16 ARTIFICIAL INTELLIGENCE AS EMPIRICAL ENQUIRY

Computer science is an empirical discipline. We would have called it an experimental science, but like astronomy, economics, and geology, some of its unique forms of observation and experience do not fit a narrow stereotype of the experimental method. Nonetheless, they are experiments. Each new machine that is built is an experiment. Actually constructing the machine poses a question to nature; and we listen for the answer by observing the machine in operation and analyzing it by all analytical and measurement means available. Each new program that is built is an experiment. It poses a question to nature, and its behavior offers clues to an answer. Neither machines nor programs are black boxes; they are artifacts that have been designed, both hardware and software, and we can open them up and look inside. We can relate their structure to their behavior and draw many lessons from a single experiment.

—A. NEWELL AND H. A. SIMON, *ACM Turing Award Lecture, 1976*

The study of thinking machines teaches us more about the brain than we can learn by introspective methods. Western man is externalizing himself in the form of gadgets.

—WILLIAM S. BURROUGHS, *Naked Lunch*

Where is the knowledge we have lost in information?

—T. S. ELIOT, *Choruses from the Rock*

16.0 Introduction

For many people, one of the most surprising aspects of work in artificial intelligence is the extent to which AI, and indeed much of computer science, turns out to be an empirical discipline. This is surprising because most people initially think of these fields in terms of their mathematical, or alternatively, their engineering foundations. From the mathematical

viewpoint, sometimes termed the "neat" perspective, there is the rationalist desire to bring standards of proof and analysis to the design of intelligent computational devices. From the engineering, or "scruffy" perspective, the task is often viewed as simply making successful artifacts that society wants to call "intelligent". Unfortunately, or fortunately, depending upon your philosophy, the complexity of intelligent software and the ambiguities inherent in its interactions with the worlds of human activity frustrate analysis from either the purely mathematical or purely engineering perspectives.

Furthermore, if artificial intelligence is to achieve the level of a science and become a critical component of the *science of intelligent systems*, a mixture of analytic and empirical methods must be included in the design, execution, and analysis of its artifacts. On this viewpoint each AI program can be seen as an experiment: it proposes a question to the natural world and the results are nature's response. Nature's response to our design and programmatic commitments shapes our understanding of formalism, mechanism, and finally, of the nature of intelligence itself (Newell and Simon 1976).

Unlike many of the more traditional studies of human cognition, we as designers of intelligent computer artifacts can inspect the internal workings of our "subjects". We can stop program execution, examine internal state, and modify structure at will. As Newell and Simon note, the structure of computers and their programs indicate their potential behavior: they may be examined, and their representations and search algorithms understood. The power of computers as tools for understanding intelligence is a product of this duality. Appropriately programmed computers are capable of both achieving levels of semantic and behavioral complexity that beg to be characterized in psychological terms as well as offer an opportunity for an inspection of their internal states that is largely denied scientists studying most other intellectual life forms.

Fortunately for continuing work in AI, as well as for establishing a science of intelligent systems, more modern psychological techniques, especially those related to neural physiology, have also shed new light on the many modes of human intelligence. We know now, for example, that human intelligent function is not monolithic and uniform. Rather it is modular and distributed. Its power is seen in the sense organs, such as the human retina, that can screen and preprocess visual information. Similarly, human learning is not a uniform and homogenous faculty. Rather learning is a function of multiple environments and differing systems, each adapted to achieve specialized goals. fMRI analysis, along with PET scans, EEG, and allied neural physical imaging procedures, all support a diverse and cooperative picture of the internal workings of actual intelligent systems.

If work in AI is going to reach the level of a science, we must also address important philosophical issues, especially those related to epistemology, or the question of how an intelligent system "knows" its world. These issues range from the question of what is the object of study of artificial intelligence to deeper issues, such as questioning the validity and utility of the physical symbol system hypothesis. Further questions include what a "symbol" is in the symbol system approach to AI and how symbols might relate to sets of weighted nodes in a connectionist model. We also question the role of rationalism expressed in the inductive bias seen in most learning programs and how this compares to the unfettered lack of structure often seen in unsupervised, reinforcement, and emergent approaches to learning. Finally, we must question the role of embodiment, situatedness, and sociological bias in problem solving. We conclude our discussion of philosophical

issues by proposing a *constructivist epistemology* that fits comfortably with both our commitment to AI as a science as well as to AI as empirical enquiry.

And so in this final chapter we return again to the questions asked in Chapter 1: What is intelligence? Can it be formalized? How can we build mechanisms that exhibit it? How can artificial and human intelligence fit into a larger context of a science of intelligent systems? In Section 16.1 we begin with a revised definition of artificial intelligence which shows how current work in AI, although rooted in the physical symbol system hypothesis of Newell and Simon, has extended both its tools, techniques, and inquiries into a much broader context. We explore these alternative approaches to the question of intelligence, and consider their powers for the design of intelligent machines and for being a component in the science of intelligence systems. In Section 16.2, we point out how many of the techniques of modern cognitive psychology, neuroscience, as well as epistemology may be used to better understand the artificial intelligence enterprise.

Finally, in Section 16.3, we discuss some of the challenges that remain both for modern AI practitioners as well as for epistemologists. For even though the traditional approaches to AI have often been guilty of a rationalist reductionism, new interdisciplinary insights and tools also have related shortcomings. For example, the creators of the genetic algorithm and the designers of a-life research define the world of intelligence from a Darwinian viewpoint: "What is, is what survives". Knowledge is also seen as "knowing how" rather than "knowing what" in a complex situated world. For the scientist, answers require explanations, and "success" or "survival" are not of themselves sufficient.

In this final chapter, we will discuss the future of AI by exploring the philosophical questions that must be addressed to create a computational science of intelligence. We conclude that AI's empirical methodology is an important tool, and perhaps one of the best available, for exploring the nature of intelligence.

16.1 Artificial Intelligence: A Revised Definition

16.1.1 Intelligence and the Physical Symbol System Hypothesis

Based on our experience of the last 15 chapters, we offer a revised definition of artificial intelligence:

> AI is the study of the mechanisms underlying intelligent behavior through the construction and evaluation of artifacts designed to enact those mechanisms.

On this definition, artificial intelligence is less a theory about the mechanisms underlying intelligence and more an empirical methodology for constructing and testing possible models for supporting such a theory. It is a commitment to the scientific method of designing, running, and evaluating experiments with the goal of model refinement and further experiment. Most importantly however, this definition, like the field of AI itself, directly attacks centuries of philosophical obscurantism about the nature of mind. It gives people who would understand what is perhaps our defining characteristic as humans an alternative to religion, superstition, Cartesian dualism, new-age placebos, or the search for

intelligence in some as yet undiscovered quirk of quantum mechanics (Penrose 1989). If the science supporting artificial intelligence has made any contribution to human knowledge, it is in confirming that intelligence is not some mystic vapor permeating men and angels, but rather the effect of a set of principles and mechanisms that can be understood and applied in the design of intelligent machines. It must be noted that our revised definition of AI does *not* define intelligence; rather it proposes a coherent role for *artificial* intelligence in exploring both the nature and expression of intelligent phenomena.

From an historical perspective, the dominant approach to artificial intelligence involved the construction of representational formalisms and their associated search-based reasoning mechanisms. The guiding principle of early AI methodology was the *physical symbol system* hypothesis, first articulated by Newell and Simon (1976). This hypothesis states:

> The necessary and sufficient condition for a physical system to exhibit general intelligent action is that it be a physical symbol system.
>
> *Sufficient* means that intelligence can be achieved by any appropriately organized physical symbol system.
>
> *Necessary* means that any agent that exhibits general intelligence must be an instance of a physical symbol system. The necessity of the physical symbol system hypothesis requires that any intelligent agent, whether human, space alien, or computer, achieve intelligence through the physical implementation of operations on symbol structures.
>
> *General intelligent action* means the same scope of action seen in human action. Within physical limits, the system exhibits behavior appropriate to its ends and adaptive to the demands of its environment.

Newell and Simon have summarized the arguments for the *necessity* as well as the *sufficiency* of this hypothesis (Newell and Simon 1976, Newell 1981, Simon 1981). In subsequent years both AI and cognitive science explored the territory delineated by this hypothesis.

The physical symbol system hypothesis has led to four significant methodological commitments: (a) the use of symbols and systems of symbols as a medium to describe the world; (b) the design of search mechanisms, especially heuristic search, to explore the space of potential inferences those symbol systems could support; and (c) the disembodiment of cognitive architecture, by which we mean it was assumed that an appropriately designed symbol system could provide a full causal account of intelligence, regardless of its medium of implementation. Finally (d), on this viewpoint, AI became empirical and constructivist: it attempted to understand intelligence by building working models of it.

On the symbol system view, tokens in a language, referred to as *symbols*, were used to denote or reference something other than themselves. Like verbal tokens in a natural language, symbols stood for or referred to things in an intelligent agent's world. Tarski (1956, Section 2.3) could offer the possibility of a *science of meaning* in these object–referent relationships.

Furthermore, AI's use of symbols goes beyond the questions addressed in a Tarskian semantics, extending symbols to represent all forms of knowledge, skill, intention, and causality. Such constructive efforts rely on the fact that symbols, together with their semantics, can be embedded in formal systems. These define a *representation language*. The ability to formalize symbolic models is essential to modeling intelligence as a running

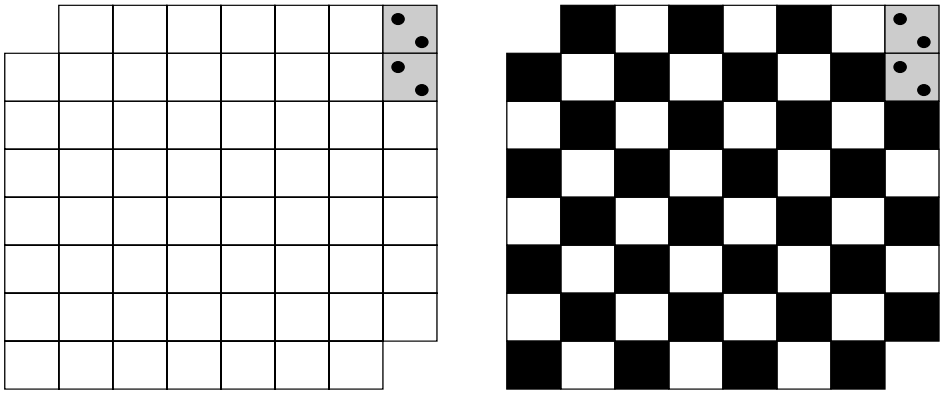

Figure 16.1 Truncated chessboard with two squares covered by a domino.

computer program. In previous chapters we studied several representations in detail: the predicate calculus, semantic networks, scripts, conceptual graphs, frames, and objects. The mathematics of formal systems allows us to argue such issues as soundness, completeness, and complexity, as well as discuss the organization of knowledge structures. The evolution of representational formalisms has allowed us to establish more complex (richer) semantic relationships. For example, inheritance systems constitute a semantic theory of taxonomic knowledge. By formally defining a class inheritance, such languages simplify the construction of intelligent programs and provide testable models of the organization of possible categories of intelligence itself.

Closely bound to representational schemata and their use in reasoning is the notion of search. Search became the step-by-step examination of problem states within a state space framework (an a priori semantic commitment) looking for solutions, subproblem goals, problem symmetries, or whatever aspect of the problem might be under consideration. Representation and search are linked because a commitment to a particular representation determines a space to be searched. Indeed, some problems can be made more difficult, or even impossible, through a poor choice of a representation language. The discussion of inductive bias later in this chapter illustrates this point.

A dramatic and often cited example of this interplay between search and representation as well as the difficulty of choosing an appropriate representation (can this process of optimizing the selection of representations be automated?) is the problem of placing dominos on a truncated chessboard. Assume that we have a chessboard and a set of dominos such that each domino will cover exactly two squares of the board. Also, assume that the board has some missing squares; in Figure 16.1 the upper left-hand corner and the lower right-hand corner have been removed.

The truncated chessboard problem asks whether there is a way of placing dominos on the board so that each square of the chessboard is covered and each domino covers exactly two squares.We might try to solve the problem by trying all placements of dominos on the board; this is the obvious search-based approach and is a natural consequence of representing the board as a simple matrix, ignoring such seemingly irrelevant features as the

color of the squares. The complexity of such a search is enormous, and would require heuristics for an efficient solution. For example, we might prune partial solutions that leave single squares isolated. We could also start by solving the problem for a smaller board, such as 2×2 and 3×3, and attempt to extend the solution to the 8×8 situation.

A more sophisticated solution, relying on a more complex representational scheme, notes that every placement of a domino must cover both a black and a white square. This truncated board has 32 black squares but only 30 white squares; thus the desired placement is not going to be possible. This raises a serious question for purely symbol-based reasoners: do we have representations that allow problem solvers to access knowledge with this degree of flexibility and creativity? How can a particular representation automatically change its structure as more is learned about a problem domain?

Heuristics are the third component, along with representation and search, of symbol-based AI. A heuristic is a mechanism for organizing search across the alternatives offered by a particular representation. Heuristics are designed to overcome the complexity of exhaustive search, the barrier to useful solutions for many classes of interesting problems. In computing, just as for humans, intelligence requires the informed choice of "what to do next". Throughout the history of AI research, heuristics have taken many forms.

The earliest problem-solving techniques, such as *hill climbing* in Samuel's checker-playing program (Section 4.1) or *means–ends analysis* in Newell, Shaw, and Simon's General Problem Solver (Section 13.1), came into AI from other disciplines, such as *operations research*, and have gradually matured into general techniques for AI problem solving. Search properties, including *admissibility, monotonicity*, and *informedness*, are important results from these early studies. These techniques are often referred to as *weak methods*. Weak methods were general search strategies intended to be applicable across entire classes of problem domains (Newell and Simon 1972, Ernst and Newell 1969). We saw these methods and their properties in Chapters 2, 3, 4, 6, and 14.

We introduced *strong methods* for AI problem solving with the rule-based expert system, model-based and case-based reasoning, and symbol-based learning of Chapters 8, 9, and 10. In contrast to weak problem solvers, strong methods focus on the information specific to each problem area, such as internal medicine or integral calculus, rather than on designing heuristic methods that generalize across problem areas. Strong methods underlie expert systems and other knowledge-intensive approaches to problem solving. Strong methods emphasize such issues as the amount of knowledge needed for problem solving, learning, and knowledge acquisition, the syntactic representation of knowledge, the management of uncertainty, and issues related to the quality of knowledge.

Why have we not built many truly intelligent symbol-based systems?

There are many criticisms that can be leveled at the *physical symbol system* characterization of intelligence. Most of these are captured by considering the issues of semantic meaning and the *grounding* of the symbols of an intelligent agent. The nature of "meaning", of course, also impacts the idea of intelligence as search through pre-interpreted symbol structures and the "utility" implicit in the use of heuristics. The notion of meaning in traditional AI is very weak, at best. Furthermore, the temptation of moving towards a more mathematics-based semantics, such as the Tarskian possible worlds approach, seems

wrong-headed. It reinforces the rationalist project of replacing the flexible and evolving intelligence of an embodied agent with a world where clear and distinct ideas are directly accessible.

The *grounding* of meaning is an issue that has forever frustrated both the proponents and critics of the AI and cognitive science enterprises. The grounding issue asks how symbols can have meaning. Searle (1980) makes just this point in his discussion of the so-called *Chinese Room*. Searle places himself in a room intended for translating Chinese sentences into English; there he receives a set of Chinese symbols, looks the symbols up in a large Chinese symbol cataloging system, and then outputs the appropriately linked sets of English symbols. Searle claims that although he himself knows absolutely no Chinese, his "system" can be seen as a Chinese-to-English translation machine.

There *is* a problem here. Although anyone who has worked in the research areas of machine translation or natural language understanding (Chapter 15), might argue that the Searle "translation machine", blindly linking one set of symbols to another set of symbols, would produce results of minimal quality, the fact remains that many current intelligent systems have a very limited ability to interpret sets of symbols in a "meaningful" fashion. This problem of too weak a supporting semantics also pervades many computationally based sensory modalities, whether they be visual, kinesthetic, or verbal.

In the areas of human language understanding, Lakoff and Johnson (1999) argue that the ability to create, use, exchange, and interpret meaning-symbols comes from a human's embodiment within an evolving social context. This context is physical, social, and right-now; it supports and enables the human ability to survive, evolve, and reproduce. It makes possible a world of analogical reasoning, the use and appreciation of humor, and the experiences of music and art. Our current generation of AI tools and techniques are very far away indeed from being able to encode and utilize any equivalent "meaning" system.

As a direct result of this weak semantic encoding, the traditional AI search/heuristic methodology explores states and contexts of states that are pre-interpreted. This means that an AI program's creator "imputes" or "lays on" to the symbols of the program various contexts of semantic meaning. A direct result of this pre-interpreted encoding is that intelligence-rich tasks, including learning and language, can only produce some computed function of that interpretation. Thus, many AI systems have very limited abilities to evolve new meaning associations as they explore their environments (Luger et al. 2002).

Finally, as a direct result of our current limited semantic modeling abilities, those applications where we are able to abstract away from a rich embodied and social context and at the same time capture the essential components of problem solving with pre-interpreted symbol systems are our most successful endeavors. Many of these were addressed throughout this book. However, even these areas remain brittle, without multiple interpretations, and with only limited ability to automatically recover from failures.

Across its brief history, the artificial intelligence research community has explored the ramifications of the physical symbol system hypothesis, and has developed its own challenges to that previously dominant view. As illustrated in the later chapters of this book, the explicit symbol system and search are not the only possible representational media for capturing intelligence. Models of computing based on the architecture of the animal brain as well as on the processes of biological evolution also provide useful frameworks for understanding intelligence in terms of scientifically knowable and empirically

reproducible processes. In the following sections of this final chapter we explore the ramifications of these approaches.

16.1.2 Connectionist or "Neural" Computing

A significant alternative to the physical symbol system hypothesis is the research into neural networks and other biologically inspired models of computing. Neural networks, for example, are computational and physically instantiated models of cognition not totally reliant on explicitly referenced and pre-interpreted symbols that characterize a world. Because the "knowledge" in a neural network is distributed across the structures of that network, it is often difficult, if not impossible, to isolate individual concepts to specific nodes and weights of the network. In fact, any portion of the network may be instrumental in the representation of different concepts. Consequently, neural networks offer a counterexample, at least to the *necessary* clause of the physical symbol system hypothesis.

Neural networks and genetic architectures shift the emphasis of AI away from the problems of symbolic representation and sound inference strategies to issues of learning and adaptation. Neural networks, like human beings and other animals, are mechanisms for adapting to the world: the structure of a trained neural network is shaped by learning, as much as by design. The intelligence of a neural network does not require that the world be recast as an explicit symbolic model. Rather, the network is shaped by its interactions with the world, reflected through the implicit traces of experience. This approach has made a number of contributions to our understanding of intelligence, giving us a plausible model of the mechanisms underlying the physical embodiment of mental processes, a more viable account of learning and development, a demonstration of the ability of simple, and local adaptations to shape a complex system in response to actual phenomena. Finally, they offer a powerful research tool for cognitive neuroscience.

Precisely because they are so different, neural nets can answer a number of questions that may be outside the expressive abilities of symbol-based AI. An important class of such questions concerns perception. Nature is not so generous as to deliver our perceptions to a processing system as neat bundles of predicate calculus expressions. Neural networks offer a model of how we might recognize "meaningful" patterns in the chaos of sensory stimuli.

Because of their distributed representation, neural networks are often more robust than their explicitly symbolic counterparts. A properly trained neural network can effectively categorize novel instances, exhibiting a human-like perception of similarity rather than strict logical necessity. Similarly, the loss of a few neurons need not seriously compromise the performance of a large neural network. This results from the often extensive redundancy inherent in network models.

Perhaps the most appealing aspect of connectionist networks is their ability to learn. Rather than attempting to construct a detailed symbolic model of the world, neural networks rely on the plasticity of their own structure to adapt directly to external experiences. They do not construct a model of the world so much as they are shaped by their experience within the world. Learning is one of the most important aspects of intelligence. It is also the problem of learning that raises some of the hardest questions for work in neural computing.

Why have we not built a brain?

In fact, the current generation of engineered connectionist systems bear very little resemblance to the human neuronal system! Because the topic of *neural plausibility* is a critical research issue, we begin with that question and then consider development and learning. Research in cognitive neuroscience (Squire and Kosslyn 1998, Gazzaniga 2000, Hugdahl and Davidson 2003) brings new insight to the understanding of human cognitive architecture. We describe briefly some findings and comment on how they relate to the AI enterprise. We consider issues from three levels: first, the neuron, second, the level of neural architecture, and finally we discuss cognitive representation or the *encoding* problem.

First, at the level of the individual neuron, Shephard (1998) and Carlson (1994) identify many different types of neuronal architectures for cells, each of which is specialized as to its function and role within the larger neuronal system. These types include *sensory receptor cells* typically found in the skin and passing input information to other cell structures, *interneurons* whose primary task is to communicate within cell clusters, *principle neurons* whose task is to communicate between cell clusters, and *motor neurons* whose task is system output.

Neural activity is electrical. Patterns of ion flows into and out of the neuron determine whether a neuron is active or resting. The typical neuron has a resting charge of −70mV. When a cell is active, certain chemicals are released from the axon terminal. These chemicals, called *neurotransmitters*, influence the postsynaptic membrane, typically by fitting into specific receptor sites, like a key into a lock, initiating further ion flows. Ion flows, when they achieve a critical level, about −50mV, produce an *action potential*, an all-or-none triggering mechanism indicating that the cell has fired. Thus neurons communicate through sequences of binary codes.

Postsynaptic changes from the action potential are of two sorts, *inhibitory*, found mainly in interneuron cell structures, or *excitatory.* These positive and negative potentials are constantly being generated throughout the synapses in the dendritic system. Whenever the net effect of all these events is to alter the membrane potentials of related neurons from −70mV to about −50mV, the threshold is crossed and massive ion flows are again initiated into those cells' axons.

Secondly, on the level of neural architecture, there are approximately 10^{10} total neurons in the cerebral cortex, a thin convoluted sheet covering the entire cerebral hemisphere. Much of the cortex is folded in on itself, increasing the total surface area. From the computational perspective we need to know not only the total number of synapses, but also the fan-in and fan-out parameters. Shephard (1998) estimates both these numbers to be about 10^5.

Finally, aside from the differences in the cells and architectures of neural and computer systems, there is a deep problem of cognitive representation. We are ignorant, for example, of how even simple memories are encoded in cortex. Of how, for example, a face is recognized, and how recognition of a face can link an agent to feelings of joy or sadness. We know a huge amount about the physical/chemical aspects of the brain, but relatively little about how the neural system encodes and uses "patterns" within its context.

One of the more difficult questions facing researchers, in both the neural and computing communities, is the role of innate knowledge in learning: can effective learning ever

occur on a *tabula rasa*, or *blank slate*, starting with no initial knowledge and learning entirely from experience? Or must learning start out with some prior inductive bias? Experience in the design of machine learning programs suggests that some sort of prior knowledge, usually expressed as an inductive bias, is necessary for learning in complex environments.

The ability of connectionist networks to converge on a meaningful generalization from a set of training data has proven sensitive to the number of artificial neurons, the network topology, and the specific learning algorithms used. Together, these factors constitute as strong an inductive bias as can be found in any symbolic representation. Research into human development supports this conclusion. There is increasing evidence, for example, that human infants inherit a range of "hard-wired" cognitive biases that enable learning of concept domains such as language and commonsense physics. Characterizing innate biases in neural networks is an active area of research (Elman et al. 1996).

The issue of innate biases becomes even more confounding when we consider more complex learning problems. For example, suppose we are developing a computational model of scientific discovery and want to model Copernicus' shift from a geocentric to heliocentric view of the universe. This requires that we represent both the Copernican and Ptolemaic views in a computer program. Although we could represent these views as patterns of activations in a neural network, our networks would tell us nothing about their behavior *as theories*. Instead, we prefer explanations such as "Copernicus was troubled by the complexity of the Ptolemaic system and preferred the simpler model of letting the planets revolve around the sun." Explanations such as this require symbols. Clearly, connectionist networks must be capable of supporting symbolic reasoning; after all, human beings are neural networks, and they seem to manipulate symbols tolerably well. Still, the neural foundation of symbolic reasoning is an important and open research problem.

Another problem is the role of development in learning. Human children cannot simply learn on the basis of available data. Their ability to learn in specific domains appears in well-defined developmental stages (Karmiloff-Smith 1992). An interesting question is whether this developmental progression is solely a result of human biology and embodiment, or whether it reflects some logically necessary limits on the ability of an intelligence to learn invariances in its world. Could developmental stages function as a mechanism for decomposing the problem of learning about the world into more manageable subproblems? Might a series of artificially imposed developmental restrictions provide artificial networks with a necessary framework for learning about a complex world?

The application of neural networks to practical problems raises a number of additional research issues. The very properties of neural networks that make them so appealing, such as adaptability and robustness in light of missing or ambiguous data, also create problems for their practical application. Because networks are trained, rather than programmed, behavior is difficult to predict. There are few guidelines for designing networks that will converge properly in a given problem domain. Finally, explanations of *why* a network arrived at a particular conclusion are often difficult to construct and may take the form of a statistical argument. These are all areas of current research.

One can ask then, whether connectionist networks and more symbolic AI are that different as models of intelligence. They both share a number of important commonalities, especially that intelligence is ultimately encoded as computation and has fundamental and

formal limits, such as the Church/Turing hypothesis (Luger 1994, Chapter 2). Both approaches also offer models of mind shaped by application to practical problems. Most importantly, however, both approaches deny philosophical dualism and place the foundations of intelligence in the structure and function of physically realized devices.

We believe that a full reconciliation of these two very different approaches to capturing intelligence is inevitable. When it is accomplished, a theory of how symbols may reduce to patterns in a network and, in turn influence future adaptation of that network, will be an extraordinary contribution. This will support a number of developments, such as integrating network-based perceptual and knowledge-intensive reasoning facilities into a single intelligence. In the meantime, however, both research communities have considerable work to do, and we see no reason why they should not continue to coexist. For those who feel uncomfortable with two seemingly incommensurable models of intelligence, even physics functions well with the intuitively contradictory notion that light is sometimes best understood as a wave and sometimes as a particle, although both viewpoints may well be subsumed by string theory (Norton 1999).

16.1.3 Agents, Emergence, and Intelligence

Agent-based computation and modular theories of cognition raise another set of interesting issues for researchers building artificial intelligences. One important school of thought in cognitive science holds that the mind is organized into sets of specialized functional units (Minsky 1985, Fodor 1983). These modules are specialists and employ a range of innate structures and functions, from "hard-wired" problem solving to inductive biases, that account for the diversity of problems they, as practical agents, must address. This makes sense: how can a single neural network or other system be trained to handle functions as diverse as perception, motor control, memory, and higher-level reasoning? Modular theories of intelligence provide both a framework for answering these questions and a direction for continued research into issues such as the nature of innate biases in individual modules as well as mechanisms of module interaction.

Genetic and emergent models of computation offer one of the newest and most exciting approaches to understanding both human and artificial intelligence. By demonstrating that globally intelligent behavior can arise from the cooperation of large numbers of restricted, independent, and individual agents, genetic and emergent theories view complex results through the interrelationships of relatively simple structures.

In an example from Holland (1995), the mechanisms that keep a large city such as New York supplied with bread demonstrate the fundamental processes underlying the emergence of intelligence in an agent-based system. It is unlikely that we could write a centralized planner that would successfully supply New Yorkers with the rich variety of daily breads to which they are accustomed. Indeed, the Communist world's unfortunate experiment with central planning revealed the limitations of such approaches! However, in spite of the practical difficulties of writing a centralized planning algorithm that will keep New York supplied with bread, the loosely coordinated efforts of the city's many bakers, truckers, suppliers of raw materials, as well as its retailers, solve the problem quite nicely. As in all agent-based emergent systems, there is no central plan. No one baker has

more than a very limited knowledge of the city's bread requirements; each baker simply tries to optimize his or her own business opportunities. The solution to the global problem emerges from the collective activities of these independent and local agents.

By demonstrating how highly goal-directed, robust, nearly optimal behaviors can arise from the interactions of local individual agents, these models provide yet another answer to old philosophical questions of the origins of mind. The central lesson of emergent approaches to intelligence is that full intelligence can and does arise from the interactions of many simple, individual, local, and embodied agent intelligences.

The second major feature of emergent models is their reliance on Darwinian selection as the basic mechanism that shapes the behavior of the individual agents. In the bakery example, it seems that each individual baker does not behave in a manner that is, in some sense, globally optimal. Rather, the source of their optimality is not of a central design; it is the simple fact that bakers who do a poor job of satisfying the needs of their local customers generally fail. It is through the tireless, persistent operations of these selective pressures that individual bakers arrive at the behaviors that lead to their individual survival as well as to a useful emergent collective behavior.

The combination of a distributed, agent-based architecture and the adaptive pressures of natural selection are a powerful model of the origins and operations of mind. Evolutionary psychologists (Cosmides and Tooby 1992, 1994; Barkow et al. 1992) have provided a model of the way in which natural selection has shaped the development of the innate structure and biases in the human mind. The basis of evolutionary psychology is a view of the mind as highly modular, as a system of interacting, highly specialized agents. Indeed, discussions of evolutionary psychology often compare the mind to a Swiss army knife, a collection of specialized tools that can be applied to solving different problems.

There is increasing evidence that human minds are, indeed, highly modular. Fodor (1983) offers a philosophical argument for the modular structure of mind. Minsky (1985) explores the ramifications of modular theories for artificial intelligence. This architecture is important to theories of the evolution of mind. It would be difficult to imagine how evolution could shape a single system as complex as a mind. It is, however, plausible that evolution, working over millions of years, could successively shape individual, specialized cognitive skills. As evolution of the brain continued, it could also work on combinations of modules, forming the mechanisms that enable the modules to interact, to share information, and to cooperate to perform increasingly complex cognitive tasks (Mithen 1996).

Theories of neuronal selection (Edelman 1992) show how these same processes can account for the adaptation of the individual neural system. Neural Darwinism models the adaptation of neural systems in Darwinian terms: the strengthening of particular circuits in the brain and the weakening of others is a process of selection in response to the world. In contrast to symbolic learning methods, which attempt to extract information from training data and use that information to build models of the world, theories of neuronal selection examine the effect of selective pressures on populations of neurons and their interactions. Edelman (1992, page 81) states:

> In considering brain science as a science of recognition I am implying that recognition is not an instructive process. No direct information transfer occurs, just as none occurs in evolutionary or immune processes. Instead, recognition is selective.

Agent technologies offer models of social cooperation as well. Using agent-based approaches, economists have constructed informative (if not completely predictive) models of economic markets. Agent technologies have exerted an increasing influence on the design of distributed computing systems, the construction of internet search tools and implementation of cooperative work environments.

Finally, agent-based models have exerted an influence on theories of consciousness. For example, Daniel Dennett (1991, 2006) has based an account of the function and structure of consciousness on an agent architecture of mind. He begins by arguing that it is incorrect to ask where consciousness is located in the mind/brain. Instead, his *multiple draft theory of consciousness* focuses on the role of consciousness in the interactions of agents in a distributed mental architecture. In the course of perception, motor control, problem solving, learning, and other mental activities, we form coalitions of interacting agents. These coalitions are highly dynamic, changing in response to the needs of different situations. Consciousness, for Dennett, serves as a binding mechanism for these coalitions, supporting agent interaction and raising critical coalitions of interacting agents to the foreground of cognitive processing.

What issues limit an agent-based approximation of intelligence?

Agent-based and "emergent" approaches have opened up a number of problems that must be solved if their promise is to be realized. For example, we have yet to fill in all the steps that have enabled the evolution of higher-level cognitive abilities such as language. Like paleontologists' efforts to reconstruct the evolution of species, tracing the development of these higher level problems will take a great deal of additional detailed work. We must both enumerate the agents that underlie the architecture of mind and trace their evolution across time.

Another important problem for agent-based theories is in explaining the interactions between modules. Although the "Swiss army knife" model of mind is a useful intuition builder, the modules that compose mind are not as independent as the blades of a pocket knife. Minds exhibit extensive, highly fluid interactions between cognitive domains: we can talk about things we see, indicating an interaction between visual and linguistic modules. We can construct buildings that enable a specific social purpose, indicating an interaction between technical and social intelligence. Poets can construct tactile metaphors for visual scenes, indicating a fluid interaction between visual and tactile modules. Defining the representations and processes that enable these inter-module interactions is an active area of research (Karmiloff-Smith 1992, Mithen 1996, Lakoff and Johnson 1999).

Practical applications of agent-based technologies are also becoming increasingly important. Using agent-based computer simulations, it is possible to model complex systems that have no closed-form mathematical description, and were heretofore impossible to study in this detail. Simulation-based techniques have been applied to a range of phenomena, such as the adaptation of the human immune system and the control of complex processes, including particle accelerators, the behavior of global currency markets, and the study of weather systems. The representational and computational issues that must be solved to implement such simulations continue to drive research in knowledge representations, algorithms, and even the design of computer hardware.

Further practical problems that agent architectures must deal with include protocols for inter-agent communication, especially when local agents often have limited knowledge of the problem at large or indeed of what knowledge other agents might already possess. Furthermore, few algorithms exist for the decomposition of larger problems into coherent agent-oriented subproblems, or indeed how limited resources might be distributed among agents. These and other agent-related issues were presented in Section 7.4.

Perhaps the most exciting aspect of emergent theories of mind is their potential for placing mental activities within a unified model of the emergence of order from chaos. Even the brief overview provided in this section has cited work using emergent theories to model a range of processes, from the evolution of the brain over time, to the forces that enable learning in individuals, to the construction of economic and social models of behavior. There is something extraordinarily appealing in the notion that the same processes of emergent order as shaped by Darwinian processes can explain intelligent behavior at a variety of resolutions, from the interactions of individual neurons, to the shaping of the modular structure of the brain, to the functioning of economic markets and social systems. It may be that intelligence has a fractal geometry, where the same emerging processes appear at whatever level of resolution we view the system at large.

16.1.4 Probabilistic Models and the Stochastic Technology

As early as the 1950s, stochastic techniques were used to address the understanding and generation of natural language expressions. Claude Shannon (1948) applied probabilistic models, including discrete Markov chains, to the task of language processing. Shannon (1951) also borrowed the notion of entropy from thermodynamics as a way of measuring the information capacity of a message. Also about this time Bell Labs created the first statistical system able to recognize the ten digits, 0, ..., 9, as spoken by an individual speaker. It functioned at 97-99% accuracy (Davis et al. 1952, Jurasky and Martin 2008).

Through the 1960s and 1970s Bayesian approaches to reasoning continued very much in the background of AI research activity. Natural language technology explored many of the symbol based approaches described in Section 7.1. Although many expert systems, for example MYCIN, created their own "certainty factor algebras" as seen in Section 9.2, several, including PROSPECTOR, took the Bayesian approach (Duda et al. 1979a). The complexity of such systems quickly become intractable, however. As we pointed out in Section 9.3, the full use of Bayes' rule for a realistic sized medical diagnosis program of 200 diseases and 2000 symptoms would require the collection and integration of eight hundred million pieces of information.

In the late 1980s Judea Pearl (1988) offered a computationally tractable model for diagnostic reasoning in the context of causal relationships within a problem domain: Bayesian belief networks. BBNs relax two constraints of the full Bayesian model. First, an implicit "causality" is assumed in stochastic inference; that is, reasoning goes from cause to effect and is not circular, i.e., an effect cannot circle back to cause itself. This supports representing BBNs as a directed acyclic graph (Sections 3.1, 9.3). Second, BBNs assume the direct parents of a node supports full causal influence on that node. All other nodes are assumed to be conditionally independent or have influence small enough to be ignored.

Pearl's research (1988, 2000) renewed interest in stochastic approaches for modeling the world. As we saw in Section 9.3, BBNs offered a very powerful representational tool for diagnostic (abductive) inference. This is true especially with the dynamic nature of both human and stochastic systems: as the world changes across time, our understanding is enriched: some causes turn out to explain more of what we see while other potential causes are "explained away". Research in the design of stochastic systems as well as their supporting inference schemes is really only in its infancy.

The late 1980s also saw new research energy applied to issues of natural language processing. As noted in Section 15.4, these stochastic approaches included new parsing, tagging, and many other techniques for disambiguating language expressions. A full range of these approaches may be found in books on speech recognition and tasks in language processing (Manning and Schutz 1999, Jurasky and Martin 2008).

With the renewed interest and successes in the stochastic approaches to characterizing intelligent behavior, a person can quite naturally wonder what its limitations might be.

Is intelligence fundamentally stochastic?

There is a tremendous attraction towards a stochastic viewpoint for agent interactions in a changing world. Many might argue that a human's "representational system" is fundamentally stochastic, i.e., conditioned by the world of perceptions and causes in which it is immersed. Certainly the behaviorist/empiricist viewpoint would find this conjecture attractive. Situated and embedded action theorists might go even further and hypothesize that the conditioned relationships an agent has with its physical and social environment offer a sufficient explanation of the successful agent's accommodation with that world. Some modern researchers are even claiming that aspects of neural function are fundamentally stochastic (Hawkins 2004, Doya et al. 2007).

Let's consider language. One of the strengths of oral and written expression, as Chomsky and others have pointed out, is its *generative* nature. This means that, within the set of vocabulary and language forms available, new and previously unexperienced expressions naturally occur. This happens both on the level of creating novel sentences as well as with individual words, verbizing, for instance: "Google it!". A stochastic account of language must demonstrate this generative power. Currently, the limitations of collected language information, whether treebanks or other data copora, can radically restrict use of the stochastic technology. This is because the collected information must offer an appropriate setting (or prior) for interpreting the current novel situation.

Stochastic models of application domains, for an aircraft engine or transmission system say, can have similar limitations. There, of necessity, will always be *closed world* or *minimal model* assumptions, Section 9.1, for realistically complex system. However, dynamic Bayesian networks are now showing promise for interpreting system state over time. But there still are limited abilities to predict novel situations across time.

On a higher explanatory level, it may be difficult for a probabilistic model to account for a shift out of its own explanatory system, or paradigm. In what sense, as noted in Section 16.1.2, can a stochastic model possibly explain theories or higher-level rearrangements of conceptual views, that is, re-evaluate issues related to the adequacy of the model itself with, perhaps, the need to shift to different viewpoints? These topics remain

important research issues and constraints on the stochastic approaches to understanding uncertainty.

But, the fact remains that, often without explicit instruction, agents "make" quite successful models. As we see from a constructivist viewpoint, Section 16.2, *some* model is a sine qua non for an agent to understand its world, i.e., if there is no a priori commitment to what the world is "about", phenomena are neither perceived nor understood!

