Thinking machines: Can there be? Are we?

Terry Winograd

1 Introduction

Futurologists have proclaimed the birth of a new species, *machina sapiens*, that will share (perhaps usurp) our place as the intelligent sovereigns of our earthly domain. These "thinking machines" will take over our burdensome mental chores, just as their mechanical predecessors were intended to eliminate physical drudgery. Eventually they will apply their "ultra-intelligence" to solving all of our problems. Any thoughts of resisting this inevitable evolution is just a form of "speciesism," born from a romantic and irrational attachment to the peculiarities of the human organism.

Critics have argued with equal fervor that "thinking machine" is an oxymoron – a contradiction in terms. Computers, with their foundations of cold logic, can never be creative or insightful or possess real judgment. No matter how competent they appear, they do not have the genuine intentionality that is at the heart of human understanding. The vain pretensions of those who seek to understand mind as computation can be dismissed as yet another demonstration of the arrogance of modern science.

Although my own understanding developed through active participation in artificial intelligence research, I have now come to recognize a larger grain of truth in the criticisms than in the enthusiastic predictions. But the story is more complex. The issues need not (perhaps cannot) be debated as fundamental questions concerning the place of humanity in the universe. Indeed, artificial intelligence has not achieved creativity, insight, and judgment. But its shortcomings are far more mundane: we have not yet been able to construct a machine with even a modicum of common sense or one that can converse on everyday topics in ordinary language.

The source of the difficulties will not be found in the details of silicon micro-circuits or of Boolean logic. The basic philosophy that has guided the research is shallow and inadequate, and has not received sufficient scrutiny. It is drawn from the traditions of rationalism and logical empiricism but has taken a novel turn away from its predecessors. This new "patchwork rationalism" will be our subject of examination.

First, we will review the guiding principles of artificial intelligence and see how they are embodied in current research. Then we will look at the fruits of that research. I will argue that "artificial intelligence" as now conceived is limited to a very particular kind of intelligence: one that can usefully be likened to bureaucracy in its rigidity, obtuseness, and inability to adapt to changing circumstances. The weakness comes not from insufficient development of the technology, but from the inadequacy of the basic tenets.

But, as with bureaucracy, weaknesses go hand in hand with unique strengths. Through a re-interpretation and re-formulation of the techniques that have been developed, we can anticipate and design appropriate and valuable uses. In conclusion I will briefly introduce an orientation I call hermeneutic constructivism and illustrate how it can lead down this alternative path of design.

2 The mechanization of rationality

In their quest for mechanical explanations of (or substitutes for) human reason, researchers in artificial intelligence are heirs to a long tradition. In his "Discourse on the method of properly guiding the reason in the search of truth in the sciences" (1637), Descartes initiated the quest for a systematic method of rationality. Although Descartes himself did not believe that reason could be achieved through mechanical devices, his understanding laid the groundwork for the symbol-processing machines of the modern age.

In 1651, Hobbes described reason as symbolic calculation:

> When a man reasoneth, he does nothing else but conceive a sum total, from addition of parcels; or conceive a remainder . . . These operations are not incident to numbers only, but to all manner of things that can be added together, and taken one out of another . . . the logicians teach the same in consequences of words; adding together two names to make an affirmation, and two affirmations to make a syllogism; and many syllogisms to make a demonstration. (*Quoted in Haugeland, 1985*)

Leibniz[1] cherished through his life the hope of discovering a kind of generalized mathematics, which he called *Characteristica Universalis*, by means of which thinking could be replaced by calculation. "If we had it," he says "we should be able to reason in metaphysics and morals in much the same way as in geometry and analysis. If controversies were to arise, there would be no more need of disputation between two philosophers than between two accountants. For it would suffice to take their pencils in their hands, to sit down to their slates, and to say to each other . . . 'Let us calculate'."

Behind this program of mechanical reason was a faith in a rational and ultimately understandable universe. The model of "Let us calculate" is that of Euclidean geometry, in which a small set of clear and self-evident postulates provides a basis for generating the right answers (given sufficient diligence) to the most complex and vexing problems. Reasonable men could be relied upon to agree on the postulates and the methods, and therefore dispute could only arise from mistaken calculation.

