INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Connectionism and the Integration of Error:
Applications in Naturalized Epistemology and Minimal Rationality

Brenda Roberts

A Thesis

in the

Special Individualized Program

Presented in Partial Fulfilment of the Requirements
for the Degree of Master of Arts at
Concordia University
Montreal, Quebec, Canada

November 1999

ABSTRACT

Connectionism and the Integration of Error:
Applications in Naturalized Epistemology and Minimal Rationality

Brenda Roberts

Traditional epistemology has it that the pursuit of knowledge is predicated on two inter-connected goals: the generation of meaningful truths and the avoidance of error. This is neatly summarized in the conventional definition of knowledge as justified true belief. In the following thesis I trace the evolution of an alternative account of knowledge which is predicated not on the avoidance of error but on the capacity to learn from error. I contend that the connectionist model of artificial intelligence provides the necessary framework for an understanding of cognition in which knowledge emerges as a dynamic product of learning. Epistemic content in this alternative is not comprised of fixed representations. Instead, content is encoded within shifting patterns of activation among large numbers of processing units. However, a connectionist approach does not give rise to a new epistemology. Rather, when integrated with Quinean naturalism, it fulfills the project of naturalized epistemology in ways that psychology could not. The convergence of connectionism and naturalized epistemology then embodies the normative principle that we should be fallibilistic about beliefs and realistic about believers.

TABLE OF CONTENTS

If a machine is expected to be infallible, it cannot also be intelligent.
Alan Turing (1946) [1]

1. Introduction: Knowledge and Error

Epistemology has it that the pursuit of knowledge is predicated on two inter-connected goals: the generation of meaningful truths and the avoidance of error. This is neatly summarized in the traditional definition of knowledge as justified true belief. On a deeper level, this tripartite definition also presumes a representational form in which content is characterized as stable and discrete. In the following thesis I sugggest an alternative account of knowledge which is predicated not on the avoidance of error but on the capacity to learn from error. In this alternative, content -- that is, the objects of knowledge and belief -- is not comprised of fixed representations. For example, our belief about grandmother doesn't correspond to a 'grandmother' neuron being fired when we think of her. Instead, 'grandmother' is encoded within shifting patterns of interaction over large areas of cognitive processing space.

I contend that the integration of error is an essential element in the generation of this distributed knowledge and that the connectionist model of artificial intelligence

---

[1] The entire quote is: "In other words then, if a machine is expected to be infallible it cannot also be intelligent. There are several theorems which say almost exactly that. But these theorems say nothing about how much intelligence may be displayed if a machine makes no pretense at infallibility."

provides a useful framework for an understanding of knowledge which then emerges as a necessarily fluid product of learning. My point is that it is error, and not certainty, that underwrites knowledge.

My strategy is to first eliminate certainty as the *sine qua non* of knowledge and then re-think the origins of knowledge using the uncertainty and error which constitute the raw material of cognition. Since a feeling of certainty is neither necessary nor sufficient for knowledge, the rejection of certainty begins with a disavowal of psychological certainty. More important, a rejection of certainty means the disavowal of infallibility regarding any proposition. While this does not entail the wholesale rejection of necessary truths such as those of mathematics and logic, it leaves the question of necessity open. Simply put, a rejection of certainty means that there is no epistemically privileged person or realm: only God has a God's eye view.

Now let me explain how error underwrites knowledge. An understanding of the relationship between intelligence and cognition is essential for this move. I equate intelligence with the capacity to deal with problems both flexibly and effectively. So intelligence is practically defined in terms of the problem at hand.[2] Cognition is defined as the group of processes responsible for knowledge, including perception, memory, and

---

[2] This explains why the development of a theory of intelligence which is part of an over-all epistemology has been ignored: the traditional definition of knowledge as justified true belief insures that intelligence is labelled a performance issue and is therefore relegated to behavioural psychology. But as we shall see in the second chapter of this thesis, the separation of epistemology and psychology is revoked by naturalized epistemology. This fact, plus the advent of technological innovations such as magnetic resonance imaging and increasingly powerful computing machinery allows us to empirically verify and simulate the flow of thought in the brain. This results in a shift from the (often uneasy) alliance between philosophy and psychology which marked naturalized epistemology at its inception, to a fully naturalized epistemology comprised of an interdisciplinary matrix that includes philosophy, cognitive science and AI.

decision-making. A relationship between the two terms emerges in which intelligence may then be seen as the *capacity* to generate knowledge while cognition is the *act* of generating knowledge. Obviously, there can be no act of cognition without a corresponding and pre-existing capacity for cognition. Given that cognition is a necessary condition for knowledge, intelligence is therefore a necessary condition for *both* cognition and knowledge. If Turing's claim that infallibility precludes intelligence is correct, (Turing 1946, 124) then error is an essential element in both intelligence and cognition and it can be asserted that error underwrites knowledge. At least that is a basic claim of this thesis.

A cursory survey of epistemology supports the suggestion that the epistemic goal of certain knowledge, defined as the impossibility of error, is both unrealistic and counterproductive. Platonic epistemology stipulates that since knowledge requires certainty and certainty requires unchanging subject matter, true knowledge can only be of unchanging forms. But the external world is defined by change. So in effect, all our knowledge of nature is demoted to the realm of mere opinion. This standard of knowledge was entrenched in modern philosophy with the indubitability requirement derived from Descartes's *Cogito* and results in a predicament whereby the only things that remain immune to doubt are my existence and the current contents of my consciousness at time *t*. But this standard leaves us trapped in our own heads, unable to acquire knowledge of the external world without the auspices of God or Reason. Epistemology thus finds itself unable to bridge the gap between the mental and material realms.

With the failure of deductive logic to mount a convincing bridge between the internal and external realms implied by Cartesian dualism, the problem of induction emerges. According to Israel Scheffler, the source of the impasse is Hume's demonstration that, "Effects cannot be simply deduced from their causes, nor can predictions be logically demonstrated on the basis of available evidence garnered from past experience."(Scheffler 1967, 228) Therefore, as Hume put it, "In vain should we pretend to determine any single event, or infer any cause or effect, without the assistance of observation and experience."(Hume 1748, 601) But experience relates events within bounded space-time frames and contributes nothing certain about evidence outside that frame. Ultimately then, inductive inference also precludes certain knowledge. Instead, an element of risk permeates induction because conclusions drawn beyond the available evidence always entail the possibility of error. Error is therefore inherent in all areas of epistemic judgement and, to borrow a phrase from Quine, the Humean problem becomes the human problem.

1.1 Error and Connectionism

D.A. Norman describes the human cognitive system as marked by error on both the input and output levels. In terms of input, we are surrounded by ambiguity, incompleteness, and false information. The fact that human cognition works well in the face of epistemic uncertainty is a testament to the brain's cognitive robustness, flexibility, and adaptivity. At the same time, the system's output is riddled with errors. This is evident in the frequency of incomplete sentences, false starts, and erroneous

words in speech, as well as the lapses in deductive reasoning such as those documented by Kahneman and Tversky (1974). These characteristics of human cognition provide essential clues about the nature of our processing environment. Norman envisioned a set of machine intelligence properties, including graceful degradation of performance, content-addressable storage, and iterative retrieval, that reflect the pervasive ambiguity of the cognitive environment. As it turns out, these properties are isomorphic with the properties of parallel distributed processing (PDP)[3] architectures (Norman 1986, 537). If the common function of machine and human intelligence is the processing of partial input by finite systems, then perhaps the nature of cognition is found in the processes which human and parallel distributed processing architectures share. Furthermore, the integration of error is a hallmark of the connectionist approach to artificial intelligence since it is by back probagation of error that spontaneous generalizations which simulate learning are produced. Therefore, Turing's insight on the connection between fallibility and intelligence applies to both machine and human intelligence: whether human or machine, the cognitive engine seems designed to run on the most widely available and infinitely renewable epistemic resource of all: error.

Prima facie, the elimination of certainty in knowledge implies the impossibility of knowledge and with it the impossibility of epistemology itself. It is crucial to stress

---

3

The terms 'connectionist', 'parallel distributed processing', 'PDP', and 'neural nets' are treated synonymously unless specific contexts such as those of computer architectures, epistemology, or cognitive science are stipulated.

that a fallibilist conceptualization of knowledge does not relegate us to universal scepticism. Rather, it is only as long as certainty about things is the ideal of knowledge and its essential characteristic that knowledge remains vulnerable to scepticism. However, just as earthquake-resistant foundations are built to shift with impact, once error and uncertainty[4] are integrated into the epistemic process, scepticism is disarmed.

There is nothing new in this development. Fallibilism emerged when the dilemma of whether or not knowledge is possible was reconstituted as a trilemma, a solution first advanced by the ancient sceptic, Pyrrho. According to Pyrrhonism, *ataraxia* (freedom from aggravation) was possible by sidestepping the quest for certain knowledge in favour of the acceptance of a plausible brand of knowledge sufficient to use as a basis for action. The trilemma approach to scepticism was revived by Fries in 1807. In Fries' incarnation, knowledge may be dogmatically assumed or justified by an endless string of justifications or else anchored to a psychological description which is justificatory without itself needing justification. (Floridi 1999) Fallibilism thus creates a third position between a dogmatic insistence on certainty and outright scepticism. This essentially pragmatic reading of fallibilism means that it is not necessary for beliefs to be grounded in certainty. Instead, it is understood that a belief, for instance the Ptolemaic account of a geocentric universe, can be strongly justified based on overwhelming evidence and yet turn out to be false. In short, error can underwrite knowledge without a

---

[4] It will be noted that I use the terms 'error', 'uncertainty', and 'risk' almost interchangeably. Of course these terms are not strictly speaking equivalent. The use of them as a group is meant to contrast with an antithetical set comprised of absolute notions such as certainty, truth, and ideal rationality.

necessary free fall into scepticism. My thesis takes the deeply entrenched fallibilism of contemporary epistemology a step further by asserting that error is not only an element of epistemology but a necessary step in the process of knowing.

1.2 The Changing Forms of Knowledge

According to Simon Blackburn's definition of epistemology, the central problems in the theory of knowledge are, "...the origin of knowledge; the place of experience in generating knowledge, and the place of reason in doing so; the relationship between knowledge and certainty, and between knowledge and the impossibility of error; the possibility of universal scepticism; and the changing forms of knowledge that arise from new conceptualizations of the world." (Blackburn 1994) A basic claim of this thesis is that the parameters of the relationship between knowledge and certainty, and between knowledge and the impossibility of error, are re-defined by eliminating certainty in favour of an account of knowledge predicated on error. We must then start from scratch and address the question of the origins of knowledge. This is explored in chapter two by using the connectionist model to describe cognition on the microstructural level.

But a connectionist description of cognition in which error defines the generation of knowledge remains just that: a description. It ignores the normative aspect which gives knowledge the quality of being correct or erroneous in the first place. It is therefore necessary to situate connectionism within an existing general theory of knowledge. In chapter three I advance the convergence of connectionism and naturalized epistemology toward such an end. This convergence fulfills Quine's project of naturalized

epistemology in ways that psychology could not, plus the reciprocal containment which characterizes naturalized epistemology is extended so that normativity is seen as an *inherent* property of description.

While a rigorous investigation of error from a connectionist standpoint argues for a fundamental re-configuration of knowledge, I stop short of endorsing eliminative materialism. The elimination of certainty does not entail the elimination of knowledge. Instead, knowledge remains but is changed in form. The convergence of connectionism and naturalized epistemology provides us with the basic principle of knowledge underwritten by error: that we ought to be fallible about beliefs and realistic about believers. The fully naturalized epistemology which concludes chapter three illustrates fallibilism about beliefs. In chapter four, realism regarding believers is illustrated using Christopher Cherniak's minimal rationality model.

Cherniak's point is that a realistic approach to our cognitive abilities means that ideal rationality standards are not only inadequate but counterproductive. I concur with this view, but, as I will demonstrate, the information processing model of cognition Cherniak relies on is flawed. I argue instead that the substitution of PDP for the traditional information processing model is necessary to develop the implications of minimal rationality to their full potential. The thesis concludes with a preliminary outline of knowledge underwritten by error.

2. Error and Connectionism

A description of connectionism in contrast to traditional artificial intelligence (AI)[5] is the most expedient way to highlight the neural net model. Good-old-fashioned-AI (GOFAI), and connectionist AI will therefore be described in terms of their common origin and divergent branching. A central element of the thesis is an investigation of how error factors into the structure of intelligence and how this element impacts the traditional concept of knowledge. The primary task of connectionism is to understand the relationship between the structure and function of the brain for it is this relationship that defines intelligence. The role of error, which is the hallmark of connectionism, allows connectionist programs to learn and represents a new approach to cognition and knowledge.

2.1 The Conceptual Evolution of Connectionism

Traditional AI and connectionism are distinct models of artificial intelligence which share a powerful common premise. This is the notion, first recognized by Alan Turing in the 1940s, that a computing machine could be indefinitely complicated and yet

---

5

Traditional AI is also variously referred to as top-down AI, serial processing, and GOFAI (good-old-fashioned-AI). Because GOFAI is the shortest term, it will be the usage which predominates in this thesis.

its constituent elements would remain utterly simple. By analogy, the human brain could also be indefinitely complex but remain reducible to simple parts. The distinction between traditional AI and connectionism rests on each model's characterization of how these elements interact to produce knowledge.

Let us accept as a pre-supposition, the classic tripartite model of cognition that is built on three interrelated levels of description: a 'top' level at which epistemic states are described in psychological vocabulary such as 'beliefs'; a 'middle' level at which they are described in a formal/mathematical way apart from any informational content; and a 'bottom' level at which they are described physically. There are two distinct approaches to the question of how new knowledge is generated in a cognitive system. With a top-down strategy we begin at the level of commonsense psychological concepts and reduce them to simpler elements. A bottom-up approach goes the opposite way and begins with the simplest elements, such as neurons, then moves upward in complexity by finding ways to interconnect these units to produce large-scale phenomena like psychological states. Top-down systems are good at tasks which rely on precision and logic. Bottom-up systems, on the other hand, excel at generating approximations and generalizations using partial and ambiguous input. GOFAI normally presumes a top-down approach while connectionism embodies a bottom-up account.

It should also be noted at the outset that there is much debate surrounding the validity of a distinction between traditional AI and connectionism.[6] To many cognitive

---

6

For example, Fodor and Pylyshyn (1988) support the classical sentential view of GOFAI; the Churchlands (1989 and 1992) and Stephen Stich (1993) support a connectionist approach. Occupying a kind of middle

psychologists and philosophers, connectionism is merely a subset of GOFAI. GOFAI is the symbolic hypothesis and connectionism is merely the *sub*-symbolic variation on the same hypothesis. The distinction, according to this view, takes place on the level of scientific explanation. This is true, of course, but it is difficult to see why this entails that the microstructural level of cognition be written off as a mere subset of GOFAI. Instead, the question remains a hierarchal one of whether or not cognition is imposed by the brain in top-down fashion or whether it emerges bottom-up from the sub-symbolic level. One of the principle axioms of GOFAI is that micro-level neural activity has no direct bearing on cognition and higher levels of cognitive activity are supposedly unaffected by the stochastically defined uncertainty which characterizes brain activity at the synaptic level. The view that we can, as John Casti puts it, "...'skim off' the rules of thought from the higher level of symbol processing and semantic networking, and just ignore what's happening down below..." (Casti 1989, 299) implies a dichotomy in AI terms that is on the scale of mind-body dualism. This dichotomy is aptly illustrated when for example, Herbert Simon says that "everything of interest in cognition happens *above* the 100-millisecond level."(Casti 1989, 299-300) Simon is defining the parameters of the GOFAI research area as confined to higher level cognitive function. On the other hand, when Douglas Hofstadler responds that everything important about cognition happens *below* the 100-millisecond mark, at the synaptic level, the line between connectionism and traditional AI is clearly drawn. The bottom line is that GOFAI preserves a gap between

---

ground are AI researchers such as Marvin Minsky (1991) who advance hybrid models. Alternative accounts with strong connectionist roots include the dynamical systems approach of Timothy Van Gelder (1996) and the associationist machines of Andy Clark (1993, 1997).

the microphysical data of neuronal cognition and the higher level stuff of beliefs; connectionism renders it superfluous.

The elemental components of traditional AI are symbols. So in the GOFAI idiom, cognition, which is used here in the broadest sense to cover a range of mental processes including reasoning and memory, as well as language, sense perception and motor function, is a function of symbol manipulation. Symbols are described as enduring entities which can be stored and retrieved from memory, and transformed according to rules. They refer to external phenomena and have semantic value or meaning. Because rules define symbolic composition and manipulation, a syntactic structure is required. (Bechtel and Abrahamsen 1991, 1) In short, GOFAI can be characterized as the symbolic hypothesis of AI in which there is not only a strict distinction between the software (ie., the symbols, rules, and propositions of knowledge) and the hardware (ie., the mechanism that runs the software), but an explicit causal connection between the two entities. The hardware specifications are based on software requirements, because the essential reason for the construction of the machine is to run the program. The machine, like the brain, has no meaningful function apart from what it does. Thus the mind is retained as an explanatory entity and is defined as what the brain does.

Connectionism is distinguished from GOFAI because it extends cognition to a deeper level wherein processes are reducible to electro-chemical activity between brain cells which excite and inhibit each other. Connectionist processes have a non-linear configguration and are distributed across the entire system. The units of the connectionist model constantly impinge on each other in shifting, dynamic patterns. In connectionism,

the distinction between hardware and software is blurred; thus, any corresponding mind/brain problem is moot. As McClelland *et al* explain, "....syntactic considerations influence semantic ones and semantic ones influence syntactic ones. We cannot say that one kind of constraint is primary". (McClelland 1986, 7) This, as we will see, is reminiscent of Quine's demonstration that the questions of epistemic priority are moot.

Thus, traditional AI preserves epistemic hierarchy and implies a kind of machine intentionality in which the program takes precedence whereas connectionism places the emphasis on architecture over programs. So while GOFAI is primarily concerned with the symbols and rules of cognition, connectionism delves more deeply into the cognitive model to ask what we know about the architecture of the brain and how this architecture shapes cognitive function. In other words, connectionist accounts focus on the generation of knowledge, that is, learning, while traditional AI focuses on the manipulation of existing knowledge.

Although connectionism may be seen as an aspect of the computational account of cognition, it is more accurate to say that it is a new form of computation based upon principles that have no pre-existing computer counterpart. (Norman 1986, 534) Traditional and connectionist AI both began from the shared belief that artificial intelligence is the most effective model for the simulation, and hence the understanding, of cognition. GOFAI, however, retains the computer metaphor of the mind while connectionism has adapted the technology to fit the neural model, effectively replacing the computer metaphor with the brain metaphor on the premise that our understanding of artificial intelligence is only as good as our understanding of the brain.

2.2 The Brain Metaphor

According to David Rumelhart, the brain metaphor of artificial intelligence is:

> ....a general and abstract model of the computational architecture of brains, [in order] to develop algorithms and procedures well-suited to this architecture, to simulate these procedures and architecture on a computer, and to explore them as hypotheses about the nature of the human information-processing system...such models are neuronally inspired, and we call computation in such a system brain-style computation. (Rumelhart 1989, 134)

Connectionism takes as its fundamental processing unit an *abstract neuron.* These units communicate by sending numbers along the lines that connect them. (Rumelhart 1989, 134-5) The biological brain upon which the connectionist system is modelled is composed of about 100 billion real neurons which receive excitory and inhibitory biochemical signals from other neurons by way of some of the 10 trillion synaptic connections in the brain and the rest of the nervous system. These connections are spread out over the surface of each of the cell bodies. Excitory or inhibitory messages enter the cell body and the extended dendritic branches, the signals are summed and the resulting signal is transmitted down the axon of the neuron, where they make contact with other neurons. These cell-to-cell connections are not discrete but continuous: they can be strong, or weak, or anything in between. The global configuration, or architecture, of these 100 trillion connections is crucial because it is the stored connection strengths (weights) that constitute the knowledge of a network at any given time.

In traditional AI, knowledge is the result of a series of *sequential* cognitive

procedures in which symbols are manipulated according to syntactic rules that are understood as distinct cognitive states. In a connectionist system, knowledge is a function of the interactions between processing units. Therefore, the activation patterns of the units is what defines content in a PDP network. Connection weights are the network's store of knowledge and can be seen as the 'mapping potential' of units that participate in distributed patterns. These potential mappings coexist across the same territory and therefore must 'compete' with each other to resolve a given set of inputs. Therefore content should not be understood as a discrete commodity to be acquired, stored, or retrieved. Rather, it is isomorphic with a continual graded process: there is no single moment in space-time at which content is 'acquired' or 'lost'. Content, learning and knowing then comprise contnual dynamic that is a function of intelligence.

2.3 Historical Evolution of Connectionism

The connectionist metaphor of the brain emerged in two steps. The foundation of the first neuronal model was laid by Warren McCullogh and Walter Pitts in 1943 with the publication of a paper in which they proposed a simple model of neuron-like computational units. They demonstrated how these binary (0/1; on/off) units could perform logical operations based on excitory and inhibitory inputs similar to synaptic activity between neurons in the brain and central nervous system. Specific configurations of these binary units could be assembled to perform the simple logical functions of, for example, Boolean operators. In 1958, Frank Rosenblatt developed a model which used binary units in layered networks. One set of units received inputs from outside the system

and sent excitations and inhibitions to another set of units which in turn sent inputs on to a third group. (Bechtel and Abrahamsen 1991, 4). Rosenblatt refined the McCollugh-Pitts model by making the strengths (weights) of the connections continuous rather than discrete. The valence of the connections could then be expressed in values *between* 0 and 1 rather than in the discrete values of 0 *or* 1. Furthermore, Rosenblatt introduced a method for altering the weights which allowed the networks to change their outputs autonomously. He called these networks 'perceptrons', and named the process by which perceptrons learn, the *perceptron convergence theorem*. It states that, "...if a set of weights existed that would procure the correct responses to a set of patterns, then through a finite number of repetitions of this training procedure, the network would in fact learn to respond correctly." (Bechtel and Abrahamsen 1991, 5)

The essential element of Rosenblatt's work is the replacement of discrete logical operators with continuous stochastic functions. Rosenblatt premised this move on the view that, "The future of information processing systems which operate on statistical, rather than logical principles seems clearly indicated."(Bechtel and Abrahamsen 1991, 6). A crucial result of the development of the perceptron was the ability to model pattern recognition on a trial and error basis. This left the way clear to build models which could recognize patterns based on partial information, if given the appropriate training. The question then was not whether or not connection strength could be spontaneously adapted within the system, in effect Rosenblatt had already demonstrated it could. The question was, how? Rosenblatt had used synapses as the neuronal model for connections in his perceptrons. Starting from the basis that it is not sufficient that a synapse could be

strengthened merely on the basis of activation, Donald Hebb wanted a mechanism that worked only when *two* activities were associated. Therefore, in 1949, Hebb posited what became known as the Hebbian mechanism: "When an axon of cell A is near enough to excite cell B and repeatedly and persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased." (Crick 1994, 183)

When trained, the Rosenblatt-Hebbian hybrids simulate information processing in a manner that is analogous to neural networks. Information is stored and defined by *connections* between the units of the network used to model synapses. The variable strengths, or weights, of these connections act as mediators between the storage and retrieval functions of systems by producing the proper activation patterns for a given input. The Hebbian mechanism simulates learning by showing that connections can spontaneously adapt to changing activation patterns within the information processing environment. Simply put, Hebb had shown that machines can 'learn'.

The publication of *Perceptrons* in 1969 by Marvin Minsky and Seymour Papert brought this work to a halt. The objective in *Perceptrons* was to study both the potential and the limitations of neural network models. One of the main limitations Minsky and Papert focused on was the inability of layered[7] neuronal networks, such as those devised by Rosenblatt, to evaluate logic problems on the order of the "exclusive OR problem"

7

It is important to note ***what*** is being layered. Some AI researchers refer to layers in terms of layers of *connections*; others to layers of ***units***. In order to avoid unnecessary complication, layers will herein refer to layers of units. So, for example, a multi-layer network contains an input layer, an output layer, and a layer of hidden units.

(XOR). XOR states that [*a* in an exclusive OR relationship to *b*] is equivalent to [(*a* and not-*b*) OR (*b* and not-*a*)]. In order for a network to compute this expression it is necessary to include hidden units between the input and output layers. Minsky and Papert's critique was widely misunderstood as destroying the credibility of perceptrons, but in fact, "...the work had simply (but elegantly) showed limitations on the power of the most limited class perceptron-like mechanisms." (Norman 1986, 535). Furthermore, it is important to note that at the time of the Minsky/Papert study there were no training procedures for multi-layered networks. Therefore programs dedicated to these tasks could not be envisioned as anything less than monumental in scale. On the basis of these connectionist scale problems, Minsky and Papert suggested that complex multi-level network models would prove sterile. It seemed that neural net architectures had reached a dead end due to limited computational resources. This resulted in a catastrophic loss of research funding for connectionist-based research in AI. (Boden 1991, 10)

In the early 80s, a refinement in the understanding of the Rosenblatt-Hebbian model was devised which triggered the implementation of learning rules used to determine synaptic weights based on variations to Hebb's rule for synaptic plasticity. (Koch 1997, 207). This allowed neural nets to go beyond the input/output functions of standard information processing models and changed the definition of knowledge production itself. As McClelland *et al* put it, "The representation of knowledge is set up in such a way that the knowledge necessarily influences the course of the processing. Using knowledge in processing is no longer a matter of finding the relevant information in memory and bringing it to bear, it is part and parcel of the processing itself."

(McClelland 1986, 32). Furthermore, the fact that memory is stored in synaptic weights meant that changing connection strengths "...can make the network converge to a different equilibrium to recalling a different memory...This allows the network to form associations over time, enabling it to learn sequences and to predict events." (Koch 1997, 207-20). Taken together, these innovations set the stage for the ascendancy of connectionism over traditional AI.

As we examine the current state of the connectionist program in greater detail in the next section, a discription of connectionist architecture and the process by which error is integrated into machine learning will be developed.

2.4 The Role of Error

The observation has often been made that while GOFAI projects are successful in modelling intellectual problems that many humans find difficult, such as those found in chess and symbolic logic, it has had far less success with human activities we find the easiest, such as recognizing a friend and retrieving his or her name. The distinction between the two sets of activities is that intellectually challenging activities involve the manipulation of discrete entities using formal rules of inference while activities we take as second nature require the ability to process partial input using *ad hoc* rules of thumb which barely reach the status of inductive inference.

We saw in the introduction that the pursuit of philosophical certainty may well be a red herring. Similarly, the use of deductive formalisms in AI has led to a focus on universal validity to the exclusion of practicality; as though nothing would do except

certainty. But in actuality the demand for inferential perfection is counter-productive when there is no guarantee that our premises are infallible in the first place. Therefore, as Marvin Minsky says, "we cannot ask our problem-solving systems to be absolutely perfect or even consistent." (Minsky 1991, 658)

The capacity to develop the commonsense knowledge required to transact the activities of daily life is predicated on the brain's ability to process information rife with ambiguity and inconsistency. This means that error is the only constant in human cognition. Similarly, the computing environment of connectionism functions on the back propagation of error. In fact, error is an essential element of learning in PDP systems. As we shall see in more detail later, neural networks learn by repeatedly adjusting the weights between units until an output with the minimal error value is reached. In order to train up the system to recognise patterns given partial or conflicting input, a circumstance which is the norm of cognition in reality, both a target vector and an acceptable error threshold are required. The difference between the two values then determines the activation state (or output) of each unit. This stochastic function allows implementation of an error term for each iteration until the weights between the units have adjusted themselves to yield an output which reaches the effective minimum.[8] It is understood that certainty is not the target. Rather, the closest plausible approximation fulfills the cognitive task at hand. Weights are altered spontaneously in the course of processing by

---

8

This is, of course, a simplified account. A full description of the generalized delta rule used to generate autonomous learning in multi-layer PDP architectures, including details regarding error surfaces in connectionist weight space, may be found in Rumelhart et al 1986b, 318-362.

this back propagation of error, a trial-and-error approach to knowledge that illustrates the integral role of error in cognition.

2.5 Advantages of Connectionism

As a model of cognitive function, connectionism holds a series of advantages over models based exclusively on sequential processing. (Bates and Elman 1993) These advantages effectively redefine knowledge as a dynamic, emergent and evolving property of neural states. As Rumelhart *et al* point out, "...almost all knowledge is *implicit* in the structure of the device that carries out the task rather than *explicit* in the states of units themselves. Knowledge is not directly accessible to interpretation by some separate processor, but is built into the processor itself and directly determines the course of processing. It is acquired through tuning of connections as these are used in processing, rather than formulated and stored as declarative facts." (Rumelhart *et al* 1986a, 75-76. Emphasis in the original). This implicit knowledge cannot be archived. Thus connectionists eschew representations which consist of symbolic atoms combined according to rules to create symbolic expressions. Instead, connectionism exploits activation patterns in geometric space. This knowledge can only be characterized on the microstructural level as a state of flux. But as we shall see, the flexibility and open-endedness of distributed representations has advantages in terms of the speed of processing, the nature of representations, and the mechanisms of learning.

2.5.1 Speed of Processing

Neurons operate on a time scale of milliseconds whereas computer components

operate within nanoseconds. The brain processes that machine intelligence simulates are incredibly complex, so efficiency dictates that programming algorithms must involve considerable parallelism. (Rumelhart 1989, 135) The brain deploys synaptic connections co-operatively and in parallel to make up in volume of processes what it lacks in speed. A simple analogy by Paul Churchland sums up the efficiency of parallel processing. It takes a typical PC about 15 minutes to run 100 billion elementary calculations. But a simple stage of the human visual system performs an equivalent number of calculations in 1/100th of a second because it performs all these computations independently and all at the same time. Churchland's analogy of this time-saving adaptation is demonstrated when a cook lines up a dozen or so green beans in parallel and cuts off all the stems in a single stroke. This analogy also answers the charge raised earlier that PDP is actually a subset of serial processing. Bates and Elman point out that it is "...still the case that most connectionist simulations are actually carried out on serial digital computers which mimic true parallel processing." (Bates and Elman 1993, 8/16) A set of would-be parallel computations is carried out in series, then batched in anticipation of the next wave of parallel simulations, in much the same way that our cook trims batches of green beans in a series of parallel operations. But this batching of parallel simulations is a design innovation which constitutes a significant departure from GOFAI systems design. To use the bean analogy, the cook does trim each end serially but the batching process means that dinner will be ready before 10 pm. Applied to the efficient use of cognitive resources, this distinction constitutes the difference between evolutionarily advantageous use and the waste of precious processing time, a concern which can be a matter of life

and death. As Paul Churchland says, "Evolution hit upon a winner when it stumbled across parallel distributed processing."(Paul Churchland 1996, 13)

2.5.2 Distributed Representations

Knowledge in GOFAI is stored as static patterns of symbols. In PDP models, knowledge is an emergent property of an activity pattern among units at time *t*. Thus, knowledge about any specific pattern is distributed across many different units. The result is that units are not representational: they have no semantic meaning and their function is defined within the context of their relationship among other units which constitute these patterns at a specific time. This will lead to a holistic account of cognition in which "...the representations are local at a global scale but global at a local scale."(Hinton *et al* 1986, 79)

Knowledge in GOFAI, on the other hand, makes a serious commitment to knowledge defined as sentential propositions justified on the basis of the discrete nature of truth (1) or falsity (0). But as we saw earlier, learning in PDP environments starts with a stochastic function which by definition is neither 0 nor 1. The implied aim of knowledge in GOFAI that it draw all and only correct conclusions is then replaced by the continuous analog nature of distributed representations and obviates the need for certainty. This is obviously closer to how intelligent agents actually go about their cognitive business. In PDP, distributed representations reflect the flux of plausible, if often deductively invalid, knowledge. The integration of uncertainty leads to two aspects of distributed representations -- coarse-coding and graceful degradation -- which are distinct from the symbol manipulation processes of serial digital computers. Each

reflects the role of error within connectionist architectures.

- *Coarse-coding*

Rather than deploying units in such a way that each unit represents information as precisely as possible with the goal of hitting an exact match, each unit in a neural net is designed to be sensitive to a *range* of possible different inputs. (Bechtel and Abrahamsen 1991, 54) In other words, each unit is *relatively* sensitive (activated) based on different *sets* of inputs. This indefinite set constitutes its receptive field. Significance is not based on the fact that a particular unit is active or non-active but on the percentage of units that are sensitive to a particular unit. A number of advantageous properties may be gained from this aspect of PDP, including a decrease in the number of units necessary for storage; a high tolerance for noise before memory is negatively impacted, and flexibility in the retrieval process. (Bechtel and Abrahamsen 1991, 56)

- *Graceful degradation*

In distributed systems, patterns built up in bits and pieces due to coarse-coding can also degrade in similar fashion. Therefore, "...computers can literally learn their way around faulty components because every unit participates in the storage of many patterns and each pattern involves many different units." (Rumelhart 1989, 152). The result is high fault tolerance because while the loss of a few components will degrade stored information, the whole batch will not be lost. This allows the system to avoid an 'all or nothing' response to damage. Therefore, PDP models require no special error recovery algorithms to work. Rather, graceful degradation is a natural by-product of the distributed nature of representations in connectionism.

2.5.3 Graded Rules

Knowledge is stored in the connections between processing units in a network and the configuration patterns of the units defines the dynamic network as a whole. Information processing consists first of the units transforming their input into output, then this output is in turn modulated by the weights of the connections as input to other units. These configurations are extremely plastic and are always competing with each other to resolve a given set of inputs. (Bates and Elman 1993, 9/16). There is never an absolute verdict on the final output. Instead, the alternatives are resolved *provisionally* with the implication that the output (knowledge) is derived from continuous (as opposed to discrete) learning functions used to adjust network weights.

It is important to note some implications of the transient quality of knowledge under graded rules. Connection weights have duration in time but this duration does not entail storage: "...retention and storage are different concepts...storage implies retention but retention does not imply storage."(Sutton 1998, 305). Furthermore, as Elman and Bates point out, "In a stochastic system of this kind, it is possible for several different networks to reach the same solution, each with a totally different set of weights. These rules are not absolute in any sense-- they can vary by degree within a given individual, and they can also vary in their internal structure from one individual to another."(Bates and Elman 1993, 9/16). The totality of connections defines the information 'content' at time *t* rather than representing information as a static symbolic, rule-bound structures.

2.5.4 Learning as Structural Change

Earlier a brief description of the back propagation of error was given . It is

important to stress now that this learning is unsupervised. The development of the generalized delta rule by Rumelhart, Hinton, and Williams allows PDP systems to change from potential states to active states *autonomously*. Rumelhart describes the two-step process:

> First, an input is applied to the network; then, after the system has processed for some time, certain units of the network are informed of the values they ought to have at this point. If they have attained the desired values the weights are unchanged. If they differ from the target values then the weights are changed according to the difference between the actual value the units have attained and the target for those units. This difference becomes an error signal. This error signal is sent back to those units that impinged on the output units. Each such unit receives an error message that is equal to the error in all the outputs to which it connects, times the weight connecting it to the output unit. Then, based on the error, the weights into these second-layer units are modified, after which the error is passed back another layer. This process continues until the error signal reaches the input units or until it has been passed back for a fixed number of times. Then a new input pattern is presented and the process repeats...such a procedure will always change its weights in such a way as to reduce the difference between the actual output values and the desired output values. (Rumelhart 1989, 151-152)

The result is that PDP systems change on the structural level due to internal processing activity which was not directly placed there by the programmer. This is a holistic approach since PDP systems with ontogenetic and internally adaptive outputs are constituted by *both* internal processes and external realities. In other words, context, in the form of past experience and new input on connection weights, spontaneously influences explicit activation patterns which are evoked on presentation of new input. The actualization, that is, the move from implicit to explicit state, of an activity pattern at time t, depends on the global state of the system, the individual connection weights, and the current input to the system. In short, machine knowledge is contextual.

Elman and Bates stress that, "...in no sense is the final product copied or programmed in."(Bates and Elman 1993, 9/16). The programmer doesn't define the final output; all she or he can do is to set the computer up beforehand in the best way possible, with a proper balance between lists of specific knowledge (algorithms) and hints about strategies and techniques (heuristics). Then the computer is on its own. Spontaneity of outcomes is the result, an output which is radically different from the certain products of deterministic if/then/else algorithms of traditional AI.

2.5.5 Non-Linear Dynamics

Two further PDP properties which result from unsupervised learning are the effect of hidden units which permit PDP systems to go beyond linear mapping into vector matrices and the development of learning functions which are heuristic in nature rather than algorithmic. "Because these networks are non-linear systems, they can behave in unexpected ways including U-shaped learning functions and sudden moments of 'insight'."(Bates and Elman 1993, 10/16) The result is that truly novel outcomes are generated based on the PDP property of spontaneous generalization. As we have seen, in connectionist architectures, knowledge is a property of the activation patterns between processing units and is distributed over a geometric construct without the mediation of symbolic representation. Similar patterns have similar effects with the result that internal generalization emerges as an inherent property of connectionist models given a degree of similarity between patterns which are described as vector space functions and not as symbols. (Rumelhart 1989, 148) This means that output can be generated based on approximate pattern recognition. Certainty in the form of an exact match is therefore

unnecessary for either information storage or retrieval.

The vector space model of connectionist architectures leads to the pre-eminence of heuristic processes over algorithmic computation. The difference between the two is this. Algorithms are externally imposed rules applied to the manipulation -- typically the classification, storage, and retrieval -- of symbols for the purpose of processing information. There is a typically hierarchal classification, along with a static series of inference rules/logic by which the manipulation of these discrete symbols is effected. As we saw earlier, common sense knowledge relies on the ability to make intuitive, often analogical, leaps in order to compensate for ambiguity and uncertainty. Algorithms are fixed rules which produce certain precise outputs. Heuristics are provisional rules-of-thumb which produce a range of outputs. Theses are usually ranked probabilistically and are geared to the specific task at hand. As we have seen, GOFAI is superior to connectionism in terms of logical precision. But as we have also seen, logic and precision have limited application to the common-sense requirements of knowledge.

The traditional definition of knowledge as justified true belief presumes a top-down approach to cognition. Yet the disorderly epistemic environment which surrounds us clearly suggests that a bottom-up account of cognition, as opposed to the rule-driven foundations of top-down processing, has much to say about a range of epistemological problems including knowledge, belief, representation, and content. Given the possibility of a bottom-up model of cognition, these elements should, at the very least, be re-thought. Thus, as Marvin Minsky suggests, we must "...know how to adapt each fragment

of knowledge to particular contexts and circumstances, and we must expect to need more and different kinds of knowledge as our concerns broaden." (Minsky 1991, 648). This calls for a re-configuration of epistemology on the order of Quine's naturalization of epistemology, as well as a new understanding of knowledge along the lines of the convergence of semantic holism and bottom-up processing models such as those of PDP architectures. In the following chapter I will describe the position of naturalized epistemology in the context of connectionism and show that a convergence between them is beneficial to both.

3. Naturalized Epistemology, Fallibilism, and Connectionism

In the previous section the connectionist program was outlined and the role of error in machine learning was investigated and shown to be an appropriate model for simulating and therefore understanding its role in the generation of knowledge. But this remains merely a description and tells us nothing about the normative implications of knowledge underwritten by error. In order to accomplish that task, it is necessary to either define a distinct epistemology derived from connectionism or else position connectionism within an existing theory of knowledge.

The basic requirement of the first option is an explanation which links microstructural cognition with its manifestations in macro-level states, such as belief, intentions, and meanings. Failing that, we can reduce mental states to micro-level physical processes, in effect dismantling the scaffolding provided by folk psychology. In the first case, we currently know too little about the brain to do more than speculate about how micro-level processes engender macro-level mental states. In the second case, the outright elimination of mental states comprised by folk psychology that has been advanced by eliminative materialists such as the Churchlands, is, I believe, both inadequate and unnecessary to the task of informing a distinct alternative epistemology for connectionism.

As defined by Patricia Churchland, eliminative materialism (EM), is comprised

of the following three claims:

- Folk psychology is a theory
- The inadequacies of folk psychology entail that it must be substantially revised *or* replaced outright (hence, "eliminative")
- What will ultimately replace folk psychology will be the conceptual framework of a matured neuroscience (hence "materialism")
  (P.S. Churchland 1986, 396. My emphasis)

The second claim in Churchland's definition means that eliminative materialism can be read in two ways: as strong EM or weak EM. Strong EM calls for the "outright replacement" of folk psychology, a replacement equivalent to the radical elimination of mental states. As Paul Churchland puts it, "A detailed neurophysiological conception of ourselves might simply displace our mentalistic self-conception in much the same way that oxidation theory (and modern chemistry generally) simply displaced the older phlogiston theory of matter transformation." (P. M. Churchland, 1979). But this tactic is inadequate because the "mature neuroscience" which replaces mental states lies somewhere in the future, hence the promissory note quality of any epistemology so derived. A strong reading is also unnecessary because the alternative weak reading of EM in which folk psychology is "substantially revised" allows us marginalize the exclusively mentalist aspects of folk psychology without completely eliminating the explanatory value of aspects of folk psychology such as beliefs, concepts, and knowledge. Instead, they may be provisionally kept, but redefined as artifacts within a connectionist account of how knowledge is generated. Having now rejected the first option of devising a distinct connectionist epistemology, we are then left with the second option which calls for integration of connectionism within an existing epistemology.

Thus, the first overall goal of this chapter is to situate connectionism within naturalized epistemology (NE) for the simple reason that, like connectionism, NE attempts to investigate the formation of knowledge from a purely scientific perspective, without the foundationalist edifice of ideal rationality and other metaphysical safeguards. Like connectionism, NE therefore segues quite naturally into the mechanics of learning (knowing-how). In structure, the overview of NE which begins this chapter, and the convergence of connectionism and NE which concludes it, underscore the distinctly non-linear structure of learning common to both connectionism and naturalized epistemology.

Convergence defines a relationship in which seemingly disparate entities share a common feature. For example, the wings of insects and those of birds are features of two distinct species yet they share a similarity of structure and function which is due to their common aerial environment. Likewise, naturalized epistemology and connectionism share a common view of knowledge as contextual and holistic yet they approach language from distinct perspectives. For Quine, language provides the empirical base for the naturalization of epistemology. Connectionism, on the other hand, relegates language to the position of cognitive artifact. We will see that with a weak reading of EM, the NE reliance of language is mitigated without abandoning knowledge and the path is clear for a convergence of connectionism and NE which is beneficial to both parties. This convergence provides the venue for a re-location of knowledge underwritten by error in which rationality and contentful inner states (i.e., beliefs) remain but the content beliefs carry and the processes they represent are radically alien to the folk psychology model.

3.1 Evolution of Naturalized Epistemology

Quine's naturalization of epistemology may be situated as the second of three revolts in epistemology. The first revolt took place after World War I and was prompted by the inadequacies of a purely metaphysical approach to knowledge. The philosophical standard bearers, known as logical positivists, included Carnap, Schlick, and Neurath. Quine was deeply influenced by these philosophers, and like all good students, came of age philosophically by refuting the key tenet of his mentors: the analytic/synthetic distinction.[9] The second revolt occured when Quine shifted the focus of linguistic analysis to language learning and undermined the certainty described by the logical positivist approach with a pragmatic appreciation for the uncertain conceptual underpinning of terms such as 'meaning' and 'reference'. Thus Quine subsumed the revolt of logical positivists against metaphysical certainty in the 1920s, with a second revolt against the alleged certainties of logical positivism in the 1950s. Although still in a promissory stage, the subsequent slow but steady displacement of folk psychology by neuroscience and AI promises a third conceptual shift which will allow us, as Paul Churchland says, to "...contrive to step out of our parochial self-conception, to transform our narrow concern with the rationality of belief into a global concern with the

---

9

This is the distinction between kinds of statements based on the nature of the evidence required to establish their truth. An analytic proposition is one in which the concept of the predicate is embedded in the subject as in, for example, the proposition 'all bachelors are men'. Analytic propositions are thus tautological and epistemologically empty with the result that the locus of knowledge cannot be analytic. Synthetic propositions, on the other hand, are epistemologically meaningful, but they are also contingent. Quine's *Two Dogma of Empiricism* dissolves the analytic/synthetic distinction and marks the beginning of his shift in focus from linguistics to language learning in epistemology. For Quine, a weak analytic/synthetic distinction may be marginally useful as a model of the different ways we know, but doesn't describe the different types of knowledge. (Quine 1953)

parameters of operation of 'epistemic engines' generally." (Churchland 1979, 6)

3.1.1 Reciprocal Containment

The dissection of the relationship between ontology, defined as the externalist, mind-independent reality described by natural science, and epistemology is essential: as Roger Gibson says, "Quine's philosophy cannot be properly understood without grasping the nature of this intimate relationship." (Gibson 1998, 45) The ontology/epistemology dynamic is tied to how language figures into ontological picture, then shifts to the mind as the device which generates language -- poorly -- to describe the world. Ontology focuses on objects of the natural world, in other words, *what* we know. These are questions of fact. Epistemology focuses on what happens in the process of knowing these objects, in other words, the question of *how* we know. These are questions of how we evaluate the sensory and cognitive input of objects.

Prior to Quine's naturalization, epistemology rested on the foundation of rational reconstruction by which it was presumed that epistemic statements (belief) could be translated into ontological statements comprised of purely observational and logical terms. Quine demonstrated that these translations are indeterminate and therefore traditional epistemology, which relies on the separation of psychological and philosophical inquiries, should be replaced by a new epistemology in which, "Epistemology, or something like it, simply falls into place as a chapter of psychology and hence of natural science." (Quine 1969a, 25)

Quine's reciprocal containment of ontology and epistemology reiterates his rejection of the analytic/synthetic distinction. As he defines it, "Epistemology in its new

setting is construed in natural science, as a chapter of psychology...[but]...the old containment remains valid too, in its own way...Our very epistemological enterprise, therefore, and the psychology wherein it is a component chapter, and the whole of natural science wherein psychology is a component book -- all this is our own construction." (Quine 1969a, 25) This bilateral containment is, Quine says, "...containment in different senses: epistemology in natural science and natural science in epistemology." (Quine 1969a, 25) By defining the relationship between epistemology and ontology as reciprocal, Quine re-configures the relationship between how we constitute the world (epistemology) and what the world is (ontology).

3.1.2 The Inherent Nature of Normativity in Description

The distinct approaches of traditional and naturalized epistemology which evolved from the analytic/synthetic distinction is underscored by their respective attitudes to descriptive and normative concerns. Within the traditional account of knowledge, the normative element is *a priori* and the descriptive element is *a postiori*. The normative element was traditionally attached to epistemology and the descriptive aspect relegated to psychology. The distinction between ontological and epistemological approaches to truth in Quine's naturalization of epistemology parallels descriptive and normative accounts of knowledge. Ontologically speaking, the world just is and truth is a function of an objective or mind-independent reality. Within the epistemological context, the world is isomorphic with how we order our experience of it and truth is therefore a

normative function of our beliefs, cognitive practices and/or capacities.[16]

Quine's reciprocal containment of epistemology and psychology provokes the normativity challenge. As he points out, the "...effect of seeing epistemology in a psychological setting is that it resolves a stubborn old enigma of epistemological priority...In the old epistemological context the conscious form had priority, for we were out to justify our knowledge of the external world by rational reconstruction. What to count as observation now can be settled in terms of the stimulation of sensory receptors, let consciousness fall where it may." (Quine 1969, 26). This externalist account of epistemology resolves the polarity between science and philosophy, but the onus on naturalized epistemology as a purely descriptive account, rather than a descriptive *and* normative program, raised the charge that naturalized epistemology cannot be normative. Based on the adage that 'is cannot imply ought', Jaegwon Kim, for example, asserts that Quine's externalism squeezes justification out of knowledge and since, from the standpoint of traditional epistemology, justification is the only truly normative term in the tripartite definition of knowledge, as justification goes, so goes epistemology. To preserve normativity within naturalized epistemology, normative tenets are then imported using the supervenience relation. (Kim 1988, 33-55)

But this importation is unnecessary. The definition of 'normativity' follows from the Latin, *norma*, a rule, and is therefore a regulative principle. In short, to define the rule

---

[16]This reflects the tension between bottom-up and top-down accounts we have described as well. And although I don't want to dwell on dichotomies, it may be noted in passing that the ontological/bottom-up pairing falls into the externalist camp while the traditional epistemological/top-down pairing is internalist.

is to prescribe the conduct. Thus, while it is understood that prescriptive and descriptive elements are *theoretically* distinct, within their regulative *qua* normative function this is a distinction without a *practical* difference. And, by extension, to define the constitutive principles of knowledge is to imply their regulative function. The epistemological form of knowing-that is then isomorphic with the scientific description of knowing-how and the prescriptive element of normativity is inherent within the description of the principles themselves.

Just as reciprocal containment subsumes epistemology within natural science, questions about how we *actually* arrive at our beliefs are then inherently relevant to questions about how we *ought* to arrive at them. Once the normative function of epistemology inheres within the description of epistemic conditions, the element of justification becomes superfluous. Because the conditions which 'justify' our beliefs and the processes that engender them are equivalent or reducible to a descriptive model -- one in which normativity is inherent -- justification as an essential trait of knowledge goes by the board.

The inherence of normativity in description is brought home by the externalist nature of NE in which cognitive states emerge by virtue of the agent's embeddedness in the external world. As Quine says, "The most notable *norm* of naturalized epistemology...is simply the watch word of empiricism: *nihil in mente quod primus in sensu.*"[11] (Quine 1990, 19. My emphasis.) The assertion that nothing in the mind comes

---

11

The use of the guiding principle of empiricism also reflects the *tabula rasa* of Locke, an approach to innate ideas that is addressed in Clark (1993b) but which is unfortunately beyond the scope of this thesis.

from what is not first in the senses reflects an externalism that is common to NE, EM, and connectionism. If, as most cognitive scientists assert, "...for every act of cognition...for example perceiving an object as a separate, three-dimensional entity among a myriad of entities...there is to be found a neuroanatomical structure and neurophysiological correlate, that is an instantiation in the workings of the brain," (Eimas and Galaburda 1990, 2) then epistemic states and psychological processes are functionally correlative and reducible to neurological processes. Their common function is the description of nomological *qua* normative patterns in nature. Since nomological is defined as 'lawlike' in reference, for example, to the laws of nature, normativity remains in inherent feature and Kim's charge that "...if justification drops out of epistemology, knowledge itself drops out of epistemology" (Kim 1988, 40-41) is undermined. Once again, the fact that NE is a descriptive account does not negate normativity because normativity remains an inherent property of description.

Furthermore, the descriptive nature of NE also indicates a direction for epistemological inquiry, in itself a normative function. As Patricia Churchland says, "If...we can determine how the brain conducts its epistemic/cognitive business, then we can proceed to get a theory concerning how to maximize efficiency in that business and hence how to maximize rationality in scientific inquiry. That is the heart of Quine's view" (P.S. Churchland 1986, 264) It is also the common goal of NE and connectionism.

3.1.3 Three Indeterminacies and Quinean Holism

Quine's rejection of apriority is also an implicit repudiation of Cartesian dualism in which the 'facts of meaning' parallel ontological 'facts of nature'. (Roth 1986, 434).

The resultant reciprocal containment of ontology and epistemology means that everything we can know is up for grabs, a free-for-all that is spelled out in Quine's description of three indeterminacies: the indeterminacy of translation, the underdetermination of theories, and the inscrutability of reference. Quine's indeterminacy of translation thesis states that, "...manuals for translating one language into another can be set up in divergent ways, all compatible with the totality of speech dispositions yet incompatible with one another."(Quine 1960, 27). The result, as Gibson and Barrett summarize, is that "..if contrasting correlates can be equally correct then there is nothing that a given sentence-to-sentence pairing proves and it is only in *patterns of usage* that meanings lie." (Barrett and Gibson 1990, xx. My emphasis.). Therefore, *contra* Chomsky, the universal grammar which supposedly underlay linguistics does not account for common theories of the world. Rather, patterns of usage provide the contextual frameworks which actually carry content. Thus, epistemology becomes a matter of semantics and the stage is set for the underdetermination of theories and inscrutability of reference theses.

With a strict boundary between semantics (what the words mean) and ontology (what the world is) blurred, we find ourselves in Neurath's boat with no Archimedean point from which to justify truth and falsity. To describe the ontology in question is nothing more than to lay out how we plan to translate one set of terms into a set of terms which has meaning for us, given our background theory. So whereas indeterminacy of translation states that there is no epistemological fact of the matter, the underdetermination of theories implies that there is no right way to attach semantic terms

to the objects they refer to and, as Quine says, the difference between 'gavagai' as rabbit and 'gavagai' as rabbit parts is a matter of "how you slice it." (Quine 1969b, 32) A fixed one-to-one word-to-world fit is lost, an outcome which discredits the museum myth that fixed labels connect words and the world and hence makes the notion of reference opaque. This also means that we are forced to re-invent the theoretic wheel each time we encounter the world, with the result that semantic and ontological relativity now seem unavoidable.

Does ontological relativity imply that meaning and reference degrade to the point where we never know what our terms or objects actually refer to? That prospect is avoided by Quine's use of proxy functions as the method for mapping disparate ontologies which allows us to change the ontology of a theory without altering the weight of the evidence. As one-to-one transformation rules, proxy functions allow unique objects in a 'new' ontology to be mapped to objects in the 'old' ontology. For example, given that for any object 'x', we can specify $f(x)$, and that dog in the old ontology is P: in the new ontology $x=f(P)$ where P is dog and x is the proxy function of a dog. It is therefore unnecessary to re-invent 'dog' from scratch each time we see Lassie. It is important to underscore that the object 'x' in the new ontology is not axiomized as a dog. Rather, we assign it a *function* of x which denotes x as a "lifelong filament of space/time occupied by a dog." (Gibson 1988, 135) The fact that 'x' refers to dog in at least one translation manual can, by proxy function, mean that it also denotes dogs in another translation manual. The upshot is that while talk of absolute reference goes by the board, talk of relative reference does not. (Gibson 1988, 137). The resultant semantic holism

can then be seen as a methodological response to ontological relativity.

The crucial point is that the patterns of inter-relationships between the three indeterminacies take precedence over the units/sentences/theories (the symbolic content) they are made up of. In the absence of an Archimedean point upon which to fix meaning and reference, semantic holism derived from these inter-relationships becomes the only method which allows us to make sense of the world. This has implications for both scientific statements and theory development which feed into the conceptual underpinning of NE. Quine's view that, "...scientific statements are not separately vulnerable to adverse observables, because it is only jointly as a theory that they imply their observable consequences." (Quine 1992, 8) means that semantic holism is then the only appropriate approach to theory development and evaluation. It is also crucial to note that although our epistemic position is taken relative to background theory, the ontological ground of that relative position is itself fixed. Hence, Quine's holism, found in his *Two Dogmas of Empiricism*. The entire passage bears quoting:

> The idea of defining a symbol in use was, as remarked, an advance over the impossible term-by-term empiricism of Locke and Hume. The statement rather than the term came with Frege to be recognized as the unit accountable to an empiricist critique. But what I am now saying is that even in taking the statement as unit we have drawn our grid too finely. The unit of empirical significance is the whole of science. (Quine 1953, 42)

In other words, Quine maintains a symbolic definition of language but extends the linguistic unit from the logical positivist statement to the whole of science. For Quine, whether knowledge is expressed in beliefs about phlogiston, chairs, or the standard model, knowledge is an interconnected web in which no one part is immune from

revision in light of experience and, at the same time, no experience can force the rejection of just one part, but must prompt a re-evaluation of the entire construct.

3.2 Fully Naturalized Epistemology

A nutshell description of connectionism views mental processing "...as the dynamic and graded evolution of activity in a neural net, each unit's activation depending on the connection strengths and activity of its neighbours." (Garson 1991, 113) Quinean holism in which content is meaningless without context is also compatible with connectionist models in which knowledge is derived from widely distributed activation patterns. The defining property of distributed representations is isomorphic with the essential feature of Quine's holism. Semantic holism dictates that meaning is contextual so a revision of even the smallest part of the whole provokes a complete overhaul of the construct. Similarly, in PDP, cognitive processing is the graded and dynamic evolution of activity in a neural net, each unit's activation depending on the connection strengths and activity of all its neighbours. (Garson 1997) As Michael Dyer explains, "the distributed nature of these representations allow models to exhibit adaptive learning, generalization, and fault tolerant properties." (Dyer 1991, 385) Similarly, in semantic holism, the truth-conditional conception of meaning means that knowledge is a function of the inter-relationships of beliefs within the whole. This dynamism is the common basis on which meanings change in NE and learning occurs in connectionism.

Far from reducing or eliminating the indeterminacy of translation, connectionism reflects the integration of all three indeterminacies in terms of both cognitive input and

output. As we saw in the indeterminacy of translation, Quine concluded that it is patterns of usage between units/sentences/theories which hold semantic content. Likewise, in connectionism, content is defined by the weights of the connections between neuronal units. In the underdetermination of theories, Quine asserts that there can be no absolute or aprioristically determined 'right' word/world fit. In connectionist models, weights are defined and redefined continuously using the backpropagation of error. This is the indeterminate/heuristic essence of bottom-up processing as opposed to deterministic rule-driven approach of top-down processing. Finally, the fact that Quine's ontological relativity is saved from incoherence by the use of proxy functions is parallelled by the use of probability functions in the process of statistical analyses which characterize how connectionist weights are determined in neural networks.

Nevertheless, the convergence of NE and connectionism also includes an essential difference. While it is a crucial aspect of NE, language learning takes a backseat in PDP, a difference that is addressed in the next section.

3.2.1 Connectionism and Language Learning

It is important to recall that Quine's naturalizing of epistemology was developed at the midpoint of the century when behavioural psychology was seen as the only viable empirical alternative to primitivism or mentalism. Quine used language-learning not because he had any particular affinity with the universalizable linguistic basis of logical positivism -- we know from *Two Dogmas* that he did not -- but because the process of acquiring language afforded the most empirically verifiable means to observe learning *at the time*. However, advances in AI and imaging technologies suggest that if Quine had

had access to these technologies[12] language-learning would have paled as a methodological tool, and with it, the importance of language learning and empirical psychology would also have waned in Quine's epistemology. Quine's epistemological mandate is to address "...the question how we physical denizens of the physical world, can have projected our scientific theory of that world from our meagre contacts with it; from the mere impacts of rays and particles on our surfaces." (Quine 1995, 16). While language provides a tool to that end, loyalty to the models of behavioural psychology foregoes the full scope of naturalized epistemology and ultimately fails to optimize the new beginning advertised in Quine's program.

A connectionist account of knowledge, on the other hand, defines language as an artifact of the cognitive process that is "...acquired, sustained, and administered by more fundamental information-processing systems of a non-linguistic kind." (P.M. Churchland 1979, 139). Language is therefore neither the foundation nor the means by which knowledge is generated. Language-like structures do not constitute the basic forms of representation in cognitive creatures and there is no correlative assumption that cognitive constraints in the manipulation of those representations occur by means of structure sensitive rules. In connectionism, language learning becomes one artifact of the cognitive engine and affords no privileged status toward the understanding of cognition.

Earlier we mentioned that Quine's indeterminacy of translation tacitly

---

12

Freeman Dyson supports this general point, though not in specific reference to Quine, when he lauds the work of philosopher of science Peter Galison. Galison's view is that tools and not concepts are the key forces driving scientific progress. For Galison, Dyson says, "...science arises out of great leaps of practical ingenuity that enable scientists to acquire new data...if the tools are bad nature's voice is muffled. If the tools are good nature will give a clear answer to a clear question." (Dyson 1999, 34)

undermines Chomsky's linguistics because in Quine's account, universal grammar rules are inadequate to explain why distinct epistemic perspectives yield isomorphic ontologies. As Paul Churchland summarized, "Chomsky's approach postulates the existence, within any linguistically competent human, of a set of *rules* for the formulation of admissible or grammatical sequences of words. And it assumes that the brain *applies* or *follows* those rules in order to comprehend and produce actual sentences." (Churchland 1996, 134. Emphasis in the original) Chomsky's implicit acceptance of the top-down classical model of cognition is rejected by Quine yet he retains the function of language within strict confines as language learning. However, Quine's holism, as well as his recognition of the inadequacies of universal grammar and the linguistic pre-occupation of his logical positivist mentors, sugests that he tacitly accepted a bottom-up approach to cognition without subscribing to the strong version of EM.

As we have seen, Quine restored the epistemic project by giving it a new setting as "...a chapter of psychology and hence of natural science." (Quine 1969, 25) In order to develop a fully naturalized epistemology, the convergence of connectionist AI and NE requires the giving up of propositional attitudes and language-like cognitive foundations and reflects the bottom-up processing view that, "there is a level of representation beneath the level of the sentential or propositional attitudes, and to the correlative idea that there is a learning dynamic that operates primarily on sublinguistic factors." (P.M. Churchland 1990, 61) A maturing neuroscience thus obviates the usefulness of language learning in NE and leads to a fully naturalized epistemology that incorporates the

connectionist view of subsymbolic learning and so completes the revolution Quine began when he naturalized epistemology.

3.2.2 Language and Semantic Content

Quine is well aware of eliminative materialism (EM) and although he approves the reclamation project of the Churchlands, he nonetheless maintains that, "...sentences are needed to bear truth-values and it is sentences that stand up as the objects of knowledge and belief -- at least they do after supplementing by the keystone of the mental content-clause 'that *p*'" (Woods 1992, 561) For their part, although the Churchlands and other eliminativist materialists agree with Quine's naturalization of epistemology, as we have seen, they reject language as the measure of cognitive virtue and assert that language gives at best a superficial representation of truth. A connectionist approach to content may thus bring the two sides together and enhance the development of a fully naturalized epistemology.

Quine's criticism of EM is well-placed and confirms my earlier position that the strong version of EM in which the entire edifice of folk psychology is eliminated is unnecessary. Quine's challenge is embedded in the premise that it is sentences that stand up as the objects of knowledge and belief...after supplementing by the keynote of the mental content-clause 'that *p*'. In other words, we need sentences in their contentful capacity, 'content' being defined as that-*p*. But what if content is re-defined in a way that is both compatible with Quinean holism and yet represents knowledge without the sentential scaffolding? First, we saw earlier that normativity is inherent in description. Therefore, to describe how-*p* is to prescribe that-*p* *and* that-*p* is not necessary for content

in a fully naturalized epistemology. Furthermore, the distributed nature of content is compatible with Quinean holism, as we also saw. The fact that knowledge arises from the connections between processing units and is not a property of the units themselves, means that knowledge in PDP systems is an emergent property of an activity pattern among units at time *t* and knowledge about any specific pattern is distributed across many units. The result is that content has no fixed meaning. There are no static objects of belief. Content, and therefore knowledge, is defined within the context of the relationships among units which constitute patterns at any given time and cannot be confined to sentential form. It is useful to code content in propositional form but that is artifactual convention and does not the imply privileged epistemic status presumed by both traditional epistemology and GOFAI.

Quine's maintenance of the classical top-down processing model implied by the language learning model and the notion of sentences as the bearers of content is therefore puzzling given that the classical model is a local approach, yet Quinean holism is a profoundly global way of looking at knowledge. Let's apply the local/global distinction to content using the connectionist network of units and contentful weights. A representation is designated local if it names the content associated with the firing of a single unit; to use the earlier example, if it corresponds to a single neuron that specifies 'grandmother'. It is said to be distributed if it names a content associated with a joint activity of several units which together indicate 'grandmother' at time *t*. The local definition of content favours the top-down approach of traditional epistemology and classical cognitive science; the global definition of content is isomorphic with bottom-up

processing and distributed content. But despite Quine's apparent loyalty to sentential content, his holistic account of the knowledge these sentences represent implies at least tacit acceptance of the distributed nature of content such as that found in connectionism. In effect, Quine's holistic account of cognition mirrors the connectionist account of Hinton *et al* in which "...representations are global at a local scale and local at a global scale." (Hinton *et al* 1986, 79). As I see it, content (knowledge) derived from the reciprocal containment of global and local processing in connectionism parallels content (knowledge) derived from the reciprocal containment of ontology and epistemology in NE.

*3.3 **Knowledge and Knowing***

We have extended the descriptive element of connectionism toward normative implications by the integration of PDP within the existing epistemological framework of NE. We have also seen that elements of weak EM, NE, and connectionism may be combined with the overall aim of situating connectionism within the existing epistemic framework of NE: connectionism may be adapted to NE by virtue of overlapping accounts of holism while at the same time, NE may also be strengthened if its reliance on language learning and propositional content is replaced by bottom-up cognitive models which employ distributed subsymbolic content.

The convergence of NE and connectionism blurs the distinction between knowledge and knowing, completes the re-configuration of epistemology that began in 1969 with Quine's *Epistemology Naturalized* and so gives normative backing to the first

principle of knowledge underwritten by error: that we ought to be fallible about beliefs and realistic about believers. In chapter four, we apply this principle to believers using Christopher Cherniak's model of minimal rationality. We will show that the connectionist model of processing is an improvement over the information processing model Cherniak applies and that the substitution of PDP for Cherniak's serial processing model leads to a more realistic understanding of knowledge underwritten by error.

4. Minimal Rationality, Cognitive Constraints, and Connectionism

Christopher Cherniak's account of minimal rationality (MR) is built on the premise that our cognitive reality is bounded and finite. Yet despite the fact that error is the medium of knowledge for cognitive agents, the prevailing philosophical trend continues to apply ideal rationality standards to cognitive processes with the result that non-ideal reasoning is then labelled as performance error. This implies that error is aberrant behaviour which is to be avoided, *qua* avoidable. In contrast, the goal of this thesis has been to investigate an alternative approach to knowledge predicated not on the avoidance of error but on the necessity of error in the learning process.

Minimal rationality assumes both the fallibilism of belief and realistic constraints on believers derived from fully naturalized epistemology in the previous chapter. Cherniak's point with MR is to show that these constraints -- the 'finetary predicament' as he calls it -- renders ideal rationality standards not only inadequate but counterproductive. I concur with this view, but, as I will demonstrate in this chapter, I believe the information processing model of cognition Cherniak uses fails to develop the implications of minimal rationality to their full potential. However, the principle that we ought to remain fallibilistic about beliefs and realistic about believers is fully realized if the PDP model is used as the foundation for MR rather than the serial information

processing model Cherniak relies on.

I will begin by outlining Cherniak's position. The salient points of his approach to rationality will be summarized and these points will then be juxtaposed with a description of the connectionist program. The purpose here is to show that minimal rationality is better served by the PDP model of cognition. Before we start, however, a few definitions are in order. First, 'rationality' is defined as a state in which a set of beliefs, desires, and decisions should cohere. To be rational, a belief or decision must at least cohere with the rest of the person's cognitive system. Cherniak's account of rationality presumes an agent-constitutive approach by which, "A person's putative beliefs must mesh with the person's desires and decisions, or else they cannot qualify as the individual's beliefs."(Cherniak 1994b, 526) The definition of 'content' in MR takes beliefs, desires, and other intentional states to be representational states. As opposed to distributed representations which fulfill the generative role of content in connectionist systems, Cherniak's definition of content therefore corresponds to the traditional one in which normativity is a property expressed in the relationship between beliefs and desires.

The view that minimal rationality is compatible with connectionist models is contrary to Cherniak's view that machine models of the mind are impracticable because they are bound by resource limitations.[13] Cherniak concludes that "...the full mind's program appears to be a type of practically unknowable thing-in-itself" (Cherniak 1988,

---

[13] Cherniak's view is more than a little paradoxical. Human rationality is limited by finite computational resources. This is supposed to argue against the practicability of AI? As we shall see, no such inference can be drawn from the resource limitations common to both human and machine intelligence; quite the contrary.

402) and extends this view to connectionist models when he says that, "Recent neurally inspired connectionist conceptions of distributed and massively parallel architectures can only exacerbate these structural unevaluability difficulties."(Cherniak 1988, 410) The result is that, "A complete computational approximation of the mind would be (1) huge, (2) 'branchy' and holistically structured, (3) quick-and-dirty (i.e., computationally tractable but formally incorrect/incomplete), and (4) 'kludge'."[14] (Cherniak 1988, 410). A machine approximation of the mind's program would be equivalent to an impossibility machine and would defy the goals of comprehension and evaluation it was designed to facilitate. Therefore, any "full-scale software description of the mind faces a predicament of diminishing returns." (Cherniak 1988, 411) The point is, however, that PDP is not a 'software description of the mind' but a model which, as we have seen, blurs the hardware/software distinction altogether. In so doing, the very properties which Cherniak considers to be obstacles to machine models, for example their 'branchy' nature and huge size, are constructively addressed by the connectionist research program. So while Cherniak's development of minimal rationality admirably depicts an alternative to idealized rationality, he fails to fully develop the implications because he mistakenly adopts the information processing (IP) model of memory favoured by traditional AI. He

---

[14] 'Kludge' is defined by Cherniak as a "radically inelegant set of procedures". (Cherniak, 1988, 402). But it should be noted that just as PDP technology allowed for the resolution of the exclusive OR (XOR) problem in early connectionist architectures, evolving technologies such as quantum computing may well address the problem. However, according to Ron Chrisley, quantum neurocomputers cannot be incorrect: "A quantum state 'means' just whatever caused it; there is no room for falsity or error." (Chrisley 1996, section 4.3) If Turing, on the other hand, is correct, and I think he is, then how can quantum neurocomputers simulate intelligence? Chrisley addresses learning in quantum neurocomputers at http://www.cogs.susx.ac.uk/users/ronc/quantum3/quantum3.html.

is therefore locked into content defined within the symbolic, rule-based paradigm. But symbolic content is also the bedrock of the idealized rationality model Cherniak has rejected. Thus, in the interest of both technological and epistemological consistency, the IP model of memory should be replaced by the PDP model in which content[15] is comprised by distributed representations. This contradicts the tacit acceptance of purely symbolic representations that Cherniak's uncritical adoption of the duplex model of memory structure implies.

4.1 From Ideal to Minimal Rationality

The move from the presupposition of ideal rationality to the development of a realistic theory of rationality begins with Cherniak's description of the finetary predicament. As Cherniak says, "...one of the most fundamental aspects of the human condition [is] that we are in the finetary predicament of having fixed finite limits on our cognitive resources." (Cherniak 1986, 127) This realistic appraisal of the limits of human rationality is at odds with the traditional view of a "...pervasive and falsely assured conception of rationality in philosophy which is so idealized that it cannot apply in an interesting way to actual human beings."(Cherniak 1986, 5)

Taking a page from Herbert Simon's notion of 'bounded rationality', Cherniak claims that "...actual human beings generally have neither perfect information about their world of alternative choices nor a capacity to use such knowledge."(Cherniak 1994b,

---

15

This mirrors the related thesis of Sutton (1998) wherein the view that memories are dynamic patterns rather than static archives is developed.

