indicates a decision on the part of the editors that such material is inappropriate, and also supports the claim that conventional AI sees connectionism as an antagonistic field. These journals do not present themselves as forums for debate between AI and connectionism.

Yet, members of the conventional AI community have opinions on connectionism. These opinions frequently appear outside AI journals, in forums devoted to cognitive science. Such statements, while located in an inter-disciplinary environment, exhibit the disciplinary identities of their authors. The writers describe themselves as computer scientists interested in cognitive science, rather than as cognitive scientists, although they frequently make rhetorical appeals to shared cognitive science research standards.

An interesting example of computer scientists' responses to connectionism appears in the issue of Behavioral and Brain Sciences containing Paul Smolensky's (1988) 'On the Proper Treatment of Connectionism.' The

article is followed by thirty-five peer commentaries, and a combined response to those commentaries by Smolensky--a standard format for this journal. Of the peer respondents, eight identify themselves as computer scientists, working either in academia or industry. Their positions can be classified as: one favorable; one non-committal; two politely skeptical; and four unfavorable. The unfavorable responses illustrate refinements and additions to the gambits seen previously in responses from biologists and psychologists.

Many of these commentaries begin by denigrating the furor that has accompanied recent reports of connectionist research. Hunter (1988) writes that "despite vociferous claims like Smolensky's, connectionism's contribution has been modest." Lehnert (1988) attributes the popularity of connectionism within cognitive science to "theorem envy" among psychologists; the connectionists, she implies, are frustrated theoretically oriented psychologists--not computer scientists--for whom the new research program offers an opportunity to speculate. Both Hunter and Lehnert echo Papert's dubbing connectionism "the new Prince Charming (1988)." Lehnert in particular implies that because the appeal of connectionism can be explained in social terms, it does not deserve consideration as an independent research field.

threat to the health of cognitive science: methodology-driven research at the expense of problem-driven research." Lehnert accuses the connectionists of exploring their methodology, rather than using the methodology to explore cognition. While maintaining her identity as a member of the artificial intelligence community, Lehnert makes a rhetorical appeal to the health of cognitive science, presumably to establish a common ground with her audience. But by characterizing connectionism as methodology-driven research, Lehnert moves towards a third response pattern, one that seems to be unique to supporters and members of the AI community.

Responses from artificial intelligence researchers and supporters frequently denigrate connectionism as merely an interesting technology. Such critiques usually rest on arguments intended to show that connectionism reduces to the symbol-processing view of cognition. In a typical example, Fodor[2] and Pylyshyn (1988) assert that: "many arguments for connectionism are best construed as claiming that cognitive architecture is <u>implemented</u> in a certain kind of network...

---

[2]Not all supporters of AI do AI research, nor are they all computer scientists. The unifying tenet of AI and its fans is a belief in the symbol-processing model of cognition, or 'computationalism.' For the sake of simplicity I characterize writers such as Jerry Fodor, a philosopher, as a supporter of AI; because of his strong support for computationalism, my characterization captures the essentials of his position. It bears repeating, at this point, that the computer is not the essence of computationalism; rather, the position asserts that the computer and the brain are similar types of entities.

Understood this way, these arguments are neutral on the question of what the architecture is." Thus, while Fodor and Pylyshyn admit that connectionism provides a novel machine, they claim that the output of that machine is, at some abstract level, equivalent to what serial computers--and brains--have always done. This 'mere-technology gambit' concedes that perhaps the connectionists can do certain tasks more conveniently, but denies that this is significant.

In the commentaries on Smolensky by computer scientists, Eric Dietrich and Chris Fields (1988) provide an example of the mere-technology gambit. They argue that Smolensky is (logically) forced into an untenable position, to which they offer an alternative. After describing this, they conclude:

> We wish to claim no credit whatsoever for this view. It was formulated over 30 years ago by Ross Ashby (1952), and it appears to us to provide... a quite adequate foundation for computational psychology and cognitive science. In particular, it shows us clearly how connectionism, viewed not as a revolution, but as a valuable addition to our methodological tools, can achieve the goals Smolensky sets out in his conclusion.

This passage illustrates the historical gambit; Dietrich and Fields imply that everything relevant has been said long ago. Additionally, the passage indicates that opponents often see connectionism as a revolutionary challenge. Their main point, however, is a more extreme version of Lehnert's response. While she claims that connectionism focuses on the methodology of modeling cognitive processes, rather than the subject of the model, Dietrich and Fields argue that connectionism only provides a new modeling technique.

Finally, the responses to Smolensky exhibit a variation of the correspondence gambit. Touretzky (1988) writes: "I cannot prove Smolensky... wrong, but I believe (his) principle and most radical claim, that formal symbolic theories of intelligence will turn out to be inadequate for explaining human performance, is very badly in need of some supporting data." In essence, Touretzky implies that the connectionist claims are unsupported by appropriate evidence.

II. Discourse Surrounding the Verb-Tense Model

Responses to connectionism identify the central source of contemporary connectionist literature as the collections Parallel Distributed Processing, Volume 1 (Rumelhart, McClelland and the PDP Research Group 1986) and Volume 2 (McClelland, Rumelhart and the PDP Research Group 1986), and Explorations in Parallel Distributed Processing

(McClelland and Rumelhart 1988). These works, henceforth referred to collectively as the PDP set, are universally accepted as seminal.

Papert (1988) refers to PDP both as "the current connectionist manifesto," and "the current bible of connectionism." Dreyfus and Dreyfus (1988a) agree, and report that the first two volumes "sold six thousand copies the day it went onto the market, and [that] thirty-thousand copies are now in print." In reviews that remain somewhat skeptical about connectionism in general, Holyoak (1987) calls the first two volumes the "focus" of the connectionist movement, while Segal (1988) calls the third a "landmark." The Emory Cognition Project Workshop on Connectionism, held in the summer of 1988, asked that participants read the first four chapters of Volume 1 (McClelland, Rumelhart and Hinton 1986; Rumelhart, Hinton and McClelland 1986; Hinton, McClelland and Rumelhart 1986; Rumelhart and McClelland 1986b) before attending. Tienson (1987), who, like Papert, calls the collection "the bible of connectionism," also directs readers to the first four chapters of Volume 1 for an overview of connectionism.

Within the PDP collections themselves, Rumelhart and McClelland's (1986a) model of verb-tense acquisition can be defended as 'typical' connectionism on two grounds: as the presentation of the model in Chapter One shows, the model displays internal features accepted as definitive of

connectionism; moreover, both sides of the debate accept it as an appropriate battlefield on which to argue the merits of connectionism.

It is often claimed that conventional AI has had limited success in reproducing natural language (e.g., Dreyfus 1972, Waltz 1982, Waldrop 1984, Rumelhart and McClelland 1986a), and supporters of AI such as Pinker and Prince (1988), Boden (1988) and Frieden (1988) all identify language as an important issue for both connectionism and conventional AI. Pinker and Prince write that "language is a crucial test case" for connectionism, and go on to state that "many observers... feel that connectionism, as a radical restructuring of cognitive theory, will stand or fall depending on its ability to account for human language." Pinker and Prince, and Boden as well, apparently endorse this test, for they argue on the basis of their critiques of Rumelhart and McClelland's verb-tense model that connectionism is inadequate.

Rumelhart and McClelland themselves present the model as part of a larger agenda concerning linguistic cognition. The structure of Rumelhart and McClelland's (1986a) account is simple. After defining a standard view, they state their opposition, present an alternative, and describe a model in order to demonstrate the plausibility of their position. Although the model provides the ostensible subject of the account, the authors place it in the context of a larger

agenda. "Put succinctly," they write, "our claim is that PDP models provide an alternative to the explicit but inaccessible rules account of the implicit knowledge of rules." This alternative is used to justify a call for "revised understanding" of language.

For Rumelhart and McClelland, the standard account of language acquisition--typified by Pinker (1984)--asserts that children learn language by subconsciously proposing rules of usage, which are judged, and then saved or rejected, on the basis of evidence provided by adult speech. This received view posits the existence of explicit linguistic rules that are irretrievable at the conscious level of cognition. By contrast, Rumelhart and McClelland "suggest that lawful behavior and judgement may be produced by a mechanism in which there is no explicit representation of the rule." They maintain that if a PDP-like network provides the mechanism for learning language, then people will behave in a manner that can be described by rules, even though these rules do not exist in the mind of the speaker.

Rumelhart and McClelland use two basic tactics to support their position; an appeal to the psychological plausibility of the model, and an appeal to the value of excess empirical content.[3] Regarding the first of these,

---

[3]While I have argued that the verb-acquisition model is representative of connectionist research, I am not prepared, on the basis of the evidence presented here, to argue that Rumelhart and McClelland's account and defense of that model is typical--although I suspect that such a claim would be born out by future

the initial section of their account describes 'The Phenomenon,' discussed in Chapter One. While short, this section is central to Rumelhart and McClelland's presentation. By citing studies from developmental psychology, they attach themselves to an established empirical tradition. Moreover, their claim that models of language should not represent only correct adult usage, but rather the acquisition of behavior patterns, including characteristic incorrect usages, has a rhetorical value. By stressing the incorrect regularization of irregular verbs (e.g., 'camed') and the regularization of spurious words (e.g., 'glumped'), and by requiring a model of cognition that makes and recovers from the same mistakes evident in human learning, Rumelhart and McClelland portray themselves as taking cues directly from the real world, and thereby doing good psychology.

The account contains frequent references to psychological research. The section titled 'The Simulations,' in which Rumelhart and McClelland compare their findings to established psychological research, contains claims such as: "The type of research just described shows up very clearly in data Bybee and Slobin

---

research. For this reason, I call their emphasis on excess empirical content and psychological plausibility 'tactics,' whose presence is clear in a particular account, rather than 'gambits,' whose use is common to a community. By contrast, I have argued that the historical, correspondence and mere-technology attacks on connectionism are indeed gambits, as their use is common to entire disciplines.

(1982) report from an elicitation task with preschool children." In the same vein, they relate that: "The two curves came together rather late, consistent with the fact that, as reported by Kuczaj (1977), these past+ed forms predominate for the most in children who are exhibiting rather few regularization errors of either type." By tying their work to existing empirical studies from within the psychological tradition, Rumelhart and McClelland attempt to make their research recognizable as a continuous outgrowth of that tradition. In addition, by adopting various conceptual tools from past researchers, such as Wickelgren's (1969) representation scheme, described in Chapter One, and Bybee and Slobin's (1982) typology of verb classes, Rumelhart and McClelland present their work in a form that can be integrated into ongoing research by their audience.

To this extent, Rumelhart and McClelland cast themselves as continuing a tradition. However, they go beyond simple citation of established research in their attempt to make their work relevant to psychology; they challenge past researchers at certain points. For example, while Rumelhart and McClelland appeal to the authority of Bybee and Slobin's research in validating their model, they question the interpretation that these 'ancestors' gave to their own data:

> Bybee and Slobin argued that Type VIII verbs were the most difficult because the past and

> present tenses were so phonologically different that the child could not easily determine that the past and present forms of the verbs actually go together. Again, our simulation showed Type VIII verbs to be the most difficult, but this had nothing to do with putting the past and present tenses together since the model was always given the present and past tenses together... Type VIII verbs are most difficult because the relationship between base form and past tense is most idiosyncratic for these verbs.

In this passage, Rumelhart and McClelland present as a strength of their account what might otherwise appear as a weakness. Because children are usually not told that one word is the past-tense of another, it seems unnatural for the words fed to the model to be labeled in this way. Given Rumelhart and McClelland's strategy, this might be a flaw in their presentation. But by presenting it as a premise in an interesting argument they cast it in a positive light. The passage shows that, rather than merely building on psychological research, Rumelhart and McClelland apparently wish their own work to influence psychology, through new interpretations of accepted data. They attempt to establish credibility not only by citing psychological research; they

try also to integrate connectionism into mainstream psychology.

The second tactic used by Rumelhart and McClelland to validate the verb-tense acquisition model is an appeal to the value of excess empirical content. They write that: "Not only can (the model) respond correctly to the 460 verbs that it was taught, but it is able to generalize rather well to the unfamiliar low-frequency verbs that it had never been presented during training." Once the network contains the relationships between nodes necessary for handling the learning set, it can correctly conjugate verbs on which it was not trained. Like children, it treats new words correctly. While Rumelhart and McClelland do not claim that this ability is serendipitous, they deny that the model was specifically designed to produce such favorable results. Similarly, they also present the tendency of the model to make errors characteristic of human learning as excess empirical content. The errors, they claim, are a natural result of the architecture of the model, and evidence for its validity.

Like the focus on psychological plausibility, Rumelhart and McClelland's emphasis on the excess empirical and theoretical content of their model is an appeal to continuity. Where the first of these tactics looks backward, and casts connectionism as the outgrowth of psychological research, the second looks forward, and

presents connectionism as a fertile basis for future psychological studies. Taken together, the two indicate that Rumelhart and McClelland strive for a peaceful relationship between their own work and psychology.

The final portion of the account combines both tactics. In their conclusion, the authors write: "In addition to our ability to account for major known features of the acquisition process, there are also a number of predictions that the model makes which have yet to be reported. These include..." By hypothesizing aspects of the phenomenon that future empirical studies of humans may discern, Rumelhart and McClelland continue to bind their work to the psychological tradition. Moreover, the postulation of hitherto undetected features of human learning counts as excess theoretical content, whether or not future research actually finds such features.

Rumelhart and McClelland end their account by restating their larger agenda: "We view this work in past-tense morphology as a step towards revised understanding of language knowledge, language acquisition, and linguistic information processing in general." Here, they counter-balance their stress on continuity between connectionism and psychology; McClelland and Rumelhart view their work as revisionary, presenting a challenge to established perspectives. Although the target of this revisionary impulse is not immediately apparent, Section III of my

Chapter Two will argue that it can be discerned between the lines of Rumelhart and McClelland's account.

Pinker and Prince's (1988) critique of the verb-tense model addresses Rumelhart and McClelland's strategies directly. Pinker and Prince--who are psychologists--take the statement of the received view as accurate, and proceed to defend it. They suggest that language is indeed a "crucial test case" for models of cognition, and argue that because the verb-tense model cannot do what Rumelhart and McClelland claim, their cognitive agenda ought to be rejected.

Pinker and Prince write that: "There is no question that (the model) is a valuable demonstration of some of the things that PDP models are capable of, but our concern is whether it is an accurate model of children." Here, Pinker and Prince use a correspondence gambit. While Rumelhart and McClelland represented themselves as modelling 'real' behavior, Pinker and Prince argue that the model does not, in fact, behave as people behave. At the same time, they echo Lehnert's version of the mere-technology gambit, implying that the model is methodology-driven research.

Pinker and Prince point out that the model never learns the appropriate past-tense of certain verbs, although humans eventually do learn the proper form. Additionally, it "easily models many rules not found in any human

language." Pinker and Prince argue that the model is too powerful, and that for this reason it must be unsound.

By attacking the psychological plausibility of the model, Pinker and Prince attempt to negate Rumelhart and McClelland's claims to excess empirical content. While they concede that "precise empirical predictions flow out of the model as it operates autonomously, rather than being continuously molded or reshaped to fit the facts by a theorist acting as a deus ex machina," they nevertheless claim that the model "makes false predictions about derivational morphology, compounding and novel words... (and that) it makes incorrect predictions about the reality of the distinction between regular verbs and exception in children and in languages." That is, Pinker and Prince claim that the model makes predictions about the nature of language that are simply wrong.

Moreover, Pinker and Prince attack the model qua theory. "One thing should be clear," they write. "Rumelhart and McClelland's model does not differ from a rule based theory in providing a more exact account of the facts... the network gives a crude, inaccurate and unrevealing description of the very facts that standard linguistic theories are designed to explain." Not only is the model false, because it does not correspond to the real world, but it is unhelpful, because it does not explain as much as the received view.

Pinker and Prince conclude their account by acknowledging Rumelhart and McClelland's larger agenda, and rejecting it completely: "the claim that the success of their model calls for a revised understanding of language and language acquisition is hardly warranted in light of the problems we have discussed... the model does not give superior or radically new answers for the questions it raises." Pinker and Prince show here that they give great importance to Rumelhart and McClelland's claim to have offered reasons for revising the received view of cognition. They take the claim that connectionism is a revolutionary successor not as background to the debate, but as a central part of it.

III. Continuity and Group Identity

Hostile accounts uniformly cast connectionism as a challenge to traditional ways of thinking about cognition, using a small number of strategies to counter connectionist claims. What I call the historical gambit portrays connectionism as an instance of an existing school of thought within the respondent's discipline. What I call the correspondence gambit attacks the relationship between connectionism and the real world, or criticizes the connectionists' interpretation of empirical data from the respondent's discipline. What I call the mere-technology gambit asserts that connectionism provides a convenient

method for modeling cognitive processes, but no revolutionary theory of cognition. The common thread in these gambits is an attack on the independent identity of connectionism.

Nevertheless, insiders perceive connectionism as presenting a different kind of challenge to psychology and biology on the one hand, and artificial intelligence on the other. In the case of biology and psychology, connectionism appeals to be acknowledged as a legitimate enterprise, synthesizing previous research. In the case of artificial intelligence research, however, connectionism is seen as a threat to AI's identity, and a challenge to the validity of what AI has accomplished so far.

This distinction can be phrased in terms of continuity. Use of the historical gambit turns the rhetorical weapon of continuity against the connectionists, by claiming that connectionism lacks an identity of its own precisely because the relationship is continuous one. By contrast, the correspondence gambit argues that the relationship between connectionism, and psychology or biology, is discontinuous, because connectionism has ignored the empirical content of those fields. Thus, biologists and psychologists take contradictory stances on the question of continuity.

Responses from AI researchers replace the correspondence gambit with the mere-technology gambit, and

admit no discontinuities at all. These responses attack the identity of connectionism, and reject its claims to be a revolutionary successor to AI. Furthermore, responses to connectionism from AI supporters are substantially more vehement than those from biology and psychology.

Consideration of Rumelhart and McClelland's account of the verb-tense model helps to explain this difference in reactions to connectionism. Rumelhart and McClelland strive to establish a friendly relationship between their work and psychology. As Pinker and Prince's remarks show, it is not apparent whether they have succeeded. But, they clearly wish to build on established psychological research, and to influence future research. Rumelhart and McClelland do not aim their revolution at the mainstream psychological community; in fact, although they do not refer to AI explicitly, that field seems to be their target.

Rumelhart and McClelland state that their purpose is to provide an alternative to the "explicit but inaccessible rules account" of language. As Chapter One pointed out, a central tenet of the symbol-processing view of cognition is that the brain and the serial computer each manipulate symbols by following rules. By claiming that language is not rule-based, Rumelhart and McClelland call into question the practice of implementing cognitive models in serial computers; this, in turn, must be taken as an attack on the validity of AI as a whole.

Identifying AI as the connectionists' target helps explain why responses to connectionism from AI are more unified in their stress on continuity than responses from biology and psychology. Simply put, AI is more directly threatened. The connectionist threat to replace AI, and the reaction to that threat, make sense only when considered in the context of the status of AI itself. Unlike biology or psychology, many insiders profess skepticism over whether AI is a successful research discipline at all. For example, John Tienson (1987) maintains that: "There is a Kuhnian crisis in (AI), brought on by a pattern of unfulfilled promises and disappointing results... Clusters of problems... have resisted serious progress for more than a decade, and the appeal of connectionism should be understood in that light." Hilary Putnam titles his (1988) article on AI in the Daedelus collection "Much Ado About Not Very Much," and asks "What is all the fuss about?" Responses--or the lack of responses--from conventional AI to connectionism must be understood in the context of frequent public inquests such as these. 'Conventional' AI does not have the same status as 'conventional' biology, nor even 'conventional' psychology, nor does it have the communal sense of security that comes with a heritage of successful research.

This analysis helps explain why positions on connectionism vary between disciplines. Although

philosophers do not universally support connectionism, they are the group most disposed to consider connectionist claims positively. This may be due the protection afforded philosophers by their role as outside analysts. The status of philosophy is not threatened by connectionist claims because philosophers traditionally maintain a certain distance from their subjects. Biologists and psychologists are equally unthreatened, but for a different reason. While biologists and psychologists are insiders to the debate, rather than outside analysts, the established identities of their disciplines protect them. While biologists and psychologists frequently respond to connectionist claims in a negative manner, the authors argue from a position of strength, and need take no coherent stand on the issue of continuity. By contrast, responses from practitioners and supporters of AI stress total continuity. The reason for this may be that neither members of the AI community nor members of other disciplines perceive AI as possessing an established group identity or general academic credibility. Finally, this explanation also explains the vehemence of Pinker and Prince's critique and of Fodor's position as well. While Pinker and Prince are psychologists, and Fodor is a philosopher, their remarks show that they have a high level of commitment to the symbol-processing model of cognition, and they react to the perceived threat of connectionism as computer scientists do.

Connectionism's threat to AI may also be usefully represented in terms of 'insiders' and 'outsiders.' AI has traditionally been outside academic orthodoxy, trying to establish credibility in the face of skepticism. The connectionist challenge, however, casts AI as an outdated insider group, and connectionism as a vital new outsider. This presents an odd dilemma for AI; while AI researchers would presumably consent to accept the mantle of established status, rhetorically conferred upon them by the connectionists, that mantle carries with it the challenge that the insider status of AI is undeserved. One might expect a certain amount of ambivalence within AI towards picking up the challenge from the self-proclaimed 'outsiders.'

In sum, analysis of the connectionism debate leads to two conclusions. The first, which will be pursued in Chapter Three, is that continuity and discontinuity play central roles in insiders' assessments of connectionism. The second is that the debate appears, at a very basic level, to concern disciplinary identities. What seems strange is the insiders' perception that connectionism threatens to invalidate and totally replace AI, and moreover that AI, which at first appears accredited and secure, actually may not have an established identity of its own.

CHAPTER THREE
A Science and Technology Studies Analysis

Introduction

Chapter One of this thesis provided background on connectionism, and described a representative network. Chapter Two considered connectionism as the focus of social negotiation. I argued in Chapter Two first that perceived continuity or discontinuity between connectionism and better established fields is a central issue in insiders' assessments of connectionism, and second, that the debate over connectionism should be analyzed in terms of group identities. In Chapter Three, I attempt to establish critical distance from my subject of study.

While a partisan of a given side of the debate might be disgruntled at the amount of time I give to an opposing position, and while it might be claimed that I have misinterpreted some position, my aim in the preceding analysis has nevertheless been representation, rather than evaluation. I have not judged the validity of connectionism. My subject has rather been the debate over the validity of connectionism, the strategies used to establish or undermine acceptance of connectionism, and above all, the roles of continuity and identity in that debate. In pursuing this aim, I have not said anything that participants in the debate could not conceivably have said themselves. Indeed, as I have shown, many participants do

'step outside' the debate in order to explain their opponents' positions in historical or sociological terms. While my position as a disinterested observer may make my account more accurate than an insider's account, it is also possible that I am unaware of some crucial aspect of the debate known to insiders. An unanswered question, then, is whether my outsider's account provides insight that could not be provided by an insider's account.

In Section I of this chapter, I present two conceptual continuities between conventional AI and connectionism. Although actors' accounts have led me to continuity as an important issue in the debate, my own version of continuity does not appear in their accounts. In Section II, I question the actors' use of continuity as an argument for or against validity; I propose the continuities presented in Section I as examples of how continuity may be phrased so that it does not constitute an argument for invalidity. In Section III, I suggest attributing the central role of continuity in the debate over connectionism to feedback from the philosophy of science into science itself. Here, again, the actors involved might not accept my account, as I suggest that the presence of continuity in the debate can be partially explained by non-cognitive factors.

I. Two Axes of Continuity

i) The role of data structures: Conventional AI maintains a continuing quest for appropriate ways of arranging information. This search for data structures has produced a tension between realism and convenience. The data structure, it is thought, should reflect the nature of the information, but also the nature of the task. That is, AI research assumes that the real world is arranged into distinct catagories, and that for a computer model to be successful, its data structures should mimic those catagories thought to exist in the world. For specialized tasks, however, researchers may not consider the arrangement of the world, but simply engineer a data structure suited to the task. Thus, one finds AI research in which claims about the realism of the categories utilized are made; one also finds cases where it is claimed only that the program using the data structure will work. In some instances, traces of both approaches may be detected.

A classic example of the former approach, which aims at representing the categories thought to exist in the world, is Schank's (1972) attempt to define a small number--originally fourteen--of concepts of primitive dependency. Schank's 'primitives' included 'Mtrans,' denoting the transfer of mental information, and 'Propel,' denoting intentional movement of another object (see Waltz 1982 or Boden 1977 for further description). The goal of this

project was to define a set of concepts with which all human action could be described; one can see it as an attempt to define 'ultimate' data structures.

Lenat, Prakash and Shepard's (1986) CYC project, which attempts to encode the contents of a desk-top encyclopedia, provides a more recent example. This group argues that even very specialized computer programs (specifically, expert systems) require a huge amount of information about the world in order to recognize exceptions to general rules. The team plans first to create a new language--CYC--engineered specifically to hold the information. This language itself can be seen as a huge data structure, embodying the programmers' suppositions concerning the structure of the world.

By contrast, computer chess-playing programs often take the second of the two approaches noted above; programmers use a conceptualization of the chess game that produces a (moderately) successful program, rather than a conceptualization similar to that of a human player. Generally, chess programs arrive at moves by constructing a 'decision tree,' a representation of the possible futures of each legal move available to the program at each successive turn; the skill of the program depends on how far into the future it follows each branch. The program thins this immense mass of potential options by rating the branches according to a scale of 'goodness.' A program called CHESS

4.5, for example, considers a tree with about 3.5 million branches (Dreyfus 1972).

This approach produces a program that plays chess, although arguably not the best chess. However, the procedure used by the program seems to have very little similarity to human cognition. As Newell, Shaw and Simon remark (1963), and Dreyfus (1972) emphasizes, "the best evidence suggests that a human player considers considerably less than 100 positions in the analysis of a move." That is, the authors suggest that humans somehow 'zero in' on a particular aspect of the game, and then construct a small decision tree. The way in which the chess program conceptualizes the world seems to have more to do with the capabilities of computers, specifically their skill at rapid calculation, than with the way in which human play chess. This genus of AI program has a pragmatic, rather than a representational, goal.

The above examples are united by a common assumption that the world is divided into categories, and that the data structure provides the key to acting correctly, whether or not it attempts to represent real-world categories. The programmers implicitly define intelligence as the ability to manipulate appropriately arranged information. Given this definition, the success of the program is thought to depend on the data structures it uses.

Data structures also occupy a central position in connectionist models. In their (1986a) account of the verb-tense acquisition model, Rumelhart and McClelland write: "the input and output target patterns--the base forms of the verbs and the target patterns of these verbs--must be represented in the model in such a way that the features provide a convenient basis for capturing the regularities embodied in the past-tense forms of English verbs." The system of Wickelfeatures that Rumelhart and McClelland adopt is a data structure, a method of conveying information in order to do a certain task.

The Wickelfeature system contains a tension between convenience and realism similar to that found in conventional AI data structures. While the model makes explicit reference to empirical data drawn from the 'real' world (e.g., Kuczaj 1977, Bybee and Slobin 1982), Rumelhart and McClelland admit that their data structure "contains several arbitrary properties." One such property addresses the issue of sequence. The nodes at the input level of a connectionist network are in fact stimulated simultaneously, and the nodes at the output level should also be interpreted as an unorganized group. Arrangement of the nodes into an ordered whole must occur in the mind of a human observer. In the case of the verb-tense model, this presents a problem, as a word is not a <u>set</u> of letters, but rather an ordered <u>sequence</u>. Thus, Rumelhart and McClelland adopt the

device of characterizing letters by referring to the preceding and succeeding letters--that is, a given node denotes '{c}A{t}' rather than 'A' as the middle character of the word 'cat.'[1] While the brain may use such a device, Rumelhart and McClelland cite no research to substantiate this; rather, their motivation seems to be the difficulty that connectionist networks have in handling sequence, a recurrent issue in connectionist literature (e.g., Hinton 1988, Rumelhart and McClelland 1986b).

In the verb-tense model, the selection of an appropriate data structure manages a crucial limitation of connectionist networks, their difficulty in representing order. One need not judge this methodological strategy as legitimate or illegitimate; it is significant, however, that both conventional AI researchers and connectionists accept it as legitimate in their own work--although they may object to its appearance in opponents' research. As the above examples illustrate, both the new models and the old share the notion that the mind--whether located in a serial computer, a connectionist network or, presumably, a brain--is in some way 'given' categories which correspond to categories existing 'out there' in the world.

---

[1]As pointed out in Chapter One, this is a slight simplification that captures the essential features of the Wickelfeature system. In fact, Rumelhart and McClelland divide each letter into six phonemes, which are then identified contextually in the manner described above.

The implicit assumption that an AI program or connectionist network will be most successful when its categories are identical to those thought to exist in the world can be rephrased by saying that both the new and the old techniques accept the role of the human programmer or modeler. The designer, rather than the putative mind, does the 'work' of defining an appropriate data structure. Again, one need not charge that this is unrealistic or duplicitous; it is often proposed (e.g., Cowan and Sharp 1988) that the programmer or modeler does for the technological model what God or evolution did for the human brain. On the other hand, alternative hypotheses exist; for example, Reeke and Edelman (1988), who are neuroscientists, propose a Neuronal Group Selection theory, in which the brain actively organizes data. Nevertheless, both conventional AI and connectionism ignore the possibility that the mind develops its own data structures; they accept that the world itself is organized into distinct categories, and that an intelligent program or connectionist network should utilize ready-made data structures embodying these categories. Furthermore, both connectionism and AI lack a unified position on whether the data structure should reflect the nature of the world or the nature of the task at hand. Taken together, these form an axis of continuity between the two fields.

ii) The role of modelers and programmers: Both connectionism and AI have shelved the question of what a model means. In the case of connectionism, it is unclear who is doing the thinking, the model or the person observing the model. An analogous debate concerning the role of the human observer of conventional AI programs has long existed, as can be seen in papers by Turing (1950), Searle (1981) and Hofstadter (1981). This is not to imply that the serial computer and connectionist network are not minds in the sense that the human brain is a mind--although that is certainly one conclusion that might be drawn. Nor do connectionists or conventional AI researchers (uniformly) ignore the issue. Rather, it is significant that both camps have accepted the shelving of the problem as acceptable, at least at this time.

For example, it was pointed out above that in the Rumelhart and McClelland model of verb-tense acquisition, the human observer performs the sequential arrangement of symbols into a past-tense form of word; the model merely offers a set of simultaneously stimulated nodes. Given the 'rules' of how to interpret Wickelfeatures, one can organize a set of units unambiguously into a word. Nevertheless, the human and not the model performs this arrangement.

A related but perhaps more important question asks whether the nodes actually mean anything at all. Two physically identical networks may contain entirely different

sorts of information. A given configuration symbolizes information about verb-tenses or relationships between family members; a given node may stand for the word 'went,' or for John's father. One cannot distinguish on physical grounds two networks with entirely different (intended) meanings. To say that a network gives us the answer to questions such as 'what is the past-tense of 'come'?' or 'who is the brother of Mary?' may be misleading, because nothing in the network contains that information. Such labels are supplied by the human observers of the network.

Similar issues arise when considering AI programs. Consider the expert system **Prospector**, described in Chapter One. At the simplest level, the program can be described as turning on lights in a computer screen, which the user interprets as words symbolizing objects in the world. The expert system would function just as efficiently were nonsense words substituted for the names of minerals; if the user knew the correlation between nonsense words and rocks, the outcome would be the same. The expert system manipulates symbols, which are interpreted by the external user as meaning something. In the cases both of PROSPECTOR, a conventional AI program, and Rumelhart and McClelland's connectionist verb-tense network, a human user interprets a collection of symbols as meaning something. It is at least a reasonable question whether the human user does something

done by neither the expert system nor the connectionist network.

A possible answer to the charge that the human observer of a connectionist network does a substantial portion of the thinking hypothesizes division of cognition into levels. Thus, the connectionist network would depict only a low level of cognition, with the 'missing' interpretation of the network occurring at a higher level, closer to consciousness. In fact, the Parallel Distributed Processing collection uses this defense, but with limited success.

The PDP set is subtitled 'Explorations in the Microstructure of Cognition.' Early on in Volume 1, McClelland, Rumelhart and Hinton (1986) write that:

> Parallel distributed processing models offer alternatives to serial models of the microstructure of cognition. They do not deny that there is a macrostructure of cognition, just as the study of subatomic particles does not deny the existence of interactions between atoms. What PDP models do is describe the internal structure of the larger units, just as subatomic physics describes the internal structure of the atoms that form the constituents of larger units of chemical structure.

If we accept this passage as significant, we should interpret the verb-tense acquisition model as an illustration of the hardware of conjugation, rather than a mind that actually conjugates. There might be a 'higher level' of cognition, a macrostructure, doing the interpretation that the network does not do.

This hypothetical explanation fails to satisfy for two reasons. First, the PDP collection contains many apparent speculations on how connectionist networks might explain macro phenomena. For instance, Rumelhart, McClelland, Smolensky and Hinton (1986) describe a hypothetical network in which the individual nodes correspond to "microfeatures" such as televisions and ceilings in a room. The authors write that they do not describe an actual network, but merely offer a hypothetical example, and they point out that in a more realistic example the concept of 'television' would be distributed over an agglomeration of units. Still, this chapter, and indeed the entire collection, lacks a convincing definition of a "microfeature," nor are the limits on what a single node or collection of nodes may symbolize made clear. As Holyoak (1987) points out, murkiness concerning what interpretations may be ascribed to a node, and the frequent use of examples drawn from macro phenomena, at best weaken (Rumelhart and McClelland's version of) the connectionist agenda, and at worst simply mislead the reader.

Even were we to take the passage quoted above at face value, it still indicates continuity between connectionism and AI, as it implies that problems such as interpretation of the hardware will be solved at the level of macrostructure. As Rumelhart and McClelland take no position on whether these macro issues will be solved in a connectionist or a serial manner, the passage amounts to begging the question.

Whether or not we accept the caveat that the studies in the PDP collection concern only the microstructure of cognition as an answer to the problem of interpreting the model, the approach taken to this issue remains a continuity between connectionism and conventional AI. If we consider the caveat unsubstantiated by the studies in the collection, the situation has not appreciably changed. We can say simply that the connectionist networks described do not satisfactorily allay uneasiness over how the interpretation problem will be resolved; in this case, the continuity is the acceptance of the postponement. If, on the other hand, we do accept the caveat, a similarity of approach exists in that both fields assume that the problem can eventually be solved, perhaps at some higher level. In each case, there does seem to be continuity in how conventional AI researchers and connectionists treat the issue of interpretation of models.

II. Continuity and Critical Distance

I have outlined two axes along which I see continuity between conventional AI and connectionism: the reliance on data structures to manage difficulties that may, in fact, arise from either the nature of the world or the nature of the technology; and the postponement of the issue of how to interpret a model. These points are, I believe, grounds on which both conventional AI and connectionism may be critiqued--not damned, but certainly questioned.

In my account, the presence of continuity does not, in fact, damn connectionism. Rather, I have suggested conceptual links between connectionism and conventional AI. My attempt to show continuity between these two fields differs from similar attempts, such as those of Boden (1988) and Pinker and Prince (1988), as those accounts consider continuity as grounds for rejecting connectionism. While I doubt the validity of connectionism as a model of cognition, I do not use continuity as an <u>argument</u> to establish invalidity, nor do I assert that if connectionism is (in some ways) similar to conventional AI, that it is somehow subsumed by the older field and therefore valueless. The axes of continuities outlined in Section I of this chapter show that continuity may be phrased so that it does not constitute an argument for invalidity.

My analysis of the gambits used in the connectionist debate suggests that perception of continuity plays an

important part in insiders' assessments of connectionism. Most obviously, what I have called the historical gambit argues that because conceptual similarities exist between connectionism and various established disciplines, connectionism lacks its own identity and is therefore not a valid research endeavor. The historical gambits used by Dietrich and Fields (1988), Dellarosa (1988) and Cowan and Sharp (1988) amount to a misuse of historical argument. In particular, there seems to be an unjustified link in insiders' accounts between discontinuity and validity. The correctness of this equation is not obvious to an outsider to the debate, and it raises the question of why the insiders should assume that the validity of connectionism depends on historical discontinuities.

The mere-technology gambit raises an identical question. While Lehnert (1988), Pinker and Prince (1988) and Fodor and Pylyshyn (1988) justify their use of the mere-technology gambit by appealing to norms of good science, the gambit rests on an appeal to continuity. By charging that connectionist architecture 'reduces' at some level to serial architecture, critics of connectionism imply that it merely continues previous research in AI and cognitive psychology. Again, the actors' accounts use continuity as an argument to undermine the validity of a new research field.

I have argued in Chapter Two that the third of the strategies, the correspondence gambit, attempts to establish

discontinuity between connectionism and the literature of the author's home discipline. As I have shown, however, the correspondence gambit appears only in responses from those disciplines relatively unthreatened by connectionist claims. Where establishing continuity is a required to protect the disciplinary identity of the author (i.e., in the case of responses from AI supporters), the mere-technology gambit replaces the correspondence gambit, allowing the author to deny any possibility of discontinuity.