The empiricists turned to physical experience and experiment as the true basis of knowledge. But in rejecting the a priori status of the propositions on which reasoning was based, they did not abandon the vision of rigorous (potentially mechanizable) logical procedures. For our purposes here, it will suffice to adopt a broader characterization, in which much of both rationalism and empiricism fall within a common "rationalistic tradition." (Winograd and Flores, 1986). This label subsumes the varied (and at times hotly opposed) inheritors of Descartes' legacy – those who seek to achieve rational reason through a precise method of symbolic calculation.

The electronic computer gave new embodiment to mechanical rationality, making it possible to derive the consequences of precisely specified rules, even when huge amounts of calculation are required. The first decades of computing emphasized the application of numerical techniques. Researchers in operations research and decision theory addressed policy questions by developing complex mathematical models of social and political systems and calculating the results of proposed alternatives.[2] Although these techniques work well in specialized cases (such as scheduling delivery vehicles or controlling the operations in a refinery), they proved inadequate for the broader problems to which they were applied. The "mathematization" of experience required simplifications that made the computer results – accurate as they might be with respect to the models – meaningless in the world.

Although there are still attempts to quantify matters of social import (for example in applying mathematical risk analysis to decisions about nuclear power), there is an overall disillusionment with the potential for adequately reducing human concerns to a precise set of numbers and equations (see for example, Davis and Hersh, 1986). The developers of artificial intelligence have rejected traditional mathematical modeling in favor of an emphasis on symbolic – rather than numerical – formalisms. Leibniz's "Let us calculate" is taken in Hobbes' broader sense to include not just numbers but also "affirmations" and "syllogisms."

3 The promise of artificial intelligence

Attempts to duplicate formal non-numerical reasoning on a machine date back to the earliest computers, but the endeavor began in earnest with the AI projects of the mid 1950s (see Gardner, 1985, for an overview of the historical perspective). The goals were ambitious: to fully duplicate the human capacities of thought and language on a digital computer. Early claims that a complete theory of intelligence would be achieved within a few decades have long since been abandoned, but the research has not diminished. For example, a recent book by Minsky (one of the founders of AI) offers computational models for phenomena as diverse as conflict, pain and pleasure, the self, the soul, consciousness, confusion, genius, infant emotion, foreign accents, and freedom of will (these are among the section headings in Minsky, 1986).

In building models of mind, there are two distinct but complementary goals. On the one hand is the quest to explain human mental processes as thoroughly and unambiguously as physics explains the functioning of ordinary mechanical devices. On the other hand is the drive to create intelligent tools – machines that apply intelligence to serve some purpose, regardless of how closely they mimic the details of human intelligence. At times these two enterprises have gone hand in hand, at others they have led down separate paths.

Researchers such as Newell and Simon (two other founding fathers of artificial intelligence) have sought precise and scientifically testable theories of more modest scope than Minsky suggests. In reducing the study of mind to the formulation of rule-governed operations on symbol systems, they focus on detailed aspects of cognitive functioning, using empirical measures such as memory capacity and reaction time. They hypothesize specific "mental architectures" and compare their detailed performance with human experimental results (e.g., Newell and Simon, 1972; Laird *et al.*, 1986). It is difficult to measure the success of this enterprise. The tasks that have been examined (such as puzzle-solving and the ability to remember abbreviations for computer commands) do not even begin to approach a representative sample of human cognitive abilities, for reasons we will examine below.

On the other side lies the goal of practical system building. In the late 1970s, the field of artificial intelligence was drastically affected by the continuing precipitous drop in computing costs. Techniques that previously demanded highly specialized and costly equipment came within the reach of commercial users. A new term, "knowledge engineering," was coined to indicate a shift to the pragmatic interests of the engineer, rather than the scientist's search for theoretical knowledge.

"Expert systems," as the new programs were called, incorporate "knowledge bases" made up of simple facts and "if . . . then" rules, as illustrated in figure 1.

FACTS:

Tank no. 23 contains sulfuric acid.

The plaintiff was injured by a portable power saw.

RULES:

If the sulfate ion test is positive, the spill material is sulfuric acid.

If the plaintiff was negligent in the use of the product,

the theory of contributory negligence applies.

Figure 1 Rules for an expert system (from D. Waterman, 1986, p. 16)

These systems do not attempt to explain human intelligence in detail, but are justified in terms of their practical applications, for which extravagant claims have been made.

> Humans need expert systems, but the problem is they don't often believe it. . . At least one high-performance medical diagnosis program sits unused because the physicians it was designed to assist didn't perceive that they needed such assistance; they were wrong, but that doesn't matter . . . There's a manifest destiny in information processing, in knowledge systems, a continent we shall all spread out upon sooner or later (*Feigenbaum and McCorduck, 1983*).

The high hopes and ambitious aspirations of knowledge engineering are well-documented, and the claims are often taken at face value, even in serious intellectual discussions. In fact, although a few widely-known systems illustrate specific potentials, the successes are still isolated pinnacles in a landscape of research prototypes, feasibility studies, and preliminary versions. It is difficult to get a clear picture of what has been accomplished and to make a realistic assessment of what is yet to come. We need to begin by examining the difficulties with the fundamental methods these programs employ.

4 The foundations of artificial intelligence

Artificial intelligence draws its appeal from the same ideas of mechanized reasoning that attracted Descartes, Leibniz, and Hobbes, but it differs from the more classical forms of rationalism in a critical way. Descartes wanted his method to stand on a bedrock of clear and self-evident truths. Logical empiricism sought truth through observation and the refinement of formal theories that predicted experimental results. Artificial intelligence has abandoned the quest for certainty and truth. The new patchwork rationalism is built upon mounds of "micro-truths" gleaned through common sense introspection, *ad hoc* programming and so-called "knowledge acquisition" techniques for interviewing experts. The grounding on this shifting sand is pragmatic in the crude sense – "If it seems to be working, it's right."

The resulting patchwork defies logic. Minsky observes:

> For generations, scientists and philosophers have tried to explain ordinary reasoning in terms of logical principles – with virtually no success. I suspect this enterprise failed because it was looking in the wrong direction: common sense works so well not because it is an approximation of logic; logic is only a small part of our great accumulation of different, useful ways to chain things together. (*Minsky, 1986*)

In the days before computing, "ways to chain things together" would have remained a vague metaphor. But the computer can perform arbitrary symbol manipulations that we interpret as having logical import. It is easy to build a program to which we enter "Most birds can fly" and "Tweety is a bird" and which then produces "Tweety can fly" according to a regular (although logically questionable) rule. The artificial intelligence methodology does not demand a logically correct answer, but one that works sufficiently often to be "heuristically adequate."

In a way, this approach is very attractive. Everyday human thought does not follow the rigid strictures of formal deduction. Perhaps we can devise some more flexible (and even fallible) system that operates according to mechanical principles, but more accurately mirrors the mind.

But this appeal is subtly deceptive. Minsky places the blame for lack of success in explaining ordinary reasoning on the rigidity of logic, and does not raise the more fundamental questions about the nature of all symbolic representations and of formal (though possibly "non-logical") systems of rules for manipulating them. There are basic limits to what can be done with symbol manipulation, regardless of how many "different, useful ways to chain things together" one invents. The reduction of mind to the interactive sum of decontextualized fragments is ultimately impossible and misleading. But before elaborating on the problems, let us first review some assumptions on which this work proceeds:

1 Intelligence is exhibited by "physical symbol systems."
2 These systems carry out symbol manipulations that correspond to some kind of "problem solving."
3 Intelligence is embodied as a large collection of fragments of "knowledge."

4.1 The physical symbol system hypothesis

The fundamental principle is the identification of intelligence with the functioning of a rule-governed symbol-manipulating device. It has been most explicitly stated by Newell and Simon:

> A physical symbol system has the necessary and sufficient means for general intelligent action . . . By "general intelligent action" we wish to indicate the same scope of intelligence we see in human action: that in any real situation behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some limits of speed and complexity. (*Newell and Simon, 1976*)

This "physical symbol system hypothesis" presupposes materialism: the claim that all of the observed properties of intelligent beings can ultimately be explained in terms of lawful physical processes. It adds the claim that these processes can be described at a level of abstraction in which all relevant aspects of physical state can be understood as the encoding of symbol structures and that the activities can be adequately characterized as systematic application of symbol manipulation rules.

The essential link is *representation* – the encoding of the relevant aspects of the world. Newell lays this out explicitly:

> An intelligent agent is embedded in a *task environment*; a *task statement* enters via a *perceptual* component and is encoded in an initial *representation*. Whence starts a cycle of activity in which a *recognition* occurs . . . of a method to use to attempt the problem. The method draws upon a memory of *general world knowledge* . . . It is clear to us all what *representation* is in this picture. It is the data structures that hold the problem and will be processed into a form that makes the solution available. Additionally, it is the data structures that hold the world knowledge and will be processed to acquire parts of the solution or to obtain guidance in constructing it. [emphasis in original] (*Newell, 1982*)

Complete and systematic symbolic representation is crucial to the paradigm. The rules followed by the machine can deal only with the symbols, not their interpretation.

4.2 Problem-solving, inference, and search

Newell's and Simon's physical symbol systems aspire not to an idealized rationality, but to "behavior appropriate to the ends of the system and adaptive to the demands of the environment." This shift reflects the formulation that won Simon a Nobel prize in economics. He supplanted decision theories based on optimization with a theory of "satisficing" – effectively using finite decision-making resources to come up with adequate, but not necessarily optimal plans of action.

As artificial intelligence developed in the 1950s and 1960s, this methodology was formalized in the techniques of "heuristic search."

The task that a symbol system is faced with, then, when it is presented with a problem and a problem space, is to use its limited processing resources to generate possible solutions, one after another, until it finds one that satisfies the problem-defining test (Newell and Simon, 1976).

The "problem space" is a formal structure that can be thought of as enumerating the results of all possible sequences of actions that might be taken by the program. In a program for playing chess, for example, the problem space is generated by the possible sequences of moves. The number of possibilities grows exponentially with the number of moves, and is beyond practical reach after a small number. However, one can limit search in this space by following heuristics that operate on the basis of local cues ("If one of your pieces could be taken on the opponent's next move, try moving it . . ."). There have been a number of variations on this basic theme, all of which are based on explicit representations of the problem space and the heuristics for operating within it.

Figure 1 illustrated some rules and facts from expert systems. These are not represented in the computer as sentences in English, but as symbols intended to correspond to the natural language terms. As these examples indicate, the domains are naturally far richer and more complex than can be captured by such simple rules. A lawyer will have many questions about whether a plaintiff was "negligent," but for the program it is a simple matter of whether a certain symbolic expression of the form "Negligent(x)" appears in the store of representations, or whether there is a rule of the form "If . . . then Negligent(x)," whose conditions can be satisfied.

There has been a great deal of technical debate over the detailed form of rules, but two principles are taken for granted in essentially all of the work:

1 Each rule is true in a limited (situation-dependent), not absolute sense.
2 The overall result derives from the synergistic combination of rules, in a pattern that need not (in fact could not in general) be anticipated in writing them.

For example, there may be cases in which the "sulfate ion test is positive" even though the spill is not sulfuric acid. The overall architecture of the rule-manipulating system may lead to a conclusion being drawn that violates one of these rules (on the basis of other rules). The question is not whether each of the rules is true, but whether the output of the program as a whole is "appropriate." The knowledge engineers hope that by devising and tuning such rules they can capture more than the deductive logic of the domain.

> "While conventional programs deal with facts, expert systems handle "lore" . . . the rules of thumb, the hunches, the intuition and capacity for judgement that are seldom explicitly laid down but which form the basis of an expert's skill, acquired over a lifetime's experience." (*Michie and Johnston, 1984*)

This *ad hoc* nature of the logic applies equally to the cognitive models of Newell and Simon, in which a large collection of separate "production rules" operates on a symbolic store or "working memory." Each production rule specifies a step to be carried out on the symbols in the store, and the overall architecture determines which will be carried out in what order. The symbols don't stand for chemical spills and law, but for hypothesized psychological features, such as the symbolic contents of short-term memory. Individual rules do things like moving an element to the front of the memory or erasing it. The cognitive modeler does not build an overall model of the system's performance on a task, but designs the individual rules in hopes that appropriate behavior will emerge from their interaction.

Minsky makes explicit this assumption that intelligence will emerge from computational interactions among a plethora of small pieces.

> I'll call 'Society of Mind' this scheme in which each mind is made of many smaller processes. The smaller processes we'll call agents. Each mental agent by itself can only do some simple thing that needs no mind or thought at all. Yet when we join these agents in societies – in certain very special ways – this leads to true intelligence. (*Minsky, 1986*)

Minsky's theory is quite different from Newell's cognitive architectures. In place of finely-tuned clockworks of precise production rules we find an impressionistic pastiche of metaphors. Minsky illustrates his view in a simple 'micro-world' of toy blocks, populated by agents such as BUILDER (which stacks up the blocks), ADD (which adds a single block to a stack), and the like:

> for example, BUILDER's agents require no sense of meaning to do their work; ADD merely has to turn on GET and PUT. Then GET and PUT do not need any subtle sense of what those turn-on signals "mean" – because they're wired up to do only what they're wired up to do. (*Minsky, 1986*)

These agents seem like simple computer subroutines – program fragments that perform a single well-defined task. But a subsequent chapter describes an interaction between the BUILDER agent and the WRECKER agent, which are parts of a PLAY-WITH-BLOCKS agent:

> inside an actual child, the agencies responsible for BUILDING and WRECKING might indeed become versatile enough to negotiate by offering support for one another's goals. "Please, WRECKER, wait a moment more till BUILDER adds just one more block: it's worth it for louder crash!". (*Minsky, 1986*).

With a simple "might indeed become versatile . . .," we have slipped from a technically feasible but limited notion of agents as subroutines, to an impressionistic description of a society of *homunculi*, conversing with each other in ordinary language. This sleight of hand is at the center of the theory. It takes an almost childish leap of faith to assume that the modes of explanation that work for the details of block manipulation will be adequate for understanding conflict, consciousness, genius, and freedom of will.

One cannot dismiss this as an isolated fantasy. Minsky is one of the major figures in artificial intelligence and he is only stating in a simplistic form a view that permeates the field.

In looking at the development of computer technology, one cannot help but be struck by the successes at reducing complex and varied tasks to systematic combinations of elementary operations. Why not, then, make the jump to the mind. If we are no more than protoplasm-based physical symbol systems, the reduction must be possible and only our current lack of knowledge prevents us from explicating it in detail, all the way from BUILDER's clever ploy down to the logical circuitry.

4.3 Knowledge as a commodity

All of the approaches described above depend on interactions among large numbers of individual elements: rules, productions, or agents. No one of these elements can be taken as representing a substantial understandable truth, but this doesn't matter since somehow the conglomeration will come out all right. But how can we have any confidence that it will? The proposed answer is a typical one of our modern society: "More is better!" "Knowledge is power, and more knowledge is more power."

A widely-used expert systems text declares:

> It wasn't until the late 1970s that AI scientists began to realize something quite important: The problem-solving power of a program comes from the knowledge it possesses, not just from the formalisms and inference schemes it employs. The conceptual breakthrough was made and can be quite simply stated. *To make a program intelligent, provide it with lots of high-quality, specific knowledge about some problem area.* [emphasis in the original] (*Waterman, 1986*)

This statement is typical of much writing on expert systems, both in the parochial perspective that inflates homily into "conceptual breakthrough" and in its use of slogans like "high-quality knowledge." Michie (the doyen of artificial intelligence in Britain) predicts:

> [Expert systems] . . . can actually help to codify and improve expert human knowledge, taking what was fragmentary, inconsistent and error-infested and turning it into knowledge that is more precise, reliable and comprehensive. This new process, with its enormous potential for the future, we call "knowledge refining." (*Michie and Johnston, 1984*)

Feigenbaum proclaims:

> The miracle product is knowledge, and the Japanese are planning to package and sell it the way other nations package and sell energy, food, or manufactured goods . . . The essence of the computer revolution is that the burden of producing the future knowledge of the world will be transferred from human heads to machine artifacts. (*Feigenbaum and McCorduck, 1983*)

Knowledge is a kind of commodity – to be produced, refined, and packaged. The knowledge engineers are not concerned with the age-old epistemological problems of what constitutes knowledge or understanding. They are hard at work on techniques of "knowledge acquisition" and see it as just a matter of sufficient money and effort:

> We have the opportunity at this moment to do a new version of Diderot's *Encyclopedia*, a gathering up of all knowledge – not just the academic kind, but the informal, experiential, heuristic kind – to be fused, amplified, and distributed, all at orders of magnitude difference in cost, speed, volume, and *usefulness* over what we have now. [emphasis in the original] (*Feigenbaum and McCorduck, 1983*)

Lenat has embarked on this task of "encod[ing] all the world's knowledge down to some level of detail." The plan projects an initial entry of about 400 articles from a desk encyclopedia (leading to 10,000 paragraphs worth of material), followed by hiring a large number of "knowledge enterers" to add "the last 99 percent." There is little concern that foundational problems might get in the way. Lenat *et al.* (1986) asserts that "AI has for many years understood enough about representation and inference to tackle this project, but no one has sat down and done it."

5 The fundamental problems

The optimistic claims for artificial intelligence have far outstripped the achievements, both in the theoretical enterprise of cognitive modeling and in the practical application of expert systems.

In cognitive modeling we seek to fit a model's performance with measured human behavior but the enterprise is fraught with methodological difficulty, as it straddles the wide chasm between the engineering bravado of computer science and the careful empiricism of experimental psychology. When a computer program duplicates to some degree some carefully restricted aspect of human behavior, what have we learned? It is all too easy to write a program that would produce that particular behavior, and all too hard to build one that covers a sufficiently general range to inspire confidence. As Pylyshyn (an enthusiastic participant in cognitive science) observes:

> Most current computational models of cognition are vastly underconstrained and *ad hoc*; they are contrivances assembled to mimic

> arbitrary pieces of behavior, with insufficient concern for explicating the principles in virtue of which such behavior is exhibited and with little regard for a precise understanding. (*Pylyshyn, 1984*)

Newell and his colleagues' painstaking attention to detailed architecture of production systems is an attempt to better constrain the computational model, in hopes that experiments can then test detailed hypotheses. As with much of experimental psychology, a highly artificial experimental situation is required to get results that can be sensibly interpreted at all. Proponents argue that the methods and theoretical foundations that are being applied to micro-behavior will eventually be extended and generalized to cover the full range of cognitive phenomena. As with Minsky, this leap from the micro-structure to the whole human is one of faith.

In the case of expert systems, there is a more immediate concern. Applied AI is widely seen as a means of managing processes that have grown too complex or too rapid for unassisted humans. Major industrial and governmental organizations are mounting serious efforts to build expert systems for tasks such as air traffic control, nuclear power plant operation, and – most distressingly – the control of weapons systems. These projects are justified with claims of generality and flexibility for AI programs. They ignore or downplay the difficulties that will make the programs almost certain to fail in just those cases where their success is most critical.

It is commonplace in the field to describe expert systems as "brittle" – able to operate only within a narrow range of situations. The problem here is not just one of insufficient engineering, but is a direct consequence of the nature of rule-based systems. We will examine three manifestations of the problem: gaps of anticipation; blindness of representation; and restriction of the domain.

5.1 Gaps of anticipation

In creating a program or knowledge base, one takes into account as many factors and connections as feasible. But in any realistically complex domain, this gives at best a spotty coverage. The person designing a system for dealing with acid spills may not consider the possibility of rain leaking into the building, or of a power failure, or that a labeled bottle does not contain what it purports to. A human expert faced with a problem in such a circumstance falls back on common sense and a general background of knowledge.

The hope of patchwork rationalism is that with a sufficiently large body of rules, the thought-through spots will successfully interpolate to the wastelands in between. Having written rule A with one circumstance in mind and rule B with another, the two rules in combination will succeed in yet a third. This strategy is the justification for the claim that AI systems are more flexible than conventional programs. There is a grain of truth in the comparison, but it is deceptive. The program applies the rules blindly with erratic results. In many cases, the price of flexibility (the ability to operate in combinations of contingencies not considered by the programmer) is irreparable and inscrutable failure.

In attempting to overcome this brittleness, expert systems are built with many thousands of rules, trying to cover all of the relevant situations and to provide representations for all potentially relevant aspects of context. One system for medical diagnosis, called CADUCEUS (originally INTERNIST) has 500 disease profiles, 350 disease variations, several thousand symptoms, and 6,500 rules describing relations among symptoms. After fifteen years of development, the system is still not on the market. According to one report, it gave a correct diagnosis in only 75 per cent of its carefully selected test cases. Nevertheless, Myers, the medical expert who developed it, "believes that the addition of another 50 [diseases] will make the system workable and, more importantly, practical" (Newquist, 1987).

Human experts develop their skills through observing and acting in many thousands of cases. AI researchers argue that this results in their remembering a huge repertoire of specialized "patterns" (complex symbolic rules) that allows them to discriminate situations with expert finesse and to recognize appropriate actions. But it is far from obvious whether the result of experience can be adequately formalized as a repertoire of discrete patterns (see discussion in Dreyfus and Dreyfus, 1986). To say that "all of the world's knowledge" could be explicitly articulated in any symbolic form (computational or not) we must assume the possibility of reducing all forms of tacit knowledge (skills, intuition, and the like) to explicit facts and rules. Heidegger and other phenomenologists have challenged this, and many of the strongest criticisms of artificial intelligence are based on the phenomenological analysis of human understanding as a "readiness-at-hand" of action in the world, rather than as the manipulation of "present-at-hand" representations (see, for example Dreyfus, 1979; Winograd and Flores, 1986).

Be that as it may, it is clear that the corresponding task in building expert systems is extremely difficult, if not theoretically impossible. The knowledge engineer attempts to provide the program with rules that correspond to the expert's experience. The rules are modified through analyzing examples in which the original rules break down. But the patchwork nature of the rules makes this extremely difficult. Failure in a particular case may not be attributable to a particular rule, but rather to a chance combination of rules that are in other circumstances quite useful. The breakdown may not even provide sharp criteria for knowing what to change, as with a chess program that is just failing to come up with good moves. The problem here is not simply one of scale or computational complexity. Computers are perfectly capable of operating on millions of elements. The problem is one of human understanding – the ability of a person to understand how a new situation experienced in the world is related to an existing set of representations, and to possible modifications of those representations.

In trying to remove the potentially unreliable "human element," expert systems conceal it. The power plant will no longer fail because a reactor-operator falls asleep, but because a knowledge engineer didn't think of putting in a rule specifying how to handle a particular failure when the emergency system is undergoing its periodic test, and the backup system is out of order. No amount of refinement and articulation can guarantee the absence of such breakdowns. The hope that a system based on patchwork rationalism will respond "appropriately" in such cases is just that: a hope, and one that can engender dangerous illusions of safety and security.

5.2 The blindness of representation

The second problem lies in the symbol system hypothesis itself. In order to characterize a situation in symbolic form, one uses a system of basic distinctions, or terms. Rules deal with the interrelations among the terms, not with their interpretations in the world.

Consider ordinