DÆDALUS

JOURNAL OF THE AMERICAN ACADEMY OF ARTS AND SCIENCES

Artificial Intelligence

Winter 1988

Issued as Volume 117, Number 1, of the *Proceedings of the American Academy of Arts and Sciences*

Winter 1988, "Artificial Intelligence." Issued as Volume 117, Number 1, of the Proceedings of the American Academy of Arts and Sciences.

© 1988 by the American Academy of Arts and Sciences.

Library of Congress Catalog Number 12-30299.

Editorial Offices: Dædalus, *Norton's Woods, 136 Irving Street, Cambridge, MA 02138.*

Dædalus (ISSN 0011-5266) *is published quarterly by the American Academy of Arts and Sciences. Subscription rate: $16, one year; $24, two years; $32, three years.*

All subscription orders, single copy orders, and change of address information must be sent in writing *to the* Dædalus *Business Office, P.O. Box 515, Canton, MA 02021, U.S.A. (Telephone: 617-828-8450). Orders for single copies must be accompanied with prepayment of $5.00.*

Printed in the United States of America.
Printing Office: 2901 Byrdhill Road, Richmond, VA 23228.

U.S.A. newsstand distribution by Eastern News Distributors, Inc., 1130 Cleveland Road, Sandusky, OH 44870.

Second-class postage paid at Canton, Massachusetts, and at additional mailing offices.

SIGILLUM ACADEMIÆ ARTIUM ET SCIENTIARUM AMERICANÆ · SUB LIBERTATE FLORENT · MDCCLXXX ·

David L. Waltz

The Prospects for Building Truly Intelligent Machines

CAN ARTIFICIAL INTELLIGENCE be achieved? If so, how soon? By what methods? What ideas from current AI research will in the long run be important contributions to a science of cognition? I believe that AI can be achieved, perhaps within our lifetimes, but that we have major scientific and engineering obstacles to hurdle if it is to come about. The methods and perspective of AI have been dramatically skewed by the existence of the common digital computer, sometimes called the von Neumann machine, and ultimately, AI will have to be based on ideas and hardware quite different from what is currently central to it. Memory, for instance, is much more important than its role in AI so far suggests, and search has far less importance than we have given it. Also, because computers lack bodies and life experiences comparable to humans', intelligent systems will probably be inherently different from humans; I speculate briefly on what such systems might be like.

OBSTACLES TO BUILDING INTELLIGENT SYSTEMS

If we are to build machines that are as intelligent as people, we have three problems to solve: we must establish a science of cognition; we must engineer the software, sensors, and effectors for a full system; and we must devise adequate hardware.

David L. Waltz, a professor of computer science at Brandeis University, is senior scientist and director of advanced information systems at Thinking Machines Corporation.

Establishing a Science of Cognition

We have no suitable science of cognition. We have only fragments of the conception, and some of those are certainly incorrect. We know very little about how a machine would have to be organized to solve the problems of intelligence. Virtually all aspects of intelligence—including perception, memory, reasoning, intention, generation of action, and attention—are still mysterious. However, even if we understood how to structure an intelligent system, we would not be able to complete the system because we also lack an appropriate science of knowledge. For some aspects of knowledge, any computational device will be on a strong footing when compared with a person. Machine-readable encyclopedias, dictionaries, and texts will eventually allow machines to absorb book knowledge quite readily. For such understanding to be deep, however, a system needs perceptual grounding and an understanding of the physical and social world. For humans, much of this knowledge is either innate or organized and gathered by innate structures that automatically cause us to attend to certain features of our experience, which we then regard as important. It will be extremely difficult to characterize and build into a system the kinds of a priori knowledge or structuring principles humans have.

Engineering the Software

Any truly intelligent system must be huge and complex. As Frederick Brooks argues, writing on his experience building the large operating system OS360 at IBM, it is not possible to speed up a software project by simply putting more and more people on it.[1] The optimum team size for building software is about five people. For this reason, and because of the sheer scope of a project of this sort—which dwarfs any that have been attempted in programming to date—hand coding will certainly be too slow and unreliable to accomplish the whole task. Consequently, a truly intelligent system will have to be capable of learning much of its structure from experience.

What structures must be built into a system to allow it to learn? This is a central question for current AI, and the answer depends on issues of knowledge representation: How should knowledge be represented? Out of what components (if any) are knowledge structures built?

Creating the Hardware

We must be able to build hardware that is well matched to AI's knowledge representation and learning needs and that compares in power with the human brain. No one should be surprised that the puny machines AI has used thus far have not exhibited artificial intelligence. Even the most powerful current computers are probably no more than one four-millionth as powerful as the human brain. Moreover, current machines are probably at least as deficient in memory capacity: today's largest computers probably have no more than about one four-millionth of the memory capacity of the human brain. Even given these extreme discrepancies, hardware will probably prove the easiest part of the overall AI task to achieve.

I begin with a discussion of traditional AI and its theoretical underpinnings in order to set the stage for a discussion of the major paradigm shifts (or splits) currently under way in and around AI. As an advocate of the need for new paradigms, I here confess my bias. I see no way that traditional AI methods can be extended to achieve humanlike intelligence. Assuming that new paradigms will replace or be merged with the traditional ones, I make some projections about how soon intelligent systems can be built and what they may be like.

LIMITS OF TRADITIONAL AI

Two revolutionary paradigm shifts are occurring within artificial intelligence. A major force behind the shifts is the growing suspicion among researchers that current AI models are not likely to be extendable to a point that will bring about human-level intelligence. The shifts are toward massively parallel computers and toward massively parallel programs that are more taught than programmed. The resultant hardware and software systems seem in many ways more brainlike than the serial von Neumann machines and AI programs that we have become used to.

For thirty years, virtually all AI paradigms were based on variants of what Herbert Simon and Allen Newell have presented as "physical symbol system" and "heuristic search" hypotheses.[2] (See also the article by Hubert and Stuart Dreyfus in this issue of *Dædalus*.)

According to the physical symbol system hypothesis, symbols (wordlike or numerical entities—the names of objects and events) are the primitive objects of the mind; by some unknown process, the brain mimics a "logical inference engine," whose most important feature is that it is able to manipulate symbols (that is, to remember, interpret, modify, combine, and expand upon them); and computer models that manipulate symbols therefore capture the essential operation of the mind. In this argument it does not matter whether the materials out of which this inference engine is built are transistors or neurons. The only important thing is that they be capable of a universal set of logical operations.[3] The physical symbol system hypothesis in turn rests on a foundation of mathematical results on computability, which can be used to show that if a machine is equivalent to a Turing machine—a simple kind of computational model devised by the pioneering British mathematician Alan Turing—then it is "universal"; that is, the machine can compute anything that can be computed. All ordinary digital computers can be shown to be universal in Turing's sense.*

In the heuristic search model, problems of cognition are instances of the problem of exploring a space of possibilities for a solution. The search space for heuristic search problems can be visualized as a branching tree: starting from the tree's root, each alternative considered and each decision made corresponds to a branching point of the tree. Heuristics, or rules of thumb, allow search to be focused first on branches that are likely to provide a solution, and thus prevent a combinatorially explosive search of an entire solution space.† Heuristic search programs are easy to implement on ordinary serial digital computers. Heuristic search has been used for a wide variety of applications, including decision making, game playing, robot planning and problem solving, natural-language processing, and the classification of perceptual objects. Heuristic search has enjoyed particular prominence, for it is at the heart of "expert systems," AI's greatest commercial success by far.

*There is perhaps one critical aspect in which all computers fail to match a Turing machine: the Turing machine includes an infinite tape, from which it reads its programs and onto which it writes its results. All computers (and presumably humans) have finite memories.

†Combinatorially explosive problems are problems in which the computational costs of solving each slightly more difficult problem grow so rapidly that no computer will ever be able to solve them; that is, even a computer with as many components as there are electrons in the universe and an instruction execution time as short as the shortest measurable physical event might require times greater than the age of the universe to consider all possible problem solutions.

In retrospect it is remarkable how seriously heuristic search has been taken as a cognitive model. When I was a graduate student in the late 1960s, the standard AI view was that for any intelligent system, the nature of a problem constrains the nature of any efficient solution, and that any system, human or computer, given a problem to solve, tends to evolve a similar, or at least an analogous, internal structure to deal with it. Thus, it was argued, studying efficient problem solutions on computers is a good way to study cognition.[4] Virtually everyone in AI at the time accepted the centrality and immutability of heuristic search machinery unquestioningly and assumed that learning should be accomplished by evolving, adapting, or adding to the heuristics and the knowledge structures of the search space. (The exceptions were the "neural net" and "perceptron" researchers, who had been actively exploring more brainlike models since the early 1950s. More on this later.)

It is now commonly recognized that the nature of the computers and computing models available to us inevitably constrains the problem-solving algorithms that we can consider. (John Backus introduced this idea to the broad computing community in his Turing Award lecture of 1977.[5]) As explained below, it has become clear that traditional AI methods do not scale up well and that new AI paradigms will therefore be needed. Despite this change in attitude, there have been few prospective replacements within AI for heuristic search (or for serial, single-processor digital computers) until very recently.

The reasons AI has focused almost exclusively on the physical symbol system and heuristic search views are deeply rooted in its history and in part reflect the myopic concentration on serial digital computers that has characterized all of computer science. The focus on heuristic search also reflects the influence of the psychological research of the 1950s. AI began at a time when psychologists were much enamored of protocol analysis, a way of examining human behavior by having subjects give accounts of their mental experience while they are solving problems.[6] Such psychological research was interpreted as evidence that the main human mechanism for problem solving is trial and error. AI adapted this model as its heuristic search paradigm. In this paradigm problems are solved by sequentially applying "operators" (elementary steps in a problem solution) and allowing "backtracking," a form of trial and error whereby a program backs up to an earlier decision point and tries new branches if the first ones explored prove fruitless.

It is difficult to see how any extension of heuristic-search–based systems could ever demonstrate common sense. In most AI systems, problem statements have come from users; the systems have not needed to decide what problems to work on. They have had relatively few actions or operators available, so search spaces have been tractable. Real-time performance hasn't generally been necessary. This way of operating will clearly not do in general. Eventually, AI must face the scale-up question: Given the immense range of possible situations a truly intelligent system could find itself in and the vast number of possible actions available to it, how could the system ever manage to search out appropriate goals and actions?

Moreover, as John McCarthy has pointed out, rule-based systems may be inherently limited by the "qualification problem": given a certain general rule, one can always alter the world situation in such a way that the rule is no longer appropriate.[7] For example, suppose we offered the rule:

bird (x) → *fly (x)* (if *x* is a bird, then *x* can fly).

Everyone knows that the rule must be amended to cover birds such as penguins and ostriches, so that it becomes:

not flightless (x) and bird (x) → *fly (x),* where
"flightless *(x)*" is true of the appropriate birds.

However, we also know a bird cannot fly if it is dead, or if its wings have been pinioned, or if its feet are embedded in cement, or if it has been conditioned by being given electric shocks each time it tries to fly.[8] There seems to be no way to ever completely specify rules for such cases. There are also serious difficulties in formulating rules for deciding which facts about the world ought to be retracted and which should still hold after particular events or actions have occurred. This is known as the "frame problem." "Nonmonotonic logic," which treats all new propositions or rules as retractable hypotheses, has been proposed for dealing with these problems.[9] However, some researchers in this area[10] are pessimistic about its potential, as am I.

By objecting to traditional AI approaches, I am not disputing the notions of universal computation or the Turing machine results, which are established mathematically beyond doubt. Rather, I dispute the heuristic search metaphor, the relationship between physical symbol systems and human cognition, and the nature and "granularity" of the units of thought. The physical symbol system hypothesis, also long shared by AI researchers, is that a vocabulary close to natural language (English, for example, perhaps supplemented by previously unnamed categories and concepts) would be sufficient to express all concepts that ever need to be expressed. My belief is that natural-language–like terms are, for some concepts, hopelessly coarse and vague, and that much finer, "subsymbolic" distinctions must be made, especially for encoding sensory inputs. At the same time, some mental units (for example, whole situations or events—often remembered as mental images) seem to be important carriers of meaning that may not be reducible to tractable structures of words or wordlike entities. Even worse, I believe that words are not in any case carriers of complete meanings but are instead more like index terms or cues that a speaker uses to induce a listener to extract shared memories and knowledge. The degree of detail and number of units needed to express the speaker's knowledge and intent and the hearer's understanding are vastly greater than the number of words used to communicate. In this sense language may be like the game of charades: the speaker transmits relatively little, and the listener generates understanding through the synthesis of the memory items evoked by the speaker's clues. Similarly, I believe that the words that seem widely characteristic of human streams of consciousness do not themselves constitute thought; rather, they represent a projection of our thoughts onto our speech-production faculties. Thus, for example, we may feel happy or embarrassed without ever forming those words, or we may solve a problem by imagining a diagram without words or with far too few words to specify the diagram.

WHAT'S THE ALTERNATIVE?

Craig Stanfill and I have argued at length elsewhere that humans may well solve problems by a process much more like *lookup* than *search,* and that the items looked up may be much more like representations of specific or stereotyped episodes and objects than like rules and facts.[11]

On the Connection Machine, built by Thinking Machines Corporation,[12] we have now implemented several types of "associative memory" systems that reason on the basis of previous experience.[13] For example, one experimental system solves medical diagnosis problems with "memory-based reasoning": given a set of symptoms and patient characteristics, the system finds the most similar previous patients and hypothesizes that the same diagnoses should be given to the new patient. "Connectionist," or neural net, models, which I shall describe later, solve similar problems, though in a very different manner. While a great deal of research is still required before such systems can become serious candidates for truly intelligent systems, I believe that these architectures may prove far easier to build and extend than heuristic search models. These new models can learn and reason by remembering and generalizing specific examples; heuristic search models, in contrast, depend on rules. It has proved difficult to collect rules from experts—people are generally not even aware of using rules. We do not know how to check sets of rules for completeness and self-consistency. Moreover, a finite set of rules cannot capture all the possible conclusions that may be drawn from a set of examples any more than a set of descriptive sentences can completely describe a picture.

It is important to note, however, that some kinds of knowledge in rule-based systems are hard to encode in our memory-based model. For instance, as currently formulated, our system does not use patients' histories and is unable to figure out that medication dose size ought to be a function of a patient's weight. Recent research strongly suggests that humans reason largely from stereotypes and from specific variations of these stereotypes. Our system does not yet demonstrate such abilities.

IMPLEMENTING ASSOCIATIVE MEMORY SYSTEMS

In the short run, associative memory models can very nicely complement AI models. Associative models have been studied for quite a while but seldom implemented (except for very small problems) because they are computationally very expensive to run on traditional digital computers. One class of associative memory implementation is called the connectionist, or neural net, model. Such systems are direct descendents of the neural net models of the 1950s. In them, thousands of processing units, each analogous to a neuron, are interconnected by links, each analogous to a synaptic connection between neurons. Each link has a "weight," or a connection strength. A system's knowledge is encoded in link weights and in the interconnection pattern of the system. Some units serve as input units, some as output units, and others as "hidden units" (they are connected only to other units and thus cannot be "seen" from either the input or the output channels).