Use of these gambits indicates an assumption by insiders that validity depends on independent identity, and that independent identity depends in turn on revolutionary discontinuity. A chief difference between insiders' accounts and my account is that while the insiders assess the validity of connectionism in terms of continuity, I link the credibility of connectionism to continuity. To be credible, a field must have an independent identity accepted by insiders, and insiders apparently define the identity of connectionism in terms of its ability to offer a revolutionary alternative to conventional AI. Credibility is thus a measure only of whether the field is accepted, not whether it ought to be accepted. The question that remains is why continuity and discontinuity so pervasively appear in insiders' accounts as arguments for or against the validity of connectionism.

III. Continuity as a Pervasive Issue in Insiders' Accounts

One might explain the use of continuity as an argument against validity by pointing to the distorted historical sense of the actors. Computer scientists, biologists and psychologists are not historians or sociologists, and they interpret the significance of the connections between events differently than those who make a specific study of such connections. This answer, while it may perhaps be true, is weaker than one would like; it reduces to a blunt assertion that we, the outsiders, see the situation better than the insiders. In some instances, only this contentious justification may be available, but in this case a more satisfying (hypothetical) answer exists.

No deep explanation is needed to explain the importance of an independent identity in gaining scientific credibility. General acceptance of the assertion that connectionism can model cognition in a useful way would lead to research funds, publicity, faculty positions, and the other necessities of capital-intensive research. It is to be expected that connectionists claim that their networks do--or could do--something new, or something important in a new way.

By the same token, it is to be expected that this would provoke a reaction in the conventional AI community. As has been pointed out, it would be a mistake to accept the claims of the AI community that their discipline is accepted

and secure--even though connectionists often present AI as the established group for the purposes of attacking them. In fact, AI is subject to frequent attacks on its own identity, as the remarks of Tienson (1987) and Putnam (1988) illustrate. When connectionists claim to provide an alternative to AI, they attack a field whose status is in question. The terms of the connectionism debate imply that if one accepts connectionist claims to validity, one also has to accept accusations that AI has failed. Connectionism and AI would apparently compete for the same niche--the same funds, the same faculty positions, the same prestige. This in itself explains much of the mood surrounding connectionism.

While illuminating, this crude sociological explanation does not explain why continuity and discontinuity should be at the heart of the debate. Why should arguments be grounded in continuity? Why should the insiders' conception of the relationship between AI and connectionism stress an exclusive relationship? One possible answer is that in this case, science itself has been influenced by the philosophy of science.

Established philosophies of science have stressed gradual theory change; such perspectives would not take arguments from continuity as necessarily condemnatory of a new research field. The past thirty years, however, have seen the emergence of various theses asserting discontinuity

in science, most notably Thomas Kuhn's (1962) Structure of Scientific Revolutions. While Kuhn's original model has been refined by subsequent science studies research, including Kuhn's later work, Kuhnian notions in crude form have seeped into science itself, causing some scientists to conceptualize themselves and their work in terms of Kuhn's 'normal science,' 'anomalies' and 'revolutionary paradigms.'

In the case of connectionism, Kuhn's terminology is immediately apparent. Smolensky (1986) talks of connectionism as offering an alternative to the "symbolic paradigm," namely the "subsymbolic paradigm." Tienson (1987) asserts a "Kuhnian crisis" in conventional AI. Holyoak, in his (1987) review of Parallel Distributed Processing, refers slightingly to the "proselytizing bent" of connectionism and notes that "talk of a Kuhnian 'paradigm shift' is in the air." Greenco (1987) refers to a "cognitive counter-revolution." Dietrich and Fields (1988) deny that connectionism is a "revolution," while Schneider (1987) asks "Connectionism: Is it a Paradigm Shift for Psychology?"

More important than the appearance of terminology, however, is the extent to which Kuhn's model of scientific change has been internalized by the actors involved. Actors seemingly accept that for connectionism to be valid, it must be revolutionary, and overturn what has come before. The concept of revolutionary discontinuity as a basis for

scientific change seems to be implicitly endorsed even by those actors who make no reference to Kuhn or his notorious model. Supporters of connectionism cast it as revision, revolution or replacement; critics consider arguments purporting to show continuity sufficient to undermine connectionism as a whole.

I suggest this explanation for the pervasive presence of continuity in insiders' debates over connectionism as an interesting hypothesis. There is no direct evidence that the insiders take their view of scientific change from Kuhn's model. Nevertheless, I would not suggest such an explanation if I did not think it plausible. The prominent role of philosophers in the connectionism debate provides circumstantial evidence for the hypothesis; one would expect less of an influence from philosophy of science in a debate over, say, chemistry, in which philosophers were uninvolved. Moreover, Kuhn's model is now, explicitly and implicitly, in various corrupt interpretations, part of general academic parlance. It is not necessary to posit Kuhn's model as the only causal element in the explanation; quite probably, scientists adopted this notion of scientific change because it resonated with existing cognitive and social interests. I suggest, therefore, that the actors involved in the connectionism debate conceptualize themselves and their work in terms of a tacit model of science, and that their claims have a sociological, as well as a cognitive dimension. That

is, the content of cognitive claims is in this episode influenced by acceptance of a philosophical account of scientific change.

CONCLUSION
Significance and Future Research

In this thesis, I have shown that actors' accounts of a new scientific field, connectionism, take continuity as an important issue in assessing that field. In one sense, I have sided with the doubters; I have argued for a certain kind of continuity, and I have expressed uneasiness over the validity of connectionism. However, I have argued that the actors have misinterpreted the nature of continuity. While I have presented two similarities between the thinking of the new field and the old field, and acknowledged that the status of connectionism seems linked to perceptions of its continuity or discontinuity in terms of better established fields, I have argued that validity and continuity are distinct issues; my version of continuity is separate from my evaluation of the connectionist program. Finally, I have presented a tentative hypothesis concerning the central presence of continuity in the debate concerning connectionism's validity; I have suggested the actors involved base their interpretation of scientific change on a crude version of a philosophical model that has seeped into common scholarly use.

My account points to a variety of more extensive projects, some combination of which could lead to a dissertation. First, I have not investigated the possibility that there are different strands of connectionism. While I do not believe that this assumption

affects my arguments, I have noticed some fragmentation of connectionism. For example, I presented Daniel Touretzky as an unfriendly critic of connectionism on the basis of his response to Paul Smolensky. However, Touretzky's citations indicate that he himself is a 'connectionist;' he may merely be an unfriendly critic of Smolensky. Future research would first have to investigate the social and intellectual fine structure of connectionism.

Second, this analysis relies solely on published books and articles. Future research would also have to extend the source base of the present study, perhaps contrasting the rhetorical strategies found in published accounts with those used in research proposals, interviews, and professional conferences.

Third, my hypothetical explanation for the presence of continuity in the connectionism debate links the cognitive and social aspects of connectionism. I imply that the cognitive and social realms are not causally separate; connectionists adopt certain cognitive positions <u>because</u> of their involvement in a struggle for credibility. This crude version of the 'social construction of knowledge' thesis asserts only a general influence of social environment on cognitive claims. Future research might ascertain whether specific cognitive positions can be attributed to environmental influences.

Fourth, future research might investigate the influence of philosophy of science on science. Such a work might also examine the extent to which philosophical models, such as that of Kuhn or Lakatos, have mistaken rhetorical strategies for cognitive criteria of judgment through an uncritical attitude towards scientist's accounts.

Beyond the illumination that my account throws on its subject, it has significance for the content, status and use of Science and Technology Studies (STS) research. Beginning with issues of the narrowest significance, my research raises questions about the effect of analytic accounts of science. It has usually been assumed that any impact from the philosophy of science on science itself will come from philosophers' explicit normative claims--suggestions that science ought to proceed in a certain way. My analysis suggests that the philosophy of science may, in fact, already have affected science, but through a purely descriptive model. A re-evaluation of the social role of the philosophy of science may be called for.

My account also has policy implications. Insiders' accounts of the merits and weaknesses of connectionism will be used to allocate the material resources of scientific research--for example, funding, laboratories and graduate students. Policy-makers recognize that insiders have material interests in the outcome of controversies over new fields, and customarily take this into account when

evaluating scientists' statements of the value of their research. But I have suggested another reason for questioning actors' accounts, namely their underlying view of scientific change. The insight that scientists' cognitive claims involve assumptions about the nature of science indicates that theory-based STS research can be made relevant to the non-academic world, and that STS merits status of an independent research discipline.

More important, my analysis exemplifies what I perceive as the best way to win credibility for STS. While analyses of past episodes form an important part of Science and Technology Studies, I believe that acceptance of STS as an independent research discipline will depend on its ability to unpack contemporary episodes in a useful way (rather than its ability to replace existing disciplines). It is presently an open question whether STS can produce accounts of science that are simultaneously relevant to the interests of--and comprehensible to--the many groups affected by science, including the general public, policy-makers, scientists themselves and those scholars that study science. Yet it seems worthwhile to try to produce such multi-stranded accounts; to the extent that I have done so, my analysis of connectionism will reflect the value of the STS approach.

Finally, my account calls into question some common assumptions about 'rhetoric' in science, and provides a key

to using actors' accounts as indicators of cognitive positions. The discourse I have examined attempts to persuade the audience; it can therefore be classified as rhetoric. However, to argue that such accounts are irrelevant to the 'real' cognitive activity of science because they are motivated only by the 'social' desire to win support mistakes the nature of actors' accounts. I suggest that the structures to which actors appeal to convince their audiences inform their own cognitive positions. We can use actors' account to discover what assumptions the actors share with their audience; in many cases, these assumptions inform the actors' own cognition. My research indicates that we can use discourse--without taking it at face value--to discover cognitive activity. Taken as a whole, my thesis suggests that the social and cognitive aspects of science are complexly bound, and in doing so addresses a central concern of Science and Technology Studies.

REFERENCES

Ashby W. (1952). Design for a Brain. New York: Chapman and Hall.

Barker P. (1989). The Reflexivity Problem in the Psychology of Science. In Gholson B., et. al. (eds.), Psychology of Science: Contributions to Metascience. Cambridge, England: Cambridge University Press, forthcoming.

Barr A. and Feigenbaum E. (1982). Handbook of Artificial Intelligence. Stanford, Calif.: HuerisTech Press.

Berko J. (1958). The Child's Learning of English Morphology. Word, 14, 150-177.

Boden M. (1977). Artificial Intelligence and Natural Man. New York: Basic Books.

_____. (1988). Computer Models of Mind. Cambridge, England: Cambridge University Press.

Bybee J. and Slobin D. (1982). Rules and Schemas in the Development and Use of the English Past Tense. Language, 58, 265-289.

Churchland P. (1986). Neurophilosophy. Cambridge, Mass.: The MIT Press.

Cowan J. and Sharp D. (1988). Neural Nets and Artificial Intelligence. Daedelus, Winter 1988, 85-122.

Dellarosa D. (1988). The Psychological Appeal of Connectionism. Behavioral and Brain Sciences, 11, 28-29.

Dennett D. (1988). When Philosophers Encounter Artificial Intelligence. Daedelus, Winter 1988, 283-294.

Dietrich E. and Fields C. (1988). Some Assumptions Underlying Smolensky's Treatment of Connectionism. Brain and Behavioral Sciences, 11, 29-31.

Dreyfus H. (1972). What Computers Can't Do: The Limits of Artificial Intelligence. New York: Harper Colophon Books.

Dreyfus H. and Dreyfus S. (1986). Mind Over Machine: The Power of Intuition and Human Expertise in the Computer Age. New York: The Free Press.

_____. (1988a). Making a Mind Versus Modelling the Brain: Artificial Intelligence Back at a Branchpoint. Daedelus, Winter 1988, 15-44.

_____.(1988b). On the Proper Treatment of Smolensky. Behavioral and Brain Sciences, 11, 31-32.

Duda R., Gashnig J. and Hart P. (1982). Model Design in the PROSPECTOR Consultant System. In Weber B. and Nilsson N. (eds.), Readings in Artificial Intelligence. Palo Alto, Calif.: Tioga Publishing Co.

Ernst G. and Newell A. (1969). GPS: A Case Study in Generality and Problem Solving. New York: Academic Press.

Egan M. (1987). Commentary: Settling Into a New Paradigm. Southern Journal of Philosophy, XXVI, Supplement, 115-118.

Fodor J. and Pylyshyn Z. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. Cognition, 28, 3-71.

Freidin R. (1988). Connectionism and the Study of Language. Behavioral and Brain Sciences, 11, 34-35.

Gellsema E. and Kanal L. (eds.) (1986). Pattern Recognition in Practice II. Amsterdam: Elsevier Press.

Graubard S. (1988). Preface to the Issue 'Artificial Intelligence.' Daedelus, Winter 1988, V-VIII.

Greenco J. (1987). The Cognition Connection. The New York Times Book Review, January 4, 1987.

Hebb D. (1949). The Organization of Behavior. New York: John Wiley.

Hinton G. (1981). Implementing Semantic Networks in Parallel Hardware. In Hinton G. and Anderson J. (eds.), Parallel Models of Associative Memory (pp.161-188). Hillsdale, NJ: Erlbaum.

_____. (1988). Representing Part-Whole Hierarchies in Connectionist Networks. Technical Report CRG-TR-88-2. (To appear also in Proceedings of the Tenth Annual Conference of the Cognitive Society, Montreal, August 1988. Hillsdale, NJ: Erlbaum.)

Hinton G., McClelland J. and Rumelhart D. (1986). Distributed Representations. In Rumelhart D., McClelland J. and the PDP Research Group (eds.), Parallel Distributed Processing, Volume 1 (pp.77-110). Cambridge, Mass.: The MIT Press.

Hofstadter D. (1981). Reflections. In Hofstadter D. and Dennett D. (eds.), The Mind's I (pp.373-83).

Holyoak K. (1987). A Connectionist View of Cognition. Science, 236, 992-996.

Hopfield J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proceedings of the National Academy of Sciences, USA, 79, 2554-2558.

Horgan T. and Tienson J. (1987). Settling into a New Paradigm. Southern Journal of Philosophy, XXVI, Supplement, 97-114.

Horgan T. (1987). Preface. Southern Journal of Philosophy, XXIV, Supplement, 1.

Hunt E. (1975). Artificial Intelligence. Orlando, Fla.: Academic Press.

Hunter L. (1988). Some Memory, but No Mind. Brain and Behavioral Sciences, 11, 37-38.

Hurlburt A. and Poggio T. (1988). Making Machines (and Artificial Intelligence) See. Daedelus, Winter 1988, 213-240.

Jones W. and Hoskins J. (1987). Back Propagation; A Generalized Delta Learning Rule. Byte, October 1987, 183-192.

Josin G. (1987). Neural Network Heuristics. Byte, October 1987, 183-192.

Kuczaj S. (1977). The Acquisition of Regular and Irregular Past Tense Verb Forms. Journal of Verbal Learning and Verbal Behavior, 16, 589-600.

Kuhn T. (1962). The Structure of Scientific Revolutions. Chicago: The University of Chicago Press.

Lakatos I. (1970). Falsification and the Methodology of Scientific Research Programs. In Lakatos I. and Musgrave A. (eds.), Criticism and the Growth of Knowledge. Cambridge, England: Cambridge University Press.

Lakatos I. and Zahar E. (1976). Why Did Copernicus' Research Programme Supercede Ptolemy's? In Westman R. (ed.), The Copernican Achievment. Los Angeles: University of California Press.

Larsen E. (1986). Neural Chips. Omni, November 1986, 113-116, 168-9.

Lehnert W. (1988). Physics, Cognition and Connectionism: An Interdisciplinary Alchemy. Brain and Behavioral Sciences, 11, 40-41.

Lenat D., Prakash M. and Shephard M. (1986). CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks. The AI Magazine, Winter 1986, 65-85.

Maratsos M. (1988). Problems of Connectionism. Science, 242, 1316-1317.

McCarthy J. (1988). Mathematical Logic in Artificial Intelligence. Daedelus, Winter 1988, 297-311.

McClelland J. and Kawamoto A. (1986). Mechanisms of Sentence Processing: Assigning Roles to Constituents. In McClelland J., Rumelhart D. and the PDP Research Group (eds.), Parallel Distributed Processing, Volume 2 (pp. 272-326). Cambridge, Mass.: The MIT Press.

McClelland J. and Rumelhart D. (1988). Explorations in Parallel Distributed Processing. Cambridge, Mass.: The MIT Press.

McClelland J., Rumelhart D. and Hinton G. (1986). The Appeal of Parallel Distributed Processing. In Rumelhart D., McClelland J. and the PDP Research Group (eds.), Parallel Distributed Processing, Volume 1 (pp.3-44). Cambridge, Mass.: The MIT Press.

McClelland J., Rumelhart D. and the PDP Research Group (eds.) (1986). Parallel Distributed Processing, Volume 2. Cambridge, Mass.: The MIT Press.

McCorduck P. (1979). Machines Who Think. San Francisco: W.H. Freeman and Co.

McCulloch W. and Pitts W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115-133.

Minsky M. and Papert S. (1969). Perceptrons. Cambridge, Mass.: The MIT Press.

Newell A., Shaw C. and Simon H. (1963). Chess-Playing Programs and the Problem of Complexity. In Fegenbaum E. and J. Feldman (eds.), Computers and Thought (p.47). New York: McGraw-Hill.

Papert S. (1988). One AI or Many? Daedelus, Winter 1988, 1-14.

Pinker S. (1984). Language Learnability and Language Development. Cambridge, Mass.: Harvard University Press.

Pinker S. and Mehler J. (eds.) (1988). Connections and Symbols. Reprinted essays from Cognition, 28. Cambridge, Mass.: The MIT Press.

Pinker S. and Prince A. (1988). On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language. Cognition, 28, 73-184.

Port O. (1986). Computers That Come Awfully Close to Thinking. Business Week, June 2, 1986, 92-96.

Putnam H. (1988). Much Ado About Not Very Much. Daedelus, Winter 1988, 269-282.

Reeke G. and Edelman G. (1988). Real Brains and Artificial Intelligence. Daedelus, Winter 1988, 143-174.

Rumelhart D. (1980). Schemata: The Building Blocks of Cognition. In Spiro R., Bruce B. and Brewer W. (eds.), Theoretical Issues in Reading Comprehension (pp.33-58). Hillsdale, NJ: Erlbaum.

Rumelhart D., Hinton G. and McClelland J. (1986). A General Framework for Parallel Distributed Processing. In Rumelhart D., McClelland J. and the PDP Research Group (eds.), Parallel Distributed Processing, Volume 1 (pp.45-76). Cambridge, Mass.: The MIT Press.

Rumelhart D. and McClelland J. (1986a). On Learning the Past Tenses of English Verbs. In McClelland J., Rumelhart D., and the PDP Research Group (eds.), Parallel Distributed Processing, Volume 2 (pp.216-271). Cambridge, Mass.: The MIT Press.

_____. (1986b). PDP Models and General Issues in Cognitive Science. In Rumelhart D., McClelland J. and the PDP Research Group (eds.), Parallel Distributed Processing, Volume 1 (pp.110-146). Cambridge, Mass.: The MIT Press.

Rumelhart D., McClelland J. and the PDP Research Group (eds.) (1986). Parallel Distributed Processing, Volume 1. Cambridge, Mass.: The MIT Press.

Rumelhart D., Smolensky P., McClelland J. and Hinton G. (1986). Schemata and Sequential Thought Processes in PDP Models. In McClelland J., Rumelhart D., and the PDP Research Group (eds.), Parallel Distributed Processing, Volume 2 (pp.7-57). Cambridge, Mass.: The MIT Press.

Schank R. (1972). Conceptual Dependency: A Theory of Natural Language Understanding. Cognition, 3, 552-631.

Schneider W. (1987). Connectionism: Is it a Paradigm Shift for Psychology? Behavior Research Methods, Instruments and Computers, 19, 73-83.

Schwartz J. (1988). The New Connectionism: Developing Relationships Between Neuroscience and Artificial Intelligence. Daedelus, Winter 1988, 123-142.

Searle J. (1981). Minds, Brains and Programs. In J. Haugeland (ed.), Philosophy, Psychology and Artificial Intelligence. Cambridge, Mass.: The MIT Press.

Segal M. (1988). Neural Network Programs. Science, 241, 1107-1108.

Serchuk A. (1987). Expert Systems: A Concurrent Internal and External Analysis. Unpublished Bachelor of Arts thesis, Vassar College. Poughkeppsie, NY.

Smolensky P. (1988). On the Proper Treatment of Connectionism. Brain and Behavioral Sciences, 11, 1-23.

_____. (1986). Information Processing in Dynamical Systems: Foundations of Harmony Theory. In Rumelhart D., McClelland J. and the PDP Research Group (eds.), Parallel Distributed Processing, Volume 1, (pp.194-281). Cambridge, Mass.: The MIT Press.

Tank D. and Hopfield J. (1987). Collective Computation in Neuronlike Circuits. Scientific American, December 1987, 104-114.

Tenebaum J., Garvey T., Weyl S. and Wolf H. (1974). An Interactive Facility for Scene Analysis Research. Stanford: Stanford Press.

Tienson J. (1987). Introduction to Connectionism. Southern Journal of Philosophy, XXVI, Supplement, 2-17.

Touretzky D. (1988). On the Proper Treatment of Thermostats. Brain and Behavioral Sciences, 11, 55-56.

Turing A. (1950). Computing Machinery and Intelligence. Mind, LIX.

Waldrop M. (1984). Natural Language Processing. Science, April 27, 1984, 372-5.

Waltz D. (1982). The State of the Art in Natural Language Processing. In W. Lehnert and M. Ringle (eds.), Strategies in Natural Language Processing. Hillsdale, NJ: Erlbaum.

_____. (1988). The Prospects for Building Truly Intelligent Machines. Daedelus, Winter 1988, 191-212.

Weizenbaum J. (1976). Computer Power and Human Reason: From Judgement to Calculation. San Francisco: W.H. Freeman and Co.

Wickelgren W. (1969). Context-sensitive Coding, Associative Memory, and Serial Order in (Speech) Behavior. Psychological Review, 76, 1-15.

Winston P. (1975). Learning Structural Definitions From Examples. In P. Winston (ed.), The Psychology of Computer Vision. New York: McGraw Hill.

Winston P., Binford T., Katz B. and Lowry M. (1984). Learning Physical Descriptions From Function Definitions, Examples and Precedents. In Brady M. and Paul R. (eds.), Robotics Research: The First International Symposium. Cambridge, Mass.: The MIT Press.

Zeidenberg M. (1987). Modelling the Brain. Byte, December, 237-243.

The two page vita has been removed from the scanned document. Page 1 of 2

The two page vita has been removed from the scanned document. Page 2 of 2