528) Cherniak explains that Simon proposed a principle to the effect that *A* "satisfices"[16] its expected utility and rational agents ought only to try to 'satisfice' rather than maximize. (Cherniak 1994b, 528) The notion of bounded, or finite, rationality is then bolstered by complexity theory and classic insolvability theorems, such as those argued by Alonso Church and Kurt Gödel. It is useful to briefly describe each of these conceptual inputs as they confirm the limits of deduction and cognition which support both the fallibilism of beliefs and realistic constraints on cognitive agents.

According to Turing, both Gödel and Church produced theorems that entail "...limitations to the powers of discrete-state machines". (Turing 1950, 503). Gödel's Incompleteness Theorem leads to the conclusion that "...in any sufficiently powerful logical system, statements can be formulated which can neither be proved nor disproved within the system, unless the system itself is consistent."(Turing 1950, 503) Church's Theorem states that because the theorems of predicate logic do not form a general recursive set, there is no decision procedure (or algorithm) for first-order logic.[17] Cherniak takes it upon himself to extend Church's Theorem to include propositional logic with the conclusion that "...a kind of practical undecideability seems to extend further down, to the most basic parts of logic, to the very core of computation...it is as if Church's Theorem applied even to the propositional calculus." (Cherniak 1986, 79) In

---

16

'Satisfice' is a term Simon coined for the rationality requirement of bounded rational agents. It is a combination of 'satisfy' and 'suffice'.

17

If both a set and its complement can be ordered as recursive functions, (i.e., as $f(0), f(1), f(2)...$) then the set is recursive. This function correlates with decidability and hence with computability. But according to Church, the set of theories of predicate calculus cannot be ordered as a recursive set. Therefore, the predicate calculus is undecidable.

other words, some problems are by their very nature computationally intractable and can only be solved "...by methods that exceed the capacity of the most powerful computer imaginable". (Stockmeyer and Chandra 1979, 140) As a result, Cherniak says, "...formally correct and complete inference procedures appear to be intrinsically intractable," (Cherniak 1994b, 529) an outcome equally true for human and machine intelligence.

Opposed to such a necessarily circumscribed description of rationality is the traditional idealized view of human rationality which has been presumed since Aristotle first defined man as a rational animal. Cherniak distills this traditional view into an idealized normative rationality standard which is defined by global consistency, deductive closure, and the normative rule that rational agents satisfy both these conditions. C. A. Hooker sums up Cherniak's description as one in which an idealized, "...fully rational agent makes all and only sound inferences from beliefs and values, which means that he uses some system of logic that is sound and complete."(Hooker 1994, 83)

Cherniak then proceeds to dismantle the idealized rationality standard. First, in terms of the global consistency requirement, complexity theory has shown that it is not possible to carry our global consistency checks on even the most powerful computers imaginable. Second, Cherniak's definition of a *Deductive Closure Condition* (DCC) by which A actually believes (or infers) all and only consequences of A's belief, presumes ideal deductive ability and is equivalent to requiring that a finite agent perform a task requiring infinite capacities. (Cherniak 1986, 12). Gödel and Church demonstrated that

inferential systems are incomplete/undecidable in both the propositional and predicate calculus, which obviously violates the requirement of the DCC. Finally, the normative form of the rule, which Cherniak calls the Ideal Inference Condition (IIC), stipulates that, "*A* would make all and only sound inferences from the belief set that are apparently appropriate". However, this entails a logical inconsistency because making 'all and only sound inferences' requires the physical impossibility of "...resources greater than those available to an ideal computer constructed from the entire universe". (Cherniak 1986, 81)

Ultimately then, the notion of ideal rationality must be discarded. At the same time, the radically opposing subjectivist view, characterized by Cherniak as the 'assent theory of belief', is also counterproductive. A position in which the believer is the final arbiter of his beliefs is equivalent to a "null" rationality requirement and is therefore without normative content. (Cherniak 1986, 6-7). Instead, "...any cognitive theory that is to satisfy the basic constraints of having significant empirical content, applying to finite creatures much more that in principle, and being applicable by finite arbiters, must include a *via media* fundamental principle that an agent has some, but not ideal, logical ability...minimal rationality conditions seem indispensable in this way for satisfactory cognitive theory. (Cherniak 1986, 87) Therefore, Cherniak sets a standard for the implementation of minimal rationality which he calls the *Minimal Inference Condition* (MIC). For an agent with minimal rationality, "If *A* has a particular belief-desire set, *A* would make some, but not necessarily all, of the sound inferences from the belief set that are apparently appropriate." (Cherniak 1986, 10)

It is important to note, as Hooker does, that Cherniak's "...arguments are

focused less on the minimal requirements than on exploring the constraints imposed on maximal achievable performance for finite rational agents."(Hooker 1994, 198) Because the MIC is fuzzy, Cherniak appropriates Hilary Putnam's notion that every cognitive concept is effectively a cluster concept: "The specification of the minimal inference condition remains combinationally vague; its structure makes every cognitive concept a cluster concept. The minimal inference condition by itself identifies not a simple defining property but a cluster of properties... The minimal condition seems to be employed probabilistically."(Cherniak 1986, 18-19) The probabilistic function of MIC as defined by Cherniak therefore cannot behave in a manner characteristic of algorithmic necessity, but rather implies distinctly indeterministic outcomes.

4.2 Feasibility and Memory Structure Constraints

The constraints which emerge as a result of Cherniak's minimal inference condition are divided into two classes he calls *Feasible Inference Theory* and *Human Memory Theory*. These ancillary theories are best understood as constraints on the implementation of human rationality.[18] The first constraint set concerns the agent's

[18] A byproduct of this approach is that once these constraints are integrated into the structure of rationality, they can be seen as the source of the logical inconsistencies isolated by the empirical work of psychologists such as Kahneman and Tversky (1974). By extension this suggests a link between the evidence of logical inconsistency and the use of heuristics. As Cherniak says, "...much recent empirical work on the psychology of people's use of 'quick and dirty' heuristics has focused on the widespread phenomena of at least apparent breakdowns in consistency. It should be emphasized that inconsistencies in a belief set need not be at all inexplicable. The logical relations among the beliefs involved in an inconsistency may be very unobvious and so not recognized; another important source of inconsistency is the structure of human memory." (Cherniak 1986, 18). I would also add that these 'apparent breakdowns in consistency' are the price we pay for a cognitive engine that processes partial input using heuristics. From an evolutionary standpoint, this is a small price to pay.

deductive abilities. The second describes the agent's memory structure and is derived from an information processing model of memory.

4.2.1 Feasible Inference Theory

The ideal agents' perfect capacity to make all and only sound inferences requires computationally intractable algorithms which, as we have seen, are beyond the range of finite human processes. However, fallible humans do manage to generate knowledge by using heuristic (i.e., non-algorithmic) procedures. It is important to note that these procedures contradict ideal deductive standards since by definition heuristics, unlike algorithms, don't work in all cases. They are, *designed* to be fallible. It is then but a short step to a normative statement of minimal rationality: "The ideal rationality models are at best silent on the normative status of these heuristics; the minimal rationality model allows us to acknowledge the basic platitude that we're finite and so we *ought* to use some such heuristics. *According to this conception, formally incorrect heuristics need not in fact be irrational at all.* They are just inadvisable or unintelligible sloppiness, because they are a means of avoiding computational paralysis while still doing better than guessing." (Cherniak 1986, 82. My emphasis.) Cherniak is mistaken in characterizing heuristics as inadvisable or unintelligible. Instead, within a connectionist framework these heuristics are not only not inadvisable but, as we have seen, they constitute an invaluable element in the processing of partial data into knowledge. The bottom line is that error is not necessarily 'irrational'. Instead, it may contribute to intelligent processing models: when paralyzed by computational kludge it may only be the integration of error that allows us to generate an applicable if not deductively valid

quick-and-dirty heuristic.

Simply put, feasible inference theory begins with the constraint that as agents with finite resources, ideal rationality is impossible. What is then more important than getting the "right" answer is knowing where to apply our finite cognitive resources. This leads to the open question of what constitutes 'feasible' inferences as opposed to ideal inferences. Since we are constantly running epistemic cost-benefit analyses (CBA) in order to choose the most likely inferences, one of the primary uses of rationality is not for deductive inference but to strategically parcel out the cognitive workload. As Hooker says, "...we have a double sense in which the exercise of reason is context-dependant. The choice of strategy is context-dependant, specifically historically dependant, through the direct and opportunistic costs and the benefits appearing in each context, [and] the choice or shaping of contexts themselves are context-dependant." (Hooker 1994, 196)

This presumes a holistic perspective to knowledge and is therefore compatible with the fully naturalized epistemology of the previous chapter. The point is not to use our rational capacity to compute a solution but to "...*become* a solution, i.e., to become internally re-arranged in such a way that appropriate solution outposts are generated... In this case we have created a procedure [strategy] which fits itself to a solution."(Hooker *et al* 1992, 533. My emphasis). This also reflects the blurring of the distinction between inter-active knowing and static knowledge acquisition illustrated by a fully naturalized epistemology. The strategy thus produced is heuristic and, as Andy Clark puts it, "The Rational Deliberator turns out to be a well-camouflaged Adaptive Responder." (Clark 1997, 33.)

Given such an approach, the feasibility theory concludes that:

- An agent must be only an adequate logician: there is no transcendentally right kind of logician.
- There are remarkably few *a priori* constraints on a rational agent's deductive abilities.
- Feasibility theory is an essential element of a predictive cognitive theory.
- Holistic interdependence of beliefs, desires, and meanings emphasized by Quine and Davidson may be extended to the computer science domain. (Cherniak 1986, 48)

4.2.2 Memory Structure

Cherniak subscribes to the information-processing model which is the foundation of serial processing. He therefore assumes the duplex memory model in his development of MR. On this basis, there is a level for the activated belief-set which is manifested in short-term memory processes, and another distinct level for the inactive belief-set of long-term memory. The duplex model of short and long-term memory is then superimposed onto the standards of information processing (IP) and information organization (IO) respectively. The cognitive performance implication for IP/short-term memory is that only short-term memory can be presumed in reasoning and therefore more rationality is needed for tasks which take place in short-term memory. On the IO/long-term memory level, since beliefs here are inert, they do not directly affect behaviour. (Cherniak 1986, 59) The upshot is that cognitive efficiency is defined by the speed and accuracy of retrieval from long to short-term memory. This recall is in turn based on the method used to classify items in long-term memory and presumes a static archival quality of representation in Cherniak's model such as that of the museum myth which Quine disproved.

For our purposes there are two important aspects of Cherniak's approach to memory. First, the method of classification is dictated by the current goals/desires of the agent which are inextricably interwoven with the agent's belief set. (Cherniak 1986, 65) Second, the finetary predicament extends to the duplex model so that efficiency is constrained by the impossibility of exhaustive searches. Exhaustive searches are themselves idealized requirements given the finite resources of our duplex human memory structure. Therefore, what is required for minimally adequate recall is a set of satisfactory *partial*, as opposed to *exhaustive*, search strategies. Just as algorithms necessarily give way to heuristic strategies in terms of minimal inference conditions, so too must exhaustive exact-match strategies give way to heuristic strategies designed to retrieve information based on incomplete input. This implies the need for a strategy of localized searching which can be met by integrating content-addressable memory in the structure of long-term memory, as well as the resulting need for a retrieval function in which there is a trade-off of hit accuracy for retrieval speed. In practice, such cost-benefit analyses explain why a judgement may seem irrational considered locally, and yet be rational when it is globally considered as part of good memory management. (Cherniak 1986, 67)

4.3 Error and Minimal Rationality

As we have seen, the cornerstone of Cherniak's account is the reality of our limited cognitive resources. This means that any valid theory of rationality ought to integrate the error which characterizes our finetary predicament. A resulting need to rely

on quick and dirty heuristics is therefore a central theme of finite rationality. (Cherniak 1986, 18, 75-6, 81-2)

In terms of the *Theory of Feasible Inference*, "...quick and dirty heuristic processes that yield approximately optimal solutions can be much more feasible..." (Cherniak 1994a, 97) than the deductive processes presumed by ideal rationality. As Cherniak says, "...we are each in the finetary predicament, with cognitive resources that are severely limited relative to the range of possible inquiries...we cannot obtain and use all available information, furthermore, it would not be advisable to attempt to do so...we must therefore try to determine the best use of our resources by deciding which information would be the most useful to seek." (Cherniak 1986, 64)

Once again we see Turing's link between fallibilism and intelligence within Cherniak's minimal reference conditions (MIC). Fallibility is required for intelligence and is, by implication, an essential feature of a realistic definition of cognition. Cherniak concludes that "...human being must evade [this] computational paralysis by using quick-and-dirty procedures."(Cherniak 1988, 410) So the *Feasible Inference Theory* replaces brittle rules with soft constraints in which, for example, logic need only be adequate. Filtering and probabilistic CBAs are used to assign weights to the feasibility ordering of inferences. In terms of the *Theory of Memory Structure*, information compartmentalization yields quick, if sometimes inaccurate, retrieval and the inevitable recall errors are the price of doing business quickly enough to do us some practical good. All these features are compatible with aspects of the connectionist model of cognition.

The limitations of our cognitive resources provoke two hallmarks of Cherniak's

rationality account: the use of heuristics over algorithms, and the criteria of context-dependance in both sets of constraints. This context-dependancy implies a holistic design for the structure of rationality based on the inter-relationships among sets (clusters) of concepts and by extension reflects the convergence of NE and connectionism described earlier.

4.4 Connectionism and Minimal Rationality

The most expedient way to develop the similarities between connectionism and MR is to define a set of 5 characteristics of connectionist systems and juxtapose the appropriate elements of Cherniak's model of minimal rationality to each. In outline, connectionist systems:

- are good at solving constraint satisfaction problems;
- are efficient mechanisms for best-match (as opposed to exact-match), pattern recognition and content-addressable memory;
- are capable of automatically implementing similarity-based generalizations;
- include simple, general mechanisms for adaptive learning;
- use the distributed nature of representations to insure 'graceful degradation' in the eventuality of damage or information overload.

4.4.1 Constraint Satisfaction Problems

As we have seen, connectionist models can be considered constraint satisfaction machines in which networks generate processing output by 'settling' into states of maximal satisfaction of the constraints implicit in the network. This "goodness-of-fit landscape" (Rumelhart 1989, 221) is a key concept in connectionism and is similar to the process in which feasible inferences are isolated using a strategy based on context-

dependancy. A second similarity is the response to real-time constraint problems. The intractability of computational models and the unrealistic nature of ideal rationality both reflect the fact that insufficient time is available for exhaustive searches given finite resources. An immediate advantage of PDP is the effect on time. The brain makes up in the volume of processes what it lacks in speed because of the fact that it deploys synaptic connections co-operatively and in parallel. Because the brain processes which connectionism seeks to characterize in machine form are incredibly complex, programming must involve considerable parallelism. This significantly reduces required computational time.

4.4.2 Best-Match Search, Pattern Recognition, and Content-Addressable Memory

Exact-match problems are intractable for GOFAI because they require exhaustive search strategies. As we have seen, for a minimally rational agent, this strategy is not realistic given real space-time constraints. Connectionist systems on the other hand, use pattern recognition heuristics and content-addressable memory to replace exhaustive search strategies. The application of content-addressable memory, in which a piece of data is located according to a property/characteristic of the object rather than its position in a list, demonstrates a blurring of the entire set of IO/IP; long-term/short-term; static/active dichotomies Cherniak presumes as part of the duplex memory structure. Information retrieval (memory recall) in a connectionist network, "...amounts to setting the values of some of the visible units and getting the problem to settle to the best interpretation (highest probability rank) of that input." (Rumelhart 1989, 222) In comparison, duplex memory and with it, the necessity of moving information from long

to short-term levels, becomes an unnecessary waste of valuable computing time. Ironically, if one were to take the duplex model and the PDP model and evaluate them according to the context-dependant CBA processes of Cherniak's own *Theory of Feasible Inferences* it would become clear that the connectionist model is the more feasible. That said, both connectionist and MR accounts nonetheless generate an explanation of why incomplete and /or uncertain input can result in adequate, albeit imperfect, information processing results. Nor is it coincidental that both accounts are in compliance with Cherniak's minimal inference conditions.

4.4.3 Automatic, Similarity-Based Generalizations

A common problem of both ideal rationality and GOFAI is brittleness, a characteristic by which deductive processes and traditional AI programs generate the correct output as long as there are no hidden variables or novel situations to integrate. The result for GOFAI is that traditional programs are not self-modifying and therefore cannot adapt to their environment. The result for ideal rationality theory is the discovery of human 'irrationality' in reasoning tests such as those described earlier by Kahneman and Tversky. The property of brittleness common to rule-based GOFAI and ideal rationality is a limitation of both approaches. Because Cherniak relies on the serial processing model, the rule-based IP approach advanced by Cherniak's memory model is limited to managing large explicit lists of data. Therefore the search capacity of traditional AI models depends entirely on whatever information organization scheme was chosen by the programmer, as well as inflexible, externally imposed rules.

But as we have seen in connectionist systems, "...knowledge is held in the

connections. Patterns represent the connection weights in such a way that similar pattern have similar effects. Therefore, similarity-based generalization emerges as an automatic property of connectionist models."(Rumelhart 1989, 223) This is closer to the way humans actually reason through everyday problems. The PDP attribute of automatic generalization also responds to the weakness of the DCC requirement of ideal rationality. Instead of requiring that an agent, given a particular belief-desire set, necessarily undertake 'all and only activities which are apparently appropriate', the agent as modeled by a connectionist system can reason in other than idealized rational ways. These methods, such as analogistic or similarity-based reasoning, are, thankfully, not infallible, since as we have seen, such "...quick and dirty shortcut strategies are required to avoid intractability." (Cherniak 1986, 95)

4.4.4 Learning

The development of the generalized delta rule described in the second chapter means that in PDP systems, adaptation in terms of changes on the structural level emerge as a result of spontaneous internal processing activity. But it is important to stress that, "Learning results do not guarantee that we can find a solution for all solvable problems." (Rumelhart 1989, 227) This parallels the minimally rational agent who, given a particular belief-desire set, undertakes some but not necessarily all, actions which are apparently appropriate. Like connectionism, minimal rationality tolerates complexity in exchange for *realistic* levels of rationality in believers.

4.4.5 Distributed Representations

Rumelhart claims that, "The ability of connectionist networks to learn leads to

the promise of computers that can literally learn their way around faulty components." (Rumelhart 1989, 227) This property is derived from the dynamic character of information in PDP systems as opposed to the static character of symbolic information representations of traditional AI models. In connectionism both memory and processing are distributed across the network. Intelligence is then an emergent property of a complex distribution system that arises from simultaneous local inter-actions. This means that "...patterns can be built up or torn down in bits and pieces, accounting for the graded nature of learning, and for the gradual patterns of breakdown that are typically displayed in brain-damaged individuals." (Bates and Elman 1993, 8/16) This is in contrast to the traditional information processing model which Cherniak uses in which memory and processing are distinct modules and memory is seen as a passive receptacle manipulated by an active processing module.

The static nature Cherniak's archival memory model is also puzzling given that his use of cluster concepts and context-dependancy so readily lends itself to distributed representations. Nevertheless, it seems that just as Quine continued to rely on language learning, Cherniak continually reverts to the rule-base symbolic representations implicit in his *Human Memory Theory*.

We have seen how elements of MR and connectionism overlap in terms of the essential role of error and context-dependancy within rationality theory, as well as the central function of heuristics in cognition. We have also demonstrated that four of the five properties of connectionist systems listed earlier overlap with elements of minimal

rationality. Yet because Cherniak's use of the duplex model of memory implies an acceptance of the symbolic representational construct used by traditional AI models as opposed to the distributed PDP model, he fails to optimize the full implications of MR. The suitability of a connectionist model rather than the static orthodox serial processing model of memory becomes apparent in light of the fifth property of PDP which MR does not instantiate – the notion of content defined as distributed representations.

As we have seen repeatedly, knowledge in connectionist systems is stored in the connections between processing units and therefore the configuration of the units linked by these connections is what constitutes content in a PDP network. Knowledge defined as 'mapping potential' is coded in distributed patterns rather than found in the firings of individual units. Because these potential mappings coexist across the same territory, they must 'compete' with each other to resolve a given set of inputs. Therefore, content is not acquired as if it were an archival *fonds* to be "...retrieved from some passive store nor is it placed in some localized buffer." (Bates and Elman 1993, 8/16) Instead, learning is a graded process in which solutions to cognitive problems emerge, a view which is compatible with the CBA analogy described in Cherniak's *Theory of Feasible Inferences* and the PDP description of learning as change at the micro-structural level.

It is therefore a mystery why, in his *Theory of Memory Structure*, Cherniak gives PDP short shrift. Although he gives tepid acknowledgment to the role of neuroscience when he suggests that "...not too much is lost by turning back to neuroanatomy and neurophysiology... Cognition is, after all, accomplished within a brain," (Cherniak 1988, 412) he still favours homuncular models based on the central

processing unit approach favored by GOFAI.[19] Cherniak maintains the duplex model complete with two distinct memory levels: a short-term memory for activated belief sets which uses information processing rules, and a long-term memory which stores inert belief sets and which is compartmentalized according to information organization rules defined by the agent. This maintenance of the duplex model requires a computational governor and by definition this is equivalent to an agent that is more than just a minimally rational agent.

The main purpose of Cherniak's duplex memory model is to account for departures from idealized rationality, which he asserts are based on long-term memory organization. But Cherniak's reliance on a homuncular account is unnecessary since these departures are adequately explained within the connectionist description of distributed representations by which meaning is not captured by a single symbolic unit, but rather arises from the interaction of a set of units in a network. This supports the compatibility of MR and connectionism and does so without resorting to homuncularity. First, in their embodiment of context-dependant processes, distributed representations reflect the holistic nature of knowledge which allows for the fallibilism of belief fixation. Second, in recognition of the realistic constraints of believers, the use of heuristics rather than algorithms is necessary for the isolation of feasible inferences which underlie knowledge in MR. Thus, once adopted by connectionism, MR embodies the fundamental

[19]Timothy Van Gelder points out that, "Homuncularity is a special kind of breakdown of a system into parts or components, each of which is responsible for a particular subtask." This implies a driver, or governor, for the system as a whole. Given that Cherniak maintains the traditional duplex model with its attendant symbolic content, the governor must be computational: "it literally computes... By manipulating symbols according to a schedule of rules." (Van Gelder 1996, 426.)

principle of a fully naturalized epistemology: that we ought to be fallibilistic about beliefs and realistic about believers.

At the same time, the drawbacks of content as symbolic representations are of a piece with the drawbacks of idealized rationality. First, symbolic representations are strongly propositional. "Thus, when this method of representation is used in no-language based applications such as image processing, it becomes difficult to explain many psychological findings. In contrast, distributed representations are well-suited to all modes of perception." (Eliasmith 1996, 1) A second drawback is that, like ideal deductive inferences, symbolic representations are 'all or nothing'. There is neither graded learning nor graceful degradation of content defined by symbolic representation. This is an extremely unrealistic model when applied to human cognitive function. In GOFAI minor damage to symbolic content causes the loss of entire concepts. However, the realistic result of damage to distributed representation is the loss of some accuracy of content but not the loss of the entire concept itself.

Therefore, although Cherniak's notion of minimal rationality sheds light on a long-standing bias of traditional philosophy and psychology toward an idealized standard of human rationality, his account in *Minimal Rationality* remains incomplete. While the *Theory of Feasible Inferences* describes a process which is of a piece with connectionist foundations and micro-level neuroanatomical models, the *Theory of Memory Structure* fails to adopt the appropriate technology and is conceptually inadequate to the task of completing the job. Forty years after Turing compared the finitude of machine intelligence with the limits of human intelligence Cherniak has shown that indeed, limits

do apply to the human intellect. However, what Cherniak seemingly fails to see — the difference in imagination and vision between his approach and Turing's — is that the limitations of both human and machine intelligence are not obstacles but clues to design solutions. It is in the response to these constraints that human intelligence develops and AI models evolve. Connectionist engineering incorporates the elements of our finetary predicament as design features of parallel distributed processing architectures, in effect turning obstacles into evolutionary advantages in much the same way that the heuristic strategy which emerges as a result of our finetary predicament allows us to use error to underwrite knowledge.

5. Conclusion: Knowledge Underwritten by Error

A brief re-iteration of the traditional definition of knowledge as justified true belief can now be used to highlight the alternative account of knowledge underwritten by error which has been developed in this thesis. We begin with the premise that two people can have opposite beliefs about *p* but by the law of non-contradiction, only one person has a *justified true* belief. That is, only one person *knows p*. The other person has a false belief or the mere opinion that-*p*. Evidence is what separates truth from mere opinion and is therefore the arbiter of truth. In this way, the addition of *logos*, or a justificatory account of the evidence, allows knowledge to go from true belief to *justified* true belief.

For justification to be compelling the evidence must be strong, a requirement that calls for a fit between our beliefs, an internal set of states, and the objects of our beliefs, the set of phenomena in the external world. Justification -- the strength of evidence -- is directly proportional to the tightness of the fit between my belief about what-is and what-*really*-is. In technical terms, that's the fit between my epistemic state and the ontologically-given world. But as we saw earlier, Quine's disolution of the analytic/synthetic distinction leads to the reciprocal containment of ontology and epistemology which in turn means that a holistic account of truth replaces a justificatory account. This holism also means that we are consigned to Neurath's boat, permanently adrift on a sea of fallibilistic knowledge. The assertion is then often made that in

naturalizing epistemology, Quine destroyed the very foundations of knowledge. For some, naturalized epistemology "...represents the gutting of philosophy. Banished outright or seriously degraded, are meanings, sense-data, ideas, knowledge, existence, and truth ...[but]... New beginnings are advertised in the naturalized sectors of the market." (Woods 1992, 570).

I would like to undercut the alarmist aspect of NE and support the 'new beginning' by suggesting a working definition of knowledge which reflects the integration of error in the PDP model, the fallibilism that characterizes belief in NE, and the realistic character of the finetary predicament of Cherniak's minimally rational agent. The difference between top-down and bottom-up processing models provides the background for this description. As we saw earlier, the traditional model of cognition is a tripartite structure comprised of a top-level which includes "an abstract formulation of what is being computed and why" (Marr 1977, 129), a middle-level which defines a particular rule for performing the process, and a lowest level account of how this rule is realized by the physical hardware. Marr's example of Fourier analysis illustrates how a top-down approach accounts for knowledge. At the top level of cognition we have knowledge of Fourier analysis. This can then be realized at mid-level by several different algorithms. Each algorithm can in turn be explained at the bottom-level by a variety of different architectures. Traditional epistemology and classical cognitive science share a common view of knowledge as cascading downward.[20] The content of this knowledge is comprised of objects derived from syntactically structured representations (symbols)

[20]One can only wonder: from where?

processed according to logico-manipulative operations (rules).

Contrast this with the connectionist model of knowledge which has neither symbols nor rules. As Andy Clark says, "connectionist models make do without stable computational objects (symbols)... [and] ... they do not operate by the application of (tacit or explicit) rules defined to apply to structures of such objects -- how could they, given that no such objects exist?" (Clark 1993a, 47) Knowledge on this account percolates upward from a microstructural base toward macro-level cognitive states. However, we are still left with the question of "...what constitutes the higher-level understanding of the processing one needs to really explain how a task is performed?" (Clark 1993a, 90) Clark's answer requires a kind of Copernican turn based on the bottom-up processing model which, "...effectively inverts the usual temporal and methodological order of explanation, much as Copernicus inverted the usual astronomical model of his day. In connectionist theory, the high-level understanding will be made to revolve around a working program which has learned how to negotiate some cognitive terrain. This inverts the classical ordering in which the high-level understanding comes first and guides the search for algorithms." (Clark 1993a, 50)

Learning how to negotiate this cognitive terrain means that we don't *know*; we can only *learn*. And we learn by trial and error. The constantly shifting content of learning is what we call 'knowledge'. It then follows that the concepts that constitute the necessarily dynamic character of knowledge in a fully naturalized epistemology are not the standard building blocks of thought. Rather, concepts are defined by their place within the overall activation pattern of thought and not vice versa. Furthermore, as we

also saw in chapters two and four, learning requires the capacity to generate adaptive heuristics to suit the occasion. Formalisms are then not only obsolete as the processes by which the conceptual building blocks become knowledge, they are antithetical to the production of knowledge because they negate the role of trial-and-error iterations necessary for learning. We have also seen that fallibilism underwrites knowledge in a fully naturalized epistemology, and we learned in chapter two that error underwrites knowledge on the microstructural level of connectionist architectures. Thus Clark makes the reasonable request that any system which can be said to grasp a concept must be capable of judging that a mistake has been made in a previous iteration or application of that concept. Clark calls this the *Requirement of Normative Depth*:

> The inner workings of an intentional system must be of a kind compatible with the description of that system as capable of making mistakes which involve the failure to respect those commitments in episodes of on-line processing.
> (Clark, 1993. 216)

It is the ability to recognize our own guiding epistemic principles and to judge our own judgements as correct or mistaken that allows us to extend our understanding of cognition to both descriptive and normative levels: to accurately describe a system as being capable of error is to ascribe normativity to that system.