Besides the philosophical arguments just presented, there are a number of practical limitations to the use of stochastic systems. The current generation of BBNs are propositional in nature. It has been only recently possible to create general laws or relationships such as $\forall X\ \text{male}(X) \rightarrow \text{smart}(X)$ with an attached distribution. Furthermore, it is desired that BBNs include recursive relationships, especially for time-series analysis. Research to develop first-order stochastic representational systems, addressing these issues of general, is important for continuing research. As noted in Section 13.3, first-order and Turing-complete stochastic modeling tools are now available (Pless et al. 2006). It is also important to explore the application of these general stochastic models to neuro/psychological data, where they have recently been applied (Burge et al. 2007, Doya et al. 2007).

We next discuss those psychological and philosophical aspects of human intelligence that impact the creation, deployment, and evaluation of an artificial intelligence.

16.2 The Science of Intelligent Systems

It is not a coincidence that a major subgroup of the artificial intelligence community has focused its research on understanding *human* intelligence. Humans provide the prototypical examples of intelligent activity, and AI engineers, even though they are usually *not* committed to "making programs that act like humans", seldom ignore human solutions. Some applications such as diagnostic reasoning are often deliberately modeled on the problem solving processes of human experts working in that area. Even more importantly, understanding human intelligence is a fascinating and open scientific challenge in itself.

Modern *cognitive science*, or *the science of intelligent systems* (Luger 1994), began with the advent of the digital computer, even though, as we saw in Chapter 1, there were many intellectual forebears of this discipline, from Aristotle through Descartes and Boole, to more modern theorists such as Turing, McCulloch and Pitts, the founders of the *neural net* model, and John von Neumann, an early proponent of a-life. The study became a science, however, with the ability to design and run experiments based on these theoretical notions, and to an important extent, this came about with the arrival of the computer. Finally, we must ask, "Is there an all inclusive science of intelligence?" We can further ask, "Can a science of intelligent systems support construction of artificial intelligences?"

In the following sections we discuss briefly how the psychological, epistemological, and sociological sciences support research and development in AI.

16.2.1 Psychological Constraints

Early research in cognitive science examined human solutions to logic problems, simple games, planning, and concept learning (Feigenbaum and Feldman 1963, Newell and

Simon 1972, Simon 1981). Coincident with their work on the Logic Theorist, Section 14.1, Newell and Simon began to compare their computational approaches with the search strategies used by human subjects. Their data consisted of *think-aloud protocols*, descriptions by human subjects of their thoughts during the process of devising a problem solution, such as a logic proof. Newell and Simon then compared these protocols with the behavior of the computer program solving the same problem. The researchers found remarkable similarities and interesting differences across both problems and subjects.

These early projects established the methodology that the discipline of *cognitive science* would employ during the following decades:

1. Based on data from humans solving particular classes of problems, design a representational scheme and related search strategy for solving the problem.
2. Run the computer-based model to produce a trace of its solution behavior.
3. Observe human subjects working on these same problems and keep track of measurable parameters of their solution process, such as those found in think-aloud protocols, eye movements, and written partial results.
4. Analyze and compare the human and computer solutions.
5. Revise the computer model for the next round of tests and comparisons with the human subjects.

This empirical methodology is described in Newell and Simon's Turing award lecture, quoted at the beginning of this chapter. An important aspect of cognitive science is the use of experiments to validate a problem-solving architecture, whether it be a production system, connectionist, emergent, or an architecture based on the interaction of distributed agents.

In recent years, an entirely new dimension has been added to this paradigm. Now, not just programs can be deconstructed and observed in the act of problem solving, but humans and other life forms can be as well. A number of new imaging techniques have been included in the tools available for observing cortical activity. These include *magnetoencephalography* (MEG), which detects the magnetic fields generated by populations of neurons. Unlike the electrical potentials generated by these populations, the magnetic field is not smeared by the skull and scalp, and thus a much greater resolution is possible.

A second imaging technology is *positron emission tomography*, or PET. A radioactive substance, typically O^{15} is injected into the bloodstream. When a particular region of the brain is active, more of this agent passes by sensitive detectors than when the region is at rest. Comparison of resting and active images can potentially reveal functional localization at a resolution of about 1cm (see Stytz and Frieder 1990).

Another technique for neural analysis is *functional magnetic resonance imaging*, or fMRI. This approach has emerged from more standard structured imaging based on nuclear magnetic resonance (NMR). Like PET, this approach compares resting with active neuronal states to reveal functional localization.

A further contribution to the localization of brain function, with an important link to the imaging techniques just mentioned, is software algorithms developed by Barak Pearlmutter and his colleagues (Pearlmutter and Parra 1997, Tang et al. 1999, 2000a, 2000b).

These researchers are able to take the complex noise patterns often seen as the output of various neural imaging techniques and break them into their separate components. This is an essential step in analysis, as the patterns of normal steady state existence, such as eye movements, breathing, and heartbeats, are interwoven with the other neuron firing patterns we want to understand.

The result of research in cognitive neuroscience (Squire and Kosslyn 1998, Shephard 1998, Gazzaniga 2000) has added greatly to our understanding of the neural components involved in intelligent activity. Even though an analysis and critique of these results is beyond the scope of this book, we list and reference several important issues:

In the area of perception and attention there is the *binding problem*. Researchers such as Anne Triesman (1993, 1998) and Jeff Hawkins (2004) note that perceptual representations depend on distributed neural codes for relating the parts and properties of objects to each other, and ask what mechanism is needed to "bind" the information relating to each object and to distinguish that object from others.

In the area of visual search, what neural mechanisms support the perception of objects embedded in large complex scenes? Some experiments show that the suppression of information from irrelevant objects plays an important role in the selection of search targets (Luck 1998). Furthermore how do we "learn" to see (Sagi and Tanne 1998)?

In the area of plasticity in perception, Gilbert (1992, 1998) contends that what we see is not strictly a reflection of the physical characteristics of a scene, but rather is highly dependent on the processes by which our brain attempts to interpret that scene.

How does the cortical system represent and index temporally related information, including interpretation of perceptions and production of motor activity (Ivry 1998)?

In memory studies, stress hormones released during emotionally arousing situations modulate memory processes (Cahill and McGaugh 1998). This relates to the *grounding* problem: how are thoughts, words, perceptions meaningful to an agent? In what sense can there possibly be a "sadness (or tears) in things," the *lacremae rerum* of Virgil?

The acoustic-phonetic aspects of speech provide important organizing principles for linking neuroscience research to cognitive and linguistic theories (Miller et al. 1998). How are syntactic and semantic components of cortex integrated (Gazzaniga 2000)?

How does an individual acquire a specific language and what neurophysiological stages support this development (Kuhl 1993, 1998)?

How is development understood, what is *critical period plasticity*, and the adult reorganizations seen in mammalian somatosensory systems (O'Leary et al. 1999)? Are developmental stages critical for "building" intelligence? See Karmiloff-Smith (1992) and Gazzaniga (2000) for further discussion.

The practice of artificial intelligence certainly does not require extensive knowledge of these and related neuro/psychological domains. However this type of knowledge can support the engineering of intelligent artifacts as well as help locate research and

development in AI within the context of the larger science of intelligence systems. Finally, the creation of a psychological, neurophysiological, and computational synthesis is truly exciting. But this requires a mature epistemology, which we discuss next.

16.2.2 Epistemological Issues

If you don't know where you are going, you will wind up somewhere else. . .

—YOGI BERRA (attributed)

The development of artificial intelligence has been shaped by a number of important challenges and questions. Natural language understanding, planning, reasoning in uncertain situations, and machine learning are all typical of those types of problems that capture some essential aspect of intelligent behavior. More importantly, intelligent systems operating in each of these domains require knowledge of purpose, practice, and performance in situated and socially embedded contexts. To better understand these issues, we examine the *epistemological commitment* of a program that is intended to be "intelligent".

The epistemological commitment reflects both the semantics supporting symbol use as well as the structures of symbols employed. The task in these situations is to discover and exploit the *invariances* existing in a problem domain. "Invariant" is a term used to describe the regularities or significant manipulable aspects of complex environments. In the present discussion, the terms *symbols* and *symbol systems* are used generically, from the explicit symbols of the Newell and Simon (1976) tradition, to the nodes and network architecture of a connectionist system, to the emergent tokens of genetic and artificial life. Although the points we make next are general across most of AI, we will focus our discussion on issues in machine learning, as we have created multiple examples and algorithms for learning throughout this book.

In spite of progress in machine learning, it remains one of the most difficult problems facing artificial intelligence. There are three issues limiting our current understanding and research progress: first, the problem of *generalization and overlearning*, second, the role of *inductive bias* in learning, and third, the *empiricist's dilemma* or addressing the idea of constraint-free learning. The last two problems are related: the implicit inductive bias of many learning algorithms is an expression of the rationalists' problem of being biased by expectations, that is, what we learn often seems to be a direct function of what we expect to learn. From the opposite viewpoint, as we saw in a-life research, where there are very few *a priori* expectations of what is to be learned, is it really sufficient to say, "Build it and it will happen"? According the attribution to Yogi Berra at the beginning of this section, it probably won't! The next sections briefly address these issues.

The generalization problem

The examples we used to introduce the various learning models–symbol-based, connectionist, and emergent–were usually very constrained. For example, connectionist architectures often contained only a few nodes or one partial hidden layer. This is appropriate in

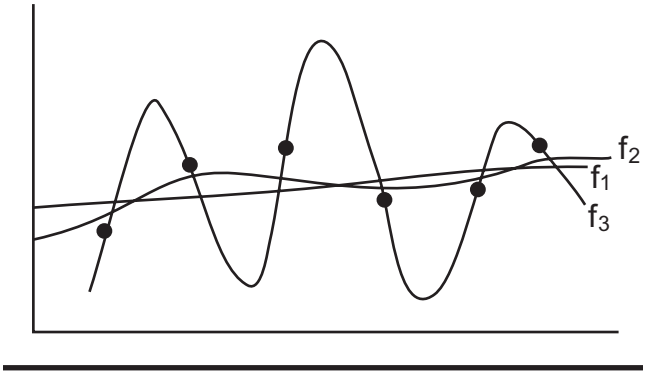

Figure 16.2 A set of data points and three function approximations.

that the main learning laws can be adequately explained in the context of a few neurons and partial layers. It can be very misleading in that neural net applications are usually considerably larger and that the problem of scale IS important. For instance, for backpropagation learning, a large number of training examples and larger networks are generally required to solve problems of any significant practical interest. Many researchers comment extensively on the matter of selecting appropriate numbers of input values, the ratio between input parameters and hidden nodes, and the training trials necessary before convergence can be expected (Hecht-Nielsen 1990, Zurada 1992, Freeman and Skapura 1991). In fact, the quality and quantity of training data are important issues for any learning algorithm. Without an appropriate bias or extensive built-in knowledge, a learning algorithm can be totally misled attempting to find patterns in noisy, sparse, or even bad data. Other than acknowledging that these are difficult, important, and open issues, this "engineering" aspect of learning is not addressed in our book.

A related problem is the issue of "sufficiency" in learning. When can we say our algorithms are sufficient for capturing the important constraints or invariants of a problem domain? Do we reserve a portion of our original data to test our learning algorithms? Does the amount of data we have relate to the quality of learning? Perhaps the sufficiency judgement is heuristic or aesthetic: we humans often see our algorithms as "good enough".

Let us illustrate this generalization problem with an example, using a form of backpropagation to induce a general function from a set of data points. Figure 16.2 might represent data points we are asking our algorithm to generalize. The lines across this set of points represent functions induced by a learning algorithm. Remember that once the algorithm is trained we will want to offer it new data points and have the algorithm produce a good generalization for these data also.

The induced function f_1 might represent a fairly accurate least mean squares fit. With further training the system might produce f_2, which seems a fairly "good" fit to the set of data points; but still, f_2 does not exactly capture the data points. Further training can produce functions that exactly fit the data but may offer terrible generalizations for further input data. This phenomena is referred to as *overtraining* a network. One of the strengths

of backpropagation learning is that in many application domains it is known to produce effective generalizations, that is, functional approximations which fit the training data well *and also* handle new data adequately. However, identifying the point where a network passes from an undertrained to an overtrained state is nontrivial. It is naive to think that one can present a neural network, or for that matter any other learning tool, with raw data and then simply step aside and watch while it produces the most effective and useful generalizations for addressing new similar problems.

We conclude by placing this generalization issue back into its epistemological context. When problem solvers make and utilize representations (whether symbols, nodes of networks, or whatever) for the solution process, they are creating invariants, and most probably systems of invariants, for exploring the problem/solution domain and producing related generalizations. This very viewpoint brought to the problem solving process biases the eventual success of the endeavour. In the next subsection, we address this issue further.

Inductive bias, the rationalist's a priori

The automated learning of Chapters 10 to 13, and for that matter most AI techniques, reflected the a priori biases of their creators. The problem of inductive bias is that the resulting representations and search strategies offer a medium for encoding an already interpreted world. They rarely offer mechanisms for questioning our interpretations, generating new viewpoints, or for backtracking and changing perspectives when they are unproductive. This implicit bias leads to the rationalist epistemological trap of seeing in the world exactly and only what we expect or are trained to see.

The role of inductive bias must be made explicit in each learning paradigm. Furthermore, just because no inductive bias is acknowledged, doesn't mean it does not exist and critically effect the parameters of learning. In symbol-based learning the inductive bias is usually obvious, for example, using a semantic net for concept learning. In Winston's (1975*a,* Section 10.1) learning algorithms, biases include the conjunctive relationship representation and the importance of using "near misses" for constraint refinement. We see similar biases in the use of particular predicates for version space search, decision trees in ID3, Section 10.3, or even rules for Meta-DENDRAL, Section 10.5.

As we have intimated throughout Chapters 10 through 13, however, many aspects of connectionist and genetic learning also assume an inductive bias. For instance, the limitations of perceptron networks led to the introduction of hidden nodes. We may well ask what contribution the hidden nodes make in solution generation. One way of understanding the role of hidden nodes is that they add dimensions to the representation space. As a simple example, we saw in Section 11.3.3 that the data points for the *exclusive-or* problem were not linearly separable in two dimensions. The learned weight on the hidden node, however, provides another dimension to the representation. In three dimensions, the points are separable using a two-dimensional plane. Given the two dimensions of the input space and the hidden node, the output layer of this network can then be seen as an ordinary perceptron, which is finding a plane that separates the points in a three-dimensioned space.

A complementary perspective is that many of the "different" learning paradigms shared (sometimes not obvious) common inductive biases. We pointed out many of these: the relationship between clustering with CLUSTER/2 in Section 10.5, the perceptron in

Section 11.2, and prototype networks in Section 11.3. We noted that counterpropagation, the coupled network that uses unsupervised competitive learning on the Kohonen layer together with supervised Hebbian learning on the Grossberg layer, is in many ways similar to backpropagation learning. In counterpropagation, clustered data on the Kohonen layer plays a role similar to the generalizations learned by the hidden nodes that use backpropagation.

In many important ways, the tools we presented are similar. In fact, even the discovery of prototypes representing clusters of data offer the complementary case to function approximation. In the first situation, we are attempting to classify sets of data; in the second, we are generating functions that explicitly divide data clusters from each other. We saw this when the minimum distance classification algorithm used by the perceptron also gave the parameters defining the linear separation of the data.

Even the generalizations that produce functions can be seen from many different viewpoints. Statistical techniques, for example, have for a long time been able to discover data correlations. Iterative expansion of Taylor series can be used to approximate most functions. Polynomial approximation algorithms have been used for over a century to approximate functions for fitting data points.

To summarize, the commitments made within a learning scheme, whether symbol-based, connectionist, emergent, or stochastic to a very large extent mediate the results we can expect from the problem solving effort. When we appreciate this synergistic effect throughout the process of the design of computational problem solvers we can often improve our chances of success as well as interpret our results more insightfully.

The empiricist's dilemma

If current approaches to machine learning, especially supervised learning, possess a dominant inductive bias, unsupervised learning, including many of the genetic and evolutionary approaches, has to grapple with the opposite problem, sometimes called *the empiricist's dilemma*. Themes of these research areas include: solutions will emerge, alternatives are evolved, populations reflects the survival of the fittest. This is powerful stuff, especially situated in the context of parallel and distributed search power. But there is a problem: How can we know we are someplace when we are not sure where we are going?

Plato, more than 2000 years ago, posed this problem in the words of the slave Meno:

> And how can you enquire, Socrates, into that which you do not already know? What will you put forth as the subject of the enquiry? And if you find out what you want, how will you ever know that this is what you did not know (Plato 1961)?

Several researchers have demonstrated that Meno was correct, see Mitchell (1997) and the *No Free Lunch* theorems of Wolpert and Macready (1995). The empiricist, in fact, does require the remnants of a rationalist's a priori to save the science!

Nonetheless, there remains great excitement about unsupervised and evolutionary models of learning; for example, in creating networks based on exemplars or energy minimization, which can be seen as fixed-point attractors or basins for complex relational invariances. We watch as data points "settle" toward attractors and are tempted to see

these new architectures as tools for modeling dynamic phenomena. What, we might ask, are the limits of computation in these paradigms?

In fact, researchers have shown (Siegelman and Sontag 1991) that recurrent networks are computationally complete, that is, equivalent to the class of Turing Machines. This Turing equivalence extends earlier results: Kolmogorov (1957) showed that for any continuous function there exists a neural network that computes that function. It has also been shown that a one hidden-layer backpropagation network can approximate any of a more restricted class of continuous functions (Hecht-Nielsen 1989). Similarly, we saw in Section 11.3, that von Neumann created finite state automata that were Turing complete. Thus connectionist networks and finite state automata appear to be but two more classes of algorithms capable of computing virtually any computable function. Furthermore, inductive biases DO apply to unsupervised as well as genetic and emergent models of learning; representational biases apply to the design of nodes, networks, and genomes, and algorithmic biases apply to the search, reward, and selection operators.

What is it then, that unsupervised learners, whether connectionist, genetic, or evolving finite state machines in their various forms, can offer?

1. One of the most attractive features of connectionist learning is that most models are data or example driven. That is, even though their architectures are explicitly designed, they learn by example, generalizing from data in a particular problem domain. But the question still arises as to whether the data is sufficient or clean enough not to perturb the solution process. And how can the designer know?
2. Genetic algorithms also support a powerful and flexible search of a problem space. Genetic search is driven both by the diversity enforced by mutation as well as by operators such as crossover and inversion, that preserve important aspects of parental information for succeeding generations. How can the program designer preserve and nurture this diversity/preservation trade-off?
3. Genetic algorithms and connectionist architectures may be viewed as instances of parallel and asynchronous processing. Do they indeed provide results through parallel asynchronous effort not possible with explicit sequential programming?
4. Even though the neural and sociological inspiration is not important for many modern practitioners of connectionist and genetic learning, these techniques do reflect many important aspects of natural evolution and selection. We saw models for error reduction learning with perceptron, backpropagation, and Hebbian models. We also saw the autoassociative Hopfield nets in Section 11.3.4. Various models of evolution were reflected in the paradigms of Chapter 12.
5. Finally, all learning paradigms are tools for empirical enquiry. As we capture the invariants of our world, are our tools sufficiently powerful and expressive to ask further questions related to the nature of perception, learning, and understanding?

In the next section we propose that a constructivist epistemology, coupled with the experimental methods of modern artificial intelligence, offer the tools and techniques for continuing the exploration of a science of intelligent systems.

The constructivist's rapprochement

Theories are like nets: he who casts, captures. . .

—L. Wittgenstein

Constructivists hypothesize that all understanding is the result of an interaction between energy patterns in the world and mental categories imposed on the world by the intelligent agent (Piaget 1954, 1970; von Glasersfeld 1978). Using Piaget's descriptions, we *assimilate* external phenomena according to our current understanding and *accommodate* our understanding to the "demands" of the phenomena.

Constructivists often use the term *schemata* to describe the *a priori* structure used to organize experience of the external world. This term is taken from the British psychologist Bartlett (1932) and its philosophical roots go back to Kant (1781/1964). On this viewpoint, observation is not passive and neutral but active and interpretive.

Perceived information, Kant's *a posteriori* knowledge, never fits precisely into our preconceived, and *a priori*, schemata. From this tension, the schema-based biases the subject uses to organize experience are either modified or replaced. The need for accommodation in the face of unsuccessful interactions with the environment drives a process of cognitive equilibration. Thus, the constructivist epistemology is fundamentally one of cognitive evolution and refinement. An important consequence of constructivism is that the interpretation of any situation involves the imposition of the observer's concepts and categories on reality (an inductive bias).

When Piaget (1954, 1970) proposed a constructivist approach to understanding, he called it *genetic epistemology.* The lack of a comfortable fit of current schemata to the world "as it is" creates a cognitive tension. This tension drives a process of schema revision. Schema revision, Piaget's *accommodation*, is the continued evolution of an agent's understanding towards *equilibration*.

Schema revision and continued movement toward equilibration is a genetic predisposition of an agent for an accommodation to the structures of society and the world. It combines both these forces and represents an embodied predisposition for survival. Schema modification is both an *a priori* result of our genetics as well as an *a posteriori* function of society and the world. It reflects the embodiment of a survival-driven agent, of a being in space and time.

There is a blending here of the empiricist and rationalist traditions, mediated by the goals of agent survival. As embodied, agents can comprehend nothing except that which first passes through their senses. As accommodating, agents survive through learning the general patterns of an external world. What is perceived is mediated by what is expected; what is expected is influenced by what is perceived: these two functions can only be understood in terms of each other. In this sense stochastic models - both Bayesian and Markovian - *are* appropriate, as prior experience conditions current interpretations (Doya et al. 2007).

Finally, we, as agents, are seldom consciously aware of the schemata that support our interactions with the world. As the sources of bias and prejudice both in science and society, we are more often than not unaware of a priori schemata. These are constitutive of our

equilibration with the world and not (usually) a perceptible element of a conscious mental life.

Finally, why is a constructivist epistemology particularly useful in addressing the problems of understanding intelligence? How can an agent within an environment understand its own understanding of that situation? We believe constructivism also addresses the *epistemological access* problem in both philosophy and psychology. For more than a century there has been a struggle in both these disciplines between two factions, the positivist, which proposes to infer mental phenomena from observable physical behavior, and a more phenomenological approach which allows the use of first person reporting to access cognitive phenomena. This factionalism exists because both modes of access to psychological phenomena require some form of model construction and inference.

In comparison to physical objects like chairs and doors, which often, naively, seem to be directly accessible, the mental states and dispositions of an agent seem to be particularly difficult to characterize. In fact, we contend that this dichotomy between the direct access to physical phenomena and the indirect access to the mental is illusory. The constructivist analysis suggests that no experience of things is possible without the use of some model or schema for organizing that experience. In scientific inquiry, as well as in our normal human experiences, this implies that *all* access to phenomena is through exploration, approximation, and continued model refinement.

So what is the project of the AI practitioner?

As AI practitioners, we are constructivists. We build, test, and refine models. But what are we approximating in our model-building? We discuss this issue in the following paragraphs but first make an epistemological observation: Rather than trying to capture the essence of "things outside" us, the AI problem-solver is best served by attempting to emulate the model building, refining, and equilibration heuristics of the intelligent agent itself.

Only the extreme solipsist (or the mentally challenged) would deny the "reality" of an extra-subject world. But what is this so called "real world"? Besides being a complex combination of "hard things" and "soft things", it is also a system of atoms, molecules, quarks, gravity, relativity, cells, DNA, and (perhaps even) superstrings. For all these concepts are but exploratory models driven by the explanatory requirements of equalibration-driven agents. Again, these exploratory models are not about an external world. Rather, they capture the dynamic equilibrating tensions of the intelligent and social agent, of a material intelligence evolving and continually calibrating itself within space and time.

But access to and creation of "the real" is also achieved through agent commitment. An embodied agent *creates the real* through an existential affirmation that a perceived model of its expectations is good-enough for addressing some of its practical needs and purposes. This act of commitment grounds the symbols and systems of symbols the agent uses in its material and social contexts. These constructs are grounded because they are affirmed as good-enough for achieving aspects of its purpose. This grounding is also seen in agent language use. Searle (1969) is correct in his notion of speech phenomena as *acts*. This *grounding* issue is part of why computers have fundamental problems with expressions of intelligence, including their demonstrations of language and learning. What disposition might a computer be given that affords it appropriate purposes and goals?

Although Dennett (1987) would impute grounding to a computer solving problems requiring and using "intelligence", the lack of *sufficient* grounding is easily seen in the computer's simplifications, brittleness, and often limited appreciation of context.

The use and grounding of symbols by animate agents implies even more. The particulars of the human agent's embodiment and social contexts mediate its interactions with its world. Auditory and visual systems sensitive to a particular bandwidth; viewing the world as an erect biped, having arms, legs, hands; being in a world with weather, seasons, sun, and darkness; part of a society with evolving goals and purposes; an individual that is born, reproduces, and dies: these are critical components that support metaphors of understanding, learning, and language; these mediate our comprehension of art, life, and love.

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date...
Shakespeare Sonnet XVIII

We conclude with a final summary of critical issues that both support and delimit our efforts at creating a science of intelligent systems.

16.3 AI: Current Challenges and Future Directions

like the geometer who strives
to square the circle and cannot find
by thinking the principle needed,

was I at that new sight. . .

—DANTE, *Paradiso*

Although the use of AI techniques to solve practical problems has demonstrated its utility, the use of these techniques to found a general science of intelligence is a difficult and continuing problem. In this final section we return to the questions that led us to enter the field of artificial intelligence and to write this book: is it possible to give a formal, computational account of the processes that enable intelligence?

The computational characterization of intelligence begins with the abstract specification of computational devices. Research through the 1930s, 40s, and 50s began this task, with Turing, Post, Markov, and Church all contributing formalisms that describe computation. The goal of this research was not just to specify what it meant to compute, but rather to specify limits on what could be computed. The Universal Turing Machine (Turing 1950) is the most commonly studied specification, although Post's rewrite rules, the basis for production system computing (Post 1943), is also an important contribution. Church's model (1941), based on partially recursive functions, offers support for modern high-level functional languages, such as Scheme, Ocaml, and Standard ML.

Theoreticians have proven that all of these formalisms have equivalent computational power in that any function computable by one is computable by the others. In fact, it is possible to show that the universal Turing machine is equivalent to any modern computational device. Based on these results, the Church–Turing hypothesis makes the even stronger argument: that no model of computation can be defined which is more powerful than these known models. Once we establish equivalence of computational specifications, we have freed ourselves from the medium of mechanizing these specifications: we can implement our algorithms with vacuum tubes, silicon, protoplasm, or tinker toys. The automated design in one medium can be seen as equivalent to mechanisms in another. This makes the empirical enquiry method even more critical, as we experiment in one medium to test our understanding of mechanisms implemented in another.

One of the possibilities is that the universal machine of Turing and Post may be too general. Paradoxically, intelligence may require a less powerful computational mechanism with more focused control. Levesque and Brachman (1985) have suggested that human intelligence may require more computationally efficient (although less expressive) representations, such as Horn clauses for reasoning, the restriction of factual knowledge to ground literals, and the use of computationally tractable truth maintenance systems. Agent-based and emergent models of intelligence also seem to espouse this philosophy.

Another point addressed by the formal equivalence of our models of mechanism is the duality issue and the mind–body problem. At least since the days of Descartes (Section 1.1), philosophers have asked the question of the interaction and integration of mind, consciousness, and a physical body. Philosophers have offered every possible response, from total materialism to the denial of material existence, even to the supporting intervention of a benign god! AI and cognitive science research reject Cartesian dualism in favor of a material model of mind based on the physical implementation or instantiation of symbols, the formal specification of computational mechanisms for manipulating those symbols, the equivalence of representational paradigms, and the mechanization of knowledge and skill in embodied models. The success of this research is an indication of the validity of this model (Johnson-Laird 1988, Dennett 1987, Luger et al. 2002.

Many consequential questions remain, however, within the epistemological foundations for intelligence in a physical system. We summarize, for a final time, several of these critical issues.

1. **The representation problem.** Newell and Simon hypothesized that the physical symbol system and search are necessary and sufficient characterizations of intelligence (see Section 16.1). Are the successes of the neural or sub-symbolic models and of the genetic and emergent approaches to intelligence refutations of the physical symbol hypothesis, or are they simply other instances of it?

   Even a weak interpretation of this hypothesis—that the physical symbol system is a *sufficient* model for intelligence—has produced many powerful and useful results in the modern field of cognitive science. What this argues is that we can implement physical symbol systems that will demonstrate intelligent behavior. Sufficiency allows creation and testing of symbol-based models for many aspects of human performance (Pylyshyn 1984, Posner 1989). But the strong interpretation—that the physical symbol system and search are *necessary* for

intelligent activity—remains open to question (Searle 1980, Weizenbaum 1976, Winograd and Flores 1986, Dreyfus and Dreyfus 1985, Penrose 1989).

2. **The role of embodiment in cognition.** One of the main assumptions of the physical symbol system hypothesis is that the particular instantiation of a physical symbol system is irrelevant to its performance; all that matters is its formal structure. This has been challenged by a number of thinkers (Searle 1980, Johnson 1987, Agre and Chapman 1987, Brooks 1989, Varela et al. 1993) who essentially argue that the requirements of intelligent action in the world require a physical embodiment that allows the agent to be fully integrated into that world. The architecture of modern computers does not support this degree of situatedness, requiring that an artificial intelligence interact with its world through the extremely limited window of contemporary input/output devices. If this challenge is correct, then, although some form of machine intelligence may be possible, it will require a very different interface than that afforded by contemporary computers. (For further comments on this topic see Section 15.0, issues in natural language understanding, and Section 16.2.2, on epistemological constraints.)

3. **Culture and intelligence.** Traditionally, artificial intelligence has focused on the individual mind as the sole source of intelligence; we have acted as if an explanation of the way the brain encodes and manipulates knowledge would be a complete explanation of the origins of intelligence. However, we could also argue that knowledge is best regarded as a social, rather than as an individual construct. In a *meme-based* theory of intelligence (Edelman 1992), society itself carries essential components of intelligence. It is possible that an understanding of the social context of knowledge and human behavior is just as important to a theory of intelligence as an understanding of the dynamics of the individual mind/brain.

4. **Characterizing the nature of interpretation.** Most computational models in the representational tradition work with an already interpreted domain: that is, there is an implicit and *a priori* commitment of the system's designers to an interpretive context. Under this commitment there is little ability to shift contexts, goals, or representations as the problem solving evolves. Currently, there is little effort at illuminating the process by which humans construct interpretations.

   The Tarskian view of semantics as a mapping between symbols and objects in a domain of discourse is certainly too weak and doesn't explain, for example, the fact that one domain may have different interpretations in the light of different practical goals. Linguists have tried to remedy the limitations of Tarskian semantics by adding a theory of pragmatics (Austin 1962). Discourse analysis, with its fundamental dependence on symbol use in context, has dealt with these issues in recent years. The problem, however, is broader in that it deals with the failure of referential tools in general (Lave 1988, Grosz and Sidner 1990).

   The semiotic tradition started by C. S. Peirce (1958) and continued by Eco, Seboek, and others (Eco 1976, Grice 1975, Sebeok 1985) takes a more radical approach to language. It places symbolic expressions within the wider context of signs and sign interpretation. This suggests that the meaning of a symbol can

only be understood in the context of its role as interpretant, that is, in the context of an interpretation and interaction with the environment (see Section 16.2.2).

5. **Representational indeterminacy.** Anderson's representational indeterminacy conjecture (Anderson 1978) suggests that it may in principle be impossible to determine what representational scheme best approximates the human problem solver in the context of a particular act of skilled performance. This conjecture is founded on the fact that every representational scheme is inextricably linked to a larger computational architecture, as well as search strategies. In the detailed analysis of human skill, it may be impossible to control the process sufficiently so that we can determine the representation; or establish a representation to the point where a process might be uniquely determined. As with the uncertainty principle of physics, where phenomena can be altered by the very process of measuring them, this is an important concern for constructing models of intelligence but need not limit their utility.

   But more importantly, the same criticisms can be leveled at the computational model itself where the inductive biases of symbol and search in the context of the Church-Turing hypothesis still under constrain a system. The perceived need of some optimal representational scheme may well be the remnant of a rationalist's dream, while the scientist simply requires models sufficiently robust to constrain empirical questions. The proof of the quality of a model is in its ability to offer an interpretation, to predict, and to be revised.

6. **The necessity of designing computational models that are falsifiable.** Popper (1959) and others have argued that scientific theories must be falsifiable. This means that there must exist circumstances under which the model is *not* a successful approximation of the phenomenon. The obvious reason for this is that *any* number of confirming experimental instances are not sufficient for confirmation of a model. Furthermore, much new research is done in direct response to the failure of existing theories.

   The general nature of the physical symbol system hypothesis as well as situated and emergent models of intelligence may make them impossible to falsify and therefore of limited use as models. The same criticism can be made of the conjectures of the phenomenological tradition (see point 7). Some AI data structures, such as the semantic network, are so general that they can model almost anything describable, or as with the universal Turing machine, any computable function. Thus, when an AI researcher or cognitive scientist is asked under what conditions his or her model for intelligence will *not* work, the answer can be difficult.

7. **The limitations of the scientific method.** A number of researchers (Winograd and Flores 1986, Weizenbaum 1976) claim that the most important aspects of intelligence are not and, in principle, cannot be modeled, and in particular not with any symbolic representation. These areas include learning, understanding natural language, and the production of speech acts. These issues have deep roots in our philosophical tradition. Winograd and Flores's criticisms, for example, are based on issues raised in phenomenology (Husserl 1970, Heidegger 1962).

Most of the assumptions of modern AI can trace their roots back from Carnap, Frege, and Leibniz through Hobbes, Locke, and Hume to Aristotle. This tradition argues that intelligent processes conform to universal laws and are, in principle, understandable.

Heidegger and his followers represent an alternative approach to understanding intelligence. For Heidegger, reflective awareness is founded in a world of embodied experience (a life-world). This position, shared by Winograd and Flores, Dreyfus, and others, argues that a person's understanding of things is rooted in the practical activity of "using" them in coping with the everyday world. This world is essentially a context of socially organized roles and purposes. This context, and human functioning within it, is not something explained by propositions and understood by theorems. It is rather a flow that shapes and is itself continuously created. In a fundamental sense, human expertise is not knowing *that*, but rather, within a world of evolving social norms and implicit purposes, knowing *how*. We are inherently unable to place our knowledge and most of our intelligent behavior into language, either formal or natural.

Let us consider this point of view. First, as a criticism of the *pure* rationalist tradition, it is correct. Rationalism asserts that all human activity, intelligence, and responsibility can, in principle at least, be represented, formalized, and understood. Most reflective people do not believe this to be the case, reserving important roles for emotion, self-affirmation and responsible commitment (at least!). Aristotle himself said, in his *Essay on Rational Action*, "Why is it that I don't feel compelled to perform that which is entailed?" There are many human activities outside the realms of science that play an essential role in responsible human interaction; these cannot be reproduced by or abrogated to machines.

This being said, however, the scientific tradition of examining data, constructing models, running experiments, and examining results with model refinement for further experiments has brought an important level of understanding, explanation, and ability to predict to the human community. The scientific method is a powerful tool for increasing human understanding. Nonetheless, there remain a number of caveats to this approach that scientists must understand.

First, scientists must not confuse the model with the phenomenon being modeled. The model allows us to progressively approximate the phenomenon: there will, of necessity, always be a "residue" that is not empirically explained. In this sense also representational indeterminacy is *not* an issue. A model is used to explore, explain, and predict; and if it allows scientists to accomplish this, it is successful (Kuhn 1962). Indeed, different models may successfully explain different aspects of a phenomenon, such as the wave and particle theories of light.

Furthermore, when researchers claim that aspects of intelligent phenomena are outside the scope and methods of the scientific tradition, this statement itself can only be verified by using that very tradition. The scientific method is the only tool we have for explaining in what sense issues may still be outside our current understanding. Every viewpoint, even that from the phenomenological tradition, if it is to have any meaning, must relate to our current notions of explanation - even to be coherent about the extent to which phenomena cannot be explained.

The most exciting aspect of work in artificial intelligence is that to be coherent and contribute to the endeavor we must address these issues. To understand problem solving, learning, and language we must comprehend the philosophical level of representations and knowledge. In a humbling way we are asked to resolve Aristotle's tension between *theoria* and *praxis*, to fashion a union of understanding and practice, of the theoretical and practical, to live between science and art.

AI practicioners are tool makers. Our representations, algorithms, and languages are tools for designing and building mechanisms that exhibit intelligent behavior. Through experiment we test both their computational adequacy for solving problems as well as our own understanding of intelligent phenomena.

Indeed, we have a tradition of this: Descartes, Leibniz, Bacon, Pascal, Hobbes, Babbage, Turing, and others whose contributions were presented in Chapter 1. Engineering, science, and philosophy; the nature of ideas, knowledge, and skill; the power and limitations of formalism and mechanism; these are the limitations and tensions with which we must live and from which we continue our explorations.

16.4 Epilogue and References

We refer the reader to the references at the end of Chapter 1 and add *Computation and Cognition* (Pylyshyn 1984) and *Understanding Computers and Cognition* (Winograd and Flores 1986). For issues in cognitive science see Newell and Simon (1972), Pylyshyn (1973, 1980), Norman (1981), Churchland (1986), Posner (1989), Luger (1994), Franklin (1995), Ballard (1997), Elman et al. (1996), and Jeannerod (1997).

Haugeland (1981, 1997), Dennett (1978, 2006) and Smith (1996) describe the philosophical foundations of a science of intelligent systems. Anderson's (1990) book on cognitive psychology offers valuable examples of information processing models. Pylyshyn (1984) and Anderson (1978, 1982, 1983a) give detailed descriptions of many critical issues in cognitive science, including a discussion of representational indeterminacy. Dennett (1991) applies the methodology of cognitive science to an exploration of the structure of consciousness itself. We also recommend books on the philosophy of science (Popper 1959, Kuhn 1962, Bechtel 1988, Hempel 1965, Lakatos 1976, Quine 1963).

Finally, *Philosophy in the Flesh*, Lakoff and Johnson (1999) suggest possible answers to the grounding problem. *The Embodied Mind* (Varela et al.1993), Suchman (1987), and *Being There* (Clark 1997) describe aspects of embodiment that support intelligence.

We leave the reader with address information for two important groups:

The Association for the Advancement of Artificial Intelligence
445 Burgess Drive, Menlo Park, CA 94025

Computer Professionals for Social Responsibility
P.O. Box 717, Palo Alto, CA 94301