Such networks display three interesting abilities. The first is *learning*. Several methods have now been devised that enable such a system, upon being given particular inputs, to be taught to produce any desired outputs. The second interesting ability is *associative recall*. Once trained to associate an output with a certain input, a network can, given some fraction of an input, produce a full pattern as its output. The third interesting property is *fault tolerance:* the network continues to operate even when some of the units are removed or damaged. In short, connectionist computing systems have many of the properties that we have associated with brains; these systems differ significantly from computers, which have traditionally been viewed as automatons with literal minds, able to do only what they are programmed to do.[14]

These networks can now be implemented efficiently on such massively parallel hardware as the Connection Machine system or by using custom chips. While associative memory systems have been simulated on traditional serial digital computers, the simulations have been very slow; a serial computer must simulate each of the computational units and links in turn and must do so many times to carry out a single calculation. A massively parallel machine can provide a separate small processor for each of the units in the associative memory system and can thus operate much more rapidly.

Stanfill and I have been exploring a functionally similar massively parallel method called memory-based reasoning. In this type of reasoning, a Connection Machine is loaded with a large data base of situations. Each situation in the data base contains both a set of attributes and an outcome. In a medical data base, for instance, the attributes would be symptoms and a patient's characteristics, and the outcome would be a diagnosis or a treatment. Each item in the data base is stored in a separate processor. When a new example to be classified is encountered, its properties are broadcast to all the processors that hold situations; each of these processors compares its situation with the input situation and computes a score of nearness to the input. The system then finds the nearest matches to the input example and, provided they are sufficiently close, uses the outcomes of these matching items to classify the new example.

Memory-based reasoning systems also have many desirable characteristics. They are fault tolerant; they can generalize well to examples that have never been seen in their exact form before; they give measurements of the closeness of the precedents to the current example, which can serve as measures of confidence for the match. If there is an exact match with a previous example, the systems can give a decision with certainty. It is easy to teach such systems: one simply adds more items to their data bases.

The complicated part of memory-based reasoning systems is the computation of nearness. To calculate the similarity of any memory example to the pattern to be classified, each memory item must first find the distance, or difference, between each of its attribute values and the attribute values of the pattern to be classified. These distances in turn depend on the statistical distribution of attribute values and on the degree of correlation between each attribute value and the outcomes with which it simultaneously occurs. All the distances for each attribute must then be combined for each memory item to arrive at its total distance from the item to be classified. Thus, computing the nearness score involves a great deal of statistical calculation across all records in the data base.[15]

What is the role of associative memory systems in traditional artificial intelligence? While they can substitute for expert systems under certain circumstances, connectionist and memory-based reasoning systems are better viewed as complements to traditional AI than as replacements for it. In one very useful mode, associative memory systems can be used to propose or hypothesize solutions to complex problems, and traditional AI systems can be used to verify that the differences between the problems that are currently being attacked and examples in the data base are unimportant. If such differences are important, the associative memory systems can propose subgoals to attempt. Thus, the associative memory process can provide a very powerful heuristic method for jumping to conclusions, while traditional AI can be used to verify or disconfirm such conclusions. Such hybrid systems could help AI models avoid the problems of searching combinatorially large spaces. Because of the computational resources required, the bulk of the computing power in an AI system of this sort would probably reside in the associative memory portion.

In the long run, however, such models are still unlikely to provide a satisfactory explanation for the operations of human thought, though I suspect they will come much closer than AI has. To my mind, the best exposition on the ultimate architecture required is Marvin Minsky's "society of mind."[16] Minsky argues persuasively, using a very wide range of types of evidence, that the brain and the mind are made up of a very large number of modules organized like a bureaucracy. Each module, or "demon," in the bureaucracy has only limited responsibilities and very limited knowledge; demons constantly watch for events of interest to themselves and act only when such events occur. These events may be external (signaled by sensory units) or purely internal (the result of other internal demons that have recognized items of interest to themselves). Actions of demons can either influence other demons or activate effectors and can thereby influence the outside world. One can make a simple analogy between a society of mind and associative memory models: in memory-based reasoning each data base item would correspond to an agent; in a connectionist model, each neural unit would correspond to an agent.

LOGICAL REASONING

I believe logical reasoning is not the foundation on which cognition is built but an emergent behavior that results from observing a sufficient number of regularities in the world. Thus, if a society of demonlike agents exhibits logical behavior, its behavior can be described by rules, although the system contains no rules to govern its operation. It operates in a regular fashion because it simulates the world's regularities.

Consider a developing infant. In the society-of-mind model, the infant first develops a large number of independent agencies that encode knowledge of the behavior of specific items in the physical world: when a block is dropped, it falls; when the child cries, its parent comes to attend; when the child touches a flame, it feels pain. Each of these examples is handled initially by a separate small bureaucracy of agents. Each bureaucracy represents the memory of some specific event. A particular agency becomes responsible for an episode because of the initial "wiring" of the brain; shortly after an agency is first activated, it changes its synaptic weights, so that any new event that activates any part of the agency will cause the entire agency to be reactivated. When similar events reactivate these agencies, new bureaucracies encoding the similarities and differences between the new and the old events are constructed out of previously unused, but closely connected (hence activated), agents. After many such incremental additions to the society of agents, a child eventually develops agents for abstract categories and rules; cuts, pinches, and burns all cause pain, and thus other agents that happen to be activated in these cases become associated with the concept of pain. Eventually, the concepts of the constant conjunction of pain with its various causes become the specialty of particular "expert" agents responsible for certain regularities in the world. Ultimately, these agents become part of the bureaucracy for the concept of causality itself. Thus agents come to reason about very general categories, no longer necessarily rooted directly in experience, and can understand abstract causal relationships. Take pain in the abstract, for example: if one breaks a law and is apprehended, one knows one will probably be punished; if one does not keep promises, one understands that other people may be angry and may retaliate; and so on.

On the surface it might seem that what is being proposed is to replace a single expert program with many expert programs, arranged in a hierarchy. However, each of the expert agents is extremely simple, in the sense that it "knows" only about one thing. The experts are connected to a perceptual system and to each other in such a way that they are triggered only when the conditions about which they are expert are actually satisfied.

While this may be a satisfactory description of the composition of the mind, it is not yet sufficiently precise to serve as a design for a very large-scale program that can organize itself to achieve intelligence. Programs that operate on the principles of the society of mind may well be the end point of many steps in the evolution of the design of intelligent systems. I believe that hybrids of associative memory and traditional AI programs for logical reasoning show the greatest promise in the near term for AI applications. It is possible that they will also prove to be useful models of cognition.[17]

LIMITS OF TRADITIONAL COMPUTER HARDWARE

Researchers' suspicion that current AI models may not be extensible to systems with human-level intelligence is not the only force driving the paradigm shift toward massively parallel computing models. Economic considerations, which transcend AI concerns, are another. Today's serial computers have begun to reach limits beyond which they cannot be speeded up at reasonable cost. For a serial, single-processor computer to operate more rapidly than at present, its processor must execute each instruction more rapidly. To accelerate processing, manufacturers have brought new, faster-acting materials into use. They have also shrunk circuits to smaller and smaller sizes so as to shorten signal paths, since internal communication speeds, and therefore overall processing rates, are limited by the speed of light. The smaller the computer, the faster its internal communications. Because each component generates heat, and because dense chips produce more heat than others, ultradense chips of exotic materials often require the addition of elaborate and expensive cooling systems. All this means that doubling the power of a serial machine usually increases its cost by more than a factor of two—sometimes much more.

In contrast, parallel designs promise the possibility of doubling power by simply doubling the number of processors, possibly for less than two times the cost, since many system components (disk storage units, power supplies, control logic, and so on) can be shared by all processors, no matter how numerous. For example, the Connection Machine system contains up to 65,536 processors. Even in its initial version, the Connection Machine is very inexpensive in terms of the number of dollars it costs per unit of computation; its cost in relation to its performance is about one-twentieth that of serial supercomputers.* Moreover, the cost of highly parallel processors is likely to drop dramatically. Initially, any chip is expensive because of

*The cost/performance figure is the cost per standard computing operation. The typical standard computing operation is either a fixed-point addition or a floating-point multiplication. Fixed-point performance is measured in millions of instructions per second (MIPS). Floating-point performance is measured in millions of floating operations per second (MFLOPS—pronounced "megaflops"). Cost/performance is measured in dollars per MIPS or dollars per MFLOPS.

low yield (only a fraction of usable chips results from initial production) and the need to recover research, design, and development costs. The price of chips follows a "learning curve," a drop-off in cost as a function of the number of chips fabricated. Memory is the prime example: the cost per bit of memory storage has dropped by a factor of ten every five years for thirty-five years running, yielding a cost that is one ten-millionth that of the 1950 price—one one-hundred millionth after adjustment for inflation! Since the processors of a massively parallel computer are mass-produced, as memory chips are, the cost of a given amount of processing power for parallel machines should drop as rapidly as the cost of memory—that is, very rapidly indeed.

The cost of computer systems involves, of course, both hardware and software. How is one to program a machine with tens of thousands or perhaps millions of processors? Clearly, human programmers cannot afford the time or the money to write a program for each processor. There seem to be two practical ways to program such machines. The first, which has been in most use to date, is to write a single program and have each processor execute it in synchrony, each processor working on its own portion of the data. This method is "data-level parallelism." A second way is to program learning machines that can turn their experiences into a different code or data for each processor.

Research in machine learning has grown dramatically during the last few years. Researchers have identified perhaps a dozen distinctly different learning methods.[18] Many massively parallel learning schemes involve the connectionist, or neural net, models mentioned earlier. Connectionist systems have usually been taught with some form of supervised learning: an input and a desired output are both presented to a system, which then adjusts the internal connection strengths among its neuronlike units so as to closely match the desired input-output behavior. Given a sufficiently large number of trials, generally on the order of tens of thousands, such systems are able to learn to produce moderately complex desired behavior. For example, after starting from a completely random state and being trained repeatedly with a 4,500-word data base of sample pronunciations, a system called NETtalk was able to learn to pronounce novel English words with fairly good accuracy.[19]

The central problem to be solved in connectionist and society-of-mind learning research is the "credit assignment problem," the problem of apportioning simple rewards and punishments among a vast number of interconnected neuronlike computing elements. To show the relevance of this problem to the ultimate goals of AI, I will couch the problem in terms of the "brain" of a robotic system that we hope will learn through its experiences.

Assume a large set (perhaps billions) of independent neural-like processing elements interconnected with many links per element. Some elements are connected to sensors, driven by the outside world; others are connected to motor systems that can influence the outside world through robotic arms and legs or wheels, which generate physical acts, as well as through language-production facilities, which generate "speech acts." At any given time a subset of these elements is active; they form a complex pattern of activation over the entire network. A short time later, the activation pattern changes because of the mutual influences among processing elements and sensory inputs.

Some activation patterns trigger motor actions. Now and then rewards or punishments are given to the system. The credit assignment problem is this: which individual elements within the mass of perhaps trillions of elements should be altered on the basis of these rewards and punishments so the system will learn to perform more effectively—that is, so the situations that have led to punishments can be avoided in the future and so the system will more often find itself in situations that lead to rewards?

The credit assignment problem has at least two aspects. The simpler is the *static* credit assignment problem, in which rewards and punishments occur shortly after the actions that cause them. Such systems receive instant gratification and instant negative feedback. The static credit assignment problem has been found reasonably tractable: units that are active can be examined, and those that have been active in the correct direction have their connections with action systems strengthened, while those that have been inappropriately active have their connection strengths reduced. If the reward or punishment occurs substantially after the fact, however, we have a *temporal* credit assignment problem, which is significantly more difficult. To solve this problem, a system must keep memories of the past states through which it has passed and have the capacity to analyze and make judgments about which earlier states were responsible for the rewards and punishments. Progress on the temporal credit assignment problem has been promising, but much remains to be done before it can be considered solved.[20]

In my estimation, these learning methods will only be suitable for producing modules of an overall intelligent system. A truly intelligent system must contain many modules. It seems very unlikely that the organization of an entire brain or mind could be automatically learned, starting with a very large, randomly interconnected system. Infants are highly organized at birth. They do not, for instance, have to learn to see or hear in any sense that we would recognize as learning. Their auditory and visual systems seem already organized to be able to extract meaningful units (objects, events, sounds, shapes, and so on). Elizabeth Spelke and her research associates have found that two-month-old infants are able to recognize the coherence of objects and that they show surprise when objects disappear or apparently move through each other.[21] At that age they cannot have learned about the properties of objects through tactile experience. It is not too surprising that such abilities can be "prewired" in the brain: newborn horses and cattle are able to walk, avoid bumping into objects, and find their mother's milk within minutes of birth. In any case, the necessity for providing intelligent systems with a priori sensory organization seems inescapable. On what other basis could we learn from scratch what the meaningful units of the world are?[22]

THE FUTURE OF ARTIFICIAL INTELLIGENCE

Any extrapolation of current trends forces one to conclude that it will take a very long time indeed to achieve systems that are as intelligent as humans. Nevertheless, the performance of the fastest computers seems destined to increase at a much greater rate than it has over the last thirty years, and the cost/performance figures for large-scale computers will certainly drop.

The effect of a great deal more processing power should be highly significant for AI. As claimed earlier, current machines probably have only one four-millionth the amount of computing power that the human brain has. However, it is quite conceivable that within about twenty-five years we could build machines with comparable power for affordable prices (for the purposes of this argument, let an affordable price be $20 million, the cost of today's most expensive supercomputer).

The Connection Machine system, currently probably the fastest in the world, can carry out the kinds of calculations we think the brain uses at the rate of about 3.6 × 10^{12} bits a second, a factor of about twenty million away from matching the brain's power (as estimated by Jack Schwartz in his article in this issue of *Dædalus*). One may build a more powerful Connection Machine system simply by plugging several of them together. The current machine costs about $4 million, so within our $20 million budget, a machine of about five times its computing power (or 1.8 × 10^{13} bits per second) could be built. Such a machine would be a factor of four million short. The stated goal of the DARPA (Defense Advanced Research Projects Agency) Strategic Computing Initiative is to achieve a thousandfold increase in computing power over the next ten years, and there is good reason to expect that this goal can be achieved. In particular, the Connection Machine system achieves its computation rates without yet using exotic materials or extreme miniaturization, the factors that have enabled us to so dramatically speed up traditional computers. If a speedup of one thousand times every ten years can be achieved, a computer comparable in processing power to the brain could be built for $20 million by 2012.

Using Schwartz's estimates, we find that the total memory capacity of the brain is 4 × 10^{16} bytes. The current Connection Machine can contain up to two gigabytes (2 × 10^{9} bytes). In today's computer world, two gigabytes of memory is considered a large amount, yet this is a factor of twenty million short, or a factor of four million short for a system with five Connection Machines.

At today's prices, two gigabytes of memory costs roughly $1 million, so to buy enough memory to match human capacity would cost on the order of $20 trillion, roughly ten times our current national debt. Given its long-term price decline of roughly a factor of ten every five years, the cost of 4 × 10^{16} bytes of memory will be in the $20 million range within thirty years, so that the time at which we might expect to build a computer with the potential to match human intelligence would be around the year 2017.* As suggested earlier,

*Well before the 2017 date, however, mass storage devices (disk units and other storage media) will certainly be capable of storing this much material at an affordable price.

however, building the hardware may be the easiest part; the need to untangle the mysteries of the structure and functioning of the mind, to gather the knowledge both innate and learned, and to engineer the software for the entire system will probably require time that goes well beyond 2017. Once we have a piece of hardware with brain-level power and appropriate a priori structure, it still might take as long as the twenty years humans require to reach adult-level mental competence! More than one such lengthy experiment is likely to be required.

What could we expect the intelligence of such powerful machines to be like? Almost certainly they will seem alien when compared with people. In some ways such machines will eclipse maximum human performance, much as pocket calculators outperform humans in arithmetic calculation. The new machines may have perfect recall of vast quantities of information, something that is not possible for people. (While humans apparently have vast amounts of memory, we are quite poor at the literal memorization of words, images, names, and details of events.) Unless deliberately programmed in, such machines would not have a repertoire of recognizable human emotions. Nor would they have motivation in any ordinary human sense. Motivation and drive seem to be based on innate mechanisms developed over eons of evolution to ensure that we make species-preserving decisions—to avoid pain, continue to eat and drink, get enough sleep, reproduce, care for our young, act altruistically (especially toward relatives and friends)—without requiring that we understand that the real reason for carrying out these actions is species preservation.[23] (It is, however, quite possible that it will prove useful to endow machines capable of problem solving and learning with the ability to experience some analogues of frustration, pleasure at achieving a goal, confusion, and other such emotion-related attitudes toward emergent phenomena in order that they can generate useful abstractions for deciding when to abandon a task, ask for advice, or give up.)

AI researchers can grasp the opportunity to build human-level intelligent machines only if they find ways to fill prodigious quantities of memory with important material. They will be able to do so only if AI can produce adequate sensory systems (for hearing, vision, touch, kinesthesia, smell, and taste). With sensory systems, AI systems will for the first time be able to learn from experience. Such experience may initially be little more than rote memory—that is, storing records of the partially digested sensory patterns seen by the system. Yet, as argued earlier, the storage of vast amounts of relatively literal material may be a key to intelligent behavior. The potential for artificial intelligence depends on the possibility of building systems that no longer require programming in the same sense that it is now required. Then we could overcome the tendency of systems development to be very slow because of software engineering difficulties.

There is also the question of what kind of "body" such an intelligence must be embedded in for it to really understand rather than to merely simulate understanding. Must the machine be wired to have emotions if it is to understand our human emotional reactions? If a machine were immortal, could it understand our reactions to our knowledge of our own mortality? Intelligent machines might be cloned by simply copying their programming or internal coding onto other identical pieces of hardware. There is no human analogue to a machine that would have experience as a unitary entity for an extended period and then, at some point during its "lifetime," suddenly become many separate entities, each with different experiences. Exactly what kind of intelligence this would be is therefore an open question.

SUMMARY

We are nearing an important milestone in the history of life on earth, the point at which we can construct machines with the potential for exhibiting an intelligence comparable to ours. It seems certain that we will be able to build hardware that is a match for human computational power for an affordable price within the next thirty years or so. Such hardware will without doubt have profound consequences for industry, defense, government, the arts, and our images of ourselves.

Having hardware with brain-level power will not in itself, however, lead to human-level intelligent systems, since the architecture and programs for such systems also present unprecedented obstacles. It is difficult to extrapolate to future effects from the rate of progress that has been made to date. Progress has been very slow, in part because the computational models that have been used have been inappropriate to the task. This inappropriateness applies most critically to the problem of learning. Without learning, systems must be handbuilt. We don't know how closely we must match human brain details to foster appropriate learning and performance. With the right architectures, it is likely that progress, both in the building of adequately powerful hardware and in programming such hardware (by teaching), will accelerate. I believe that the construction of truly intelligent machines is sufficiently likely to justify beginning study and policy planning now. In that way we can maximize their benefits and minimize their negative effects on society.

ENDNOTES

[1] Frederick P. Brooks, *The Mythical Man-Month: Essays on Software Engineering* (Reading, Mass.: Addison-Wesley, 1974).

[2] Allen Newell and Herbert Simon, *Human Problem Solving* (Engelwood Cliffs, N.J.: Prentice-Hall, 1972).

[3] The Boolean operations AND, OR, and NOT constitute a universal set. NAND (NOT AND) also is universal by itself, as is NOR (NOT OR). For a derivation of this result see Marvin L. Minsky, *Computation: Finite and Infinite Machines* (Cambridge: MIT Press, 1967).

[4] Herbert A. Simon, *The Sciences of the Artificial* (Cambridge: MIT Press, 1965).

[5] John Backus, "Can Programming Be Liberated from the von Neumann Style? A Functional Style and its Algebra of Programs," *Communications of the ACM* 21 (8) (August 1978):613–41.

[6] George A. Miller, Eugene Galanter, and Karl Pribram, *Plans and the Structure of Behavior* (New York: Holt, Rinehart, and Winston, 1954).

[7] John McCarthy, "Epistemological Problems in Artificial Intelligence," in *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* (Los Altos, Calif.: Morgan-Kaufmann, August 1977), 1038–44.

[8] Example from Marvin Minsky, personal communication.

[9] John McCarthy, "Circumscription—A Form of Nonmonotonic Reasoning," *Artificial Intelligence* 13 (1) (1980):27–39; and Drew V. McDermott and Jon Doyle, "Nonmonotonic Logic I," *Artificial Intelligence* 13 (1) (1980):41–72.

[10] Steve Hanks and Drew V. McDermott, "Default Reasoning, Nonmonotonic Logics, and the Frame Problem," in *Proceedings of the Fifth National Conference on Artificial Intelligence* (Los Altos, Calif.: Morgan-Kaufmann, August 1986), 328–33.

[11] Craig Stanfill and David L. Waltz, "Toward Memory-Based Reasoning," *Communications of the ACM* 29 (12) (December 1986): 1213–28.

[12] W. Daniel Hillis, *The Connection Machine* (Cambridge: MIT Press, 1986).

[13] David E. Rumelhart, James L. McClelland, and the PDP Research Group, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vols. 1, 2 (Cambridge: MIT Press, 1986).

[14] For extended treatment of such systems, see Rumelhart and McClelland, *Parallel Distributed Processing,* and the special issue on connectionist models of *Cognitive Science* 9 (1) (1985).

[15] For details, see Stanfill and Waltz, "Toward Memory-Based Reasoning."

[16] Marvin L. Minsky, *The Society of Mind* (New York: Simon and Schuster, 1986); *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence,* by A. Harry Klopf (Washington, D.C.: Hemisphere, 1982), presents a compatible neural theory.

[17] There is also a fairly extensive literature on learning and knowledge acquisition. It is based on the heuristic search and physical symbol system paradigms. Broadly speaking, these learning algorithms fall into three categories. The first type is statistical and uses a large number of processors to find patterns or regularities in data bases of examples—in medical diagnosis, weather forecasting, and decision making, for example. Three systems that fall into this category are the ID3 system of Ross Quinlan and the systems built by Ryszard Michalski, both of which are described in Ryszard S. Michalski, Jaime Carbonell, and Thomas Mitchell, eds., *Machine Learning: An Artificial Intelligence Approach* (Los Altos, Calif.: Tioga Publishing Company, 1983), and the "memory-based reasoning" system of Craig Stanfill and David Waltz (see Stanfill and Waltz, "Toward Memory-Based Reasoning"). A second type of learning algorithm uses "production rules" (sometimes termed "if-then rules") and learns by adding to and modifying an existing set of such rules. The rules are changed by providing "experience," which may include "rewards and punishments." Such systems can also be taught by giving them correct examples from which they can learn rules by rote. Two systems of this sort are the genetic algorithms of John Holland (see John H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence* [Ann Arbor: University of Michigan Press, 1975]) and the SOAR system of Allen Newell and Paul Rosenbloom (see John E. Laird, Paul S. Rosenbloom, and Allen Newell, "Chunking and SOAR: The Anatomy of a General Learning Mechanism," *Machine Learning* 1 [1] [1986]:11– 46). The third branch is "explanation-based learning" (see Gerald F. DeJong and Raymond A. Mooney, "Explanation-Based Learning: An Alternative View," *Machine Learning* 1 [2] [April 1986]:145–76). An explanation-based learning system attempts to build causal structures, or "schemata," as explanations of new phenomena and as elements for building new schemata.

[18] One of the most successful learning methods is the centerpiece of McClelland and Rumelhart's *Parallel Distributed Processing.* Other systems include Stephen Grossberg's, Andrew Barto's, and Geoffrey Hinton's. See Stephen Grossberg, "Competitive Learning: From Interactive Activation to Adaptive Resonance," *Cognitive Science* 11 (1) (January-March 1987):23–64; Andrew G. Barto, "Learning by Statistical Cooperation of Self-Interested Neuron-like Computing Elements," *Human Neurobiology* 4 (1985):229–56; and Geoffrey Hinton, "The Boltzmann Machine," in Geoffrey E. Hinton and John A. Anderson, eds., *Parallel Models of Associative Memory* (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1981).

[19] Terrence J. Sejnowski and Charles R. Rosenberg, "NETtalk: A Parallel Network that Learns to Read Aloud," Technical Report JHU/EECS-86–01 (Baltimore, Md.: Johns Hopkins University, Electrical Engineering and Computer Science, 1986).

[20] Ronald J. Williams, "Reinforcement-Learning Connectionist Systems: A Progress Report" (unpublished manuscript, College of Computer Science, Northeastern University, November 1986).

[21] Elizabeth Spelke, "Perceptual Knowledge of Objects in Infancy," in Jacques Mehler, Edward C. T. Walker, and Merrill Garrett, eds., *Perspectives on Mental Representation: Experimental and Theoretical Studies of Cognitive Processes and Capacities* (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1962).

[22] In the *Critique of Pure Reason* Immanuel Kant argues essentially this point: that "the innate forms of human perception and the innate categories of human understanding impose an invariant order on the initial chaos of raw sensory experience." This is quoted from Paul M. Churchland in *Matter and Consciousness* (Cambridge: MIT Press, 1984), 84.

[23] See Isaac Asimov, *I, Robot* (New York: The New American Library of World Literature, 1950), 6, for an early exploration of the need for a kind of ethics for robots, embodied in three laws of robotics:

> 1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm; 2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law; 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Ironic, alas, for the first highly intelligent mobile robot will probably be embedded in tanks and fighter aircraft.