This approach to knowledge is antithetical to a prevailing pre-occupation of epistemology centred on the relationship between knowledge and certainty, and its corollary, the relationship between knowledge and the impossibility of error. I have argued, on the contrary, that certainty and the impossibility of error are not defining features of knowledge. Instead, because most, if not all, of what is known has to be

learned, and uncertainty and error characterize both the input provided to the brain by the external world and the output of learning, the traditional definition of knowledge is both unrealistic and counterproductive. Instead, as I have argued, the integration of error according to a connectionist account of learning leads to a fully naturalized epistemology that provides a re-configuration of knowledge. Furthermore, PDP holds the most promise for integrating design solutions which will continue to allow us to explore the evolution of the brain on an environmental scale. Inevitably, such an exploration implies the need for a fundamental re-examination of the notions of rationality, certainty, and knowledge.

How we learn about the world in the face of error is isomorphic with how we construct the bridge between the evidence of nature and the hypotheses we posit to explain it. The first principle of design, be it in civil engineering or evolution, is that form follows function. The structure of knowledge should then reflect the function of knowledge. Therefore, like other higher order mammals, and *contra* Aristotle, humans do not naturally desire knowledge *per se*. Rather, our cognitive apparatus is designed by evolution to evaluate our environment in terms of risk. It is the impetus to avoid risk that precedes and informs the desire for knowledge. This function suggests that the cognitive engine comes equipped by evolutionary design to process partial and erroneous data in the production of knowledge.

The replacement of knowledge, traditionally typified by stable inner representational states, with a process-oriented and ability-based view of PDP, means that knowledge can no longer be seen as static but actively distributed. Learning then, "...appears as the continuous ontogenetic structural coupling of an organism to its

medium through a process that follows a direction determined by the selection exerted on its changes of structure by the implementation of the behaviour that it generates through the structure already selected in it by its previous plastic interactions." (Maturana 1978, 45) From a pragmatic point of view, knowledge serves to perpetuate species survival by the mitigation of risk and uncertainty for the purpose of evolutionary advantage. Thus, as Chris Stary and Markus Peschl state in *Android Epistemology*:

> The aim of generating knowledge is to correlate input and output activations in order to generate an adequate behaviour enabling the cognitive system to "fit" into its environment. This means to behave in such a way that it survives and does not act in such a way as to lead to the system's death... We think of this very basic form of knowledge as being the basis for higher level cognitive processes. (Stary and Peschl 1995, 195)

The pursuit of certainty has led to an abstraction away from the very body and the very world in which our brains evolved to guide us. (Clark 1997, xii) The reconfigured account of knowledge underwritten by error explored in this thesis represents a small step toward re-integration.

REFERENCES

Barrett, Robert B., and Roger F. Gibson. (1990). Perspectives on Quine. Cambridge MA: Basil Blackwell.

Bates, Elizabeth A., and Jeffrey Elman. (1996). "Connectionism and the Study of Change". In Brain Development and Cognition: A Reader. Mark Johnson, ed. 623-642. Oxford: Blackwell Publishers, 1993. http://crl.ucsd.edu/~elman/Papers/bates_elman/bates_elman.html. Last modified Sept. 14, 1997. 16 pages

Bechtel, William and Adele Abrahamsen. (1991). Connectionism and the Mind: An Introduction to Parallel Processing in Networks. Cambridge MA: Basil Blackwell.

Blackburn, Simon. (1994). Oxford Dictionary of Philosophy. Oxford: Oxford University Press. S.v. "Epistemology".

Boden, Margaret A. (1991)."Horses of a Different Color?" In Philosophy and Connectionist Theory. William Ramsey, Stephen P. Stich, and David E. Rumelhart, eds. 3-19 Hillsdale NJ: Erlbaum.

Casti, John L. (1989) Paradigms Lost: Tackling the Unanswered Mysteries of Modern Science. New York: Avon Books.

Cherniak, Christopher. (1986) Minimal Rationality. Cambridge MA: MIT Press.

_____. (1988). "Undebuggability and Cognitive Science". Communications of the ACM. 31:4 (April 1988) 402-412.

_____. (1994a). "Philosophy and Computational Neuroanatomy". Philosophical Studies. Vol. 73. 89-107.

_____. (1994b) A Companion to the Philosophy of Mind. S. Guttenplan, ed. Cambridge MA: Basil Blackwell. S.v. "Rationality". 526-531.

Chrisley, Ronald L. (1996). Learning in Non-superpositional Quantum Neurocomputers, http://www.cogs.susx.ac.uk/users/ronc/quantum3.quantum3.html Last update: Nov, 20, 1996.

Churchland, P.M. (1979)Scientific Realism and the Plasticity of Mind. New York: Cambridge University Press.

_____. (1989) Neurocomputational Perspective. Cambridge MA: MIT Press.

_____. (1990). "On the Nature of Theories: A Neurocomputational Perspective". In Scientific Theories. Minnesota Studies in the Philosophy of Science, Vol. XIV. 59-101. Minneapolis: University of Minnesota Press.

_____. (1996) The Engine of Reason, the Seat of the Soul. Cambridge MA: MIT Press.

Churchland, P.S. (1986). Neurophilosophy: Toward a Unified Science of the Mind/Brain. Cambridge MA: MIT Press.

_____. (1992). The Computational Brain. Cambridge MA: MIT Press.

Clark, Andy. (1993a) Associative Engines: Connectionism, Concepts, and Representational Change. Cambridge MA: MIT Press.

_____. (1993b). *Minimal Rationalism.* Mind 192: 408 (October) 587-610.

_____ (1997). Being There: Putting Brain, Body and World Together Again. Cambridge MA: MIT Press.

Crick, Francis. (1994). The Astonishing Hypothesis: The Scientific Search for the Soul. New York: Scribners.

Dyer, Michael G. (1991). "Connectionism vs. Symbolism". In Connectionism and the Philosophy of Mind. 382-416. Boston: Kluwer Academic Publishing.

Dyson, Freeman. (1999). "Miracles of Rare Devices". Sciences. 39:2 (March/April)

Eimas, Peter and Peter Galaburda. (1990) Neurobiology of Cognition. Cambridge MA: MIT Press.

Eliasmith, Chris. (1996) Dictionary of the Philosophy of Mind. S.v. "Distributed Representations". http://artsci.wustl.edu/~philos/MindDict/distributedrepresentation.html. 3 pages.

Floridi, Luciano. (1999) 1914-1945 Grounds For Knowledge. http://www.wolfson.ox.ac.uk/~floridi/cambridge.htm. 9 pages

Fodor, J.A. and Z. Pylyshyn. (1988). "Connectionism and cognitive architecture: a critical analysis". Cognition. 28: 3-71.

Garson, James W. (1991). "What Connectionists Cannot Do: the Threat to Classical AI". In Connectionism and the Philosophy of Mind. 113-142. Boston: Kluwer Academic Publishing.

_____. (1997) Stanford Encyclopedia of Philosophy. S.v. "Connectionism", http://plato.stanford.edu/entries/connectionism Last updated July 1997.

Gibson, Roger F. (1988) Enlightened Empiricism: An Examination of W.V. Quine's Theory of Knowledge. Tampa FL: University of South Florida Press.

Hahn, Lewis and Paul Schilpp, eds. (1986) Philosophy of W.V. Quine. Library of Living Philosophers, Vol. LVIII. LaSalle IL: Open Court.

Hinton, G.E., et al. (1986). "Distributed Representations". In Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol 1:Foundations. David E. Rumelhart, James L. McClelland, and the PDP Research Group. 77-109. Cambridge MA: MIT Press.

Horgan, Terence and John Tienson, eds. (1991) Connectionism and the Philosophy of Mind. Boston: Kluwer Academic.

Hooker, C.A. (1994). "Idealization, Naturalism, and Rationality: Some Lessons From Minimal Rationality". Synthese. 99:2 (May)

Hooker, C.A., H.B. Penfold, and R.J. Evans. (1992). "Control, Connectionism and Cognition: Towards a New Regulatory Paradigm". British Journal for the Philosophy of Science. 43:4 (Dec. 1992). 517-536.

Hume, David. (1748) "An Enquiry Concerning Human Understanding." Section IV, Part I. In Edwin A. Burtt, ed. English Philosophers From Bacon to Mill. New York: Modern Library, 1967

Kim, Jaegwon. (1988) "What is 'Naturalized Epistemology'?" In Naturalizing Epistemology. Hilary Kornblith, ed. Cambridge MA: MIT Press, 1994.

Koch, Christof. (1997). "Computation and the Single Neuron". Nature. Vol. 385 (January 16, 1997) 207-210

Kornblith, Hilary, ed. (1994) Naturalizing Epistemology, 2nd. Edition. Cambridge MA: MIT Press.

Marr, D. (1977). "Artificial Intelligence: A personal view". In Mind Design. J. Haugeland, ed. Cambridge MA: MIT Press.

Maturana, HR. (1978). "Biology of Language". In Psychology and Biology of Language. G.A. Miller and E. Lemeberg, eds. New York: Academic Press.

McClelland, J.L., D.E. Rumelhart, and G. E. Hinton. (1986). "The Appeal of Parallel Distributed Processing." In Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol 1: Foundations. David E. Rumelhart, James L. McClelland, and the PDP Research Group. 3-44. Cambridge MA: MIT Press.

Minsky, Marvin. (1991). "Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy". In Computation and Intelligence. George F. Luger, ed. 647-674. Menlo Park CA: AAAI Press and MIT Press, 1995.

Norman, D. A. (1986). "Reflections on Cognition and Parallel Distributed Processing." In Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models. James L. McClelland, David E. Rumelhart, and the PDP Research Group. 535-546. Cambridge MA: MIT Press.

Quine, W.V. (1953). "Two Dogmas of Empiricism". In From a Logical Point of View. 20-46. Cambridge MA: Harvard University Press.

_____. (1960). Word and Object. Cambridge MA: MIT Press.

_____. (1969a). "Epistemology Naturalized" In Naturalizing Epistemology, 2nd ed. Hilary Kornblith, ed. 15-31. Cambridge MA: MIT Press, 1994.

_____. (1969b). "Ontological Relativity". In Ontological Relativity and Other Essays. 26-28. New York: Columbia University Press.

_____. (1990). Pursuit of Truth. Cambridge MA: Harvard University Press.

_____. (1992). "Structure and Nature". Journal of Philosophy. 89:1.

_____. (1995). From Stimulus to Science. Cambridge MA: Harvard University Press.

Roth, Paul A. (1986). "Semantics Without Foundations". In Philosophy of W.V. Quine. Library of Living Philosophers, Vol. LVIII. Lewis Hahn and Paul Schilpp, eds. 533-458 LaSalle IL: Open Court.

Rumelhart, David E. (1989). "The Architecture of Mind: A Connectionist Approach". In Mind Design II. John Haugeland, ed. 203-232. Cambridge MA: MIT Press.

_____. (1992). "Towards a Microstructural Account of Human Reasoning". In Connectionism: Theory and Practice. 69-83. New York: Oxford University Press.

Rumelhart, David E., et al (1986a). "A General Framework for Parallel Distributed Processing". In Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol 1: Foundations. David E. Rumelhart, James L. McClelland, and the PDP Research Group. 45-76. Cambridge MA: MIT Press.

_____. (1986b). "Learning Internal Representations in Propagation of Error". In Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol 1: Foundations. David E. Rumelhart, James L. McClelland, and the PDP Research Group. 318-362. Cambridge MA: MIT Press.

Scheffler, Israel. (1967). The Anatomy of Inquiry: Philosophical Studies in the Theory of Science. New York: Alfred Knopf.

Stary, Chris and Markus F. Peschl. (1995). "Towards Constructivist Unification of Machine Learning and Parallel Distributed Processing." In Android Epistemology. Kenneth Ford, Clark Glymour, and Patrick J. Hayes, eds. Menlo Park CA: AAAI Press and MIT Press, 1995.

Stich, S.P. (1993). From Folk Psychology to Cognitive Science. Cambridge MA: MIT Press.

_____. (1996). "Deconstructing the Mind." In Deconstructing the Mind. 3-90. New York: Oxford University Press..

Stockmeyer, Larry J. and Ashok K. Chandra. (1979). "Intrinsically Difficult Problems." Scientific American. Vol. 240 (May 1979) 140-159.

Sutton, John. (1998) Philosophy and Memory Traces: Descartes to Connectionism. Cambridge: Cambridge University Press.

Turing, A.M. (1946). "ACE Reports of 1946 and Other Papers." B.E. Carpenter and R.W. Doran, eds. Cambridge MA: MIT Press.

Turing, A.M. (1950). "Computing Machinery and Intelligence." Mind. 59:236 (Oct. 1950) 433-460.

Tversky, Amos and Daniel Kahneman. (1974). "Judgement Under Uncertainty: Heuristics and Biases." Science. Vol. 185 (27 September, 1974) 1124-1131.

Van Gelder, Timothy. (1991). "Classical Questions, Radical Answers: Connectionism and Structure of Mental Representations." In Connectionism and the Philosophy of Mind. Terence Horgan and John Tienson, eds. 355-379. Boston: Kluwer Academic.

_____. (1996) "Dynamics and Cognition." In Mind Design II. John Haugeland, ed. 421-450. Cambridge MA: MIT Press.

Woods, John. (1992). "'Critical Notice' of W.V. Quine: Pursuit of Truth." Canadian Journal of Philosophy. 22:4. (December) 547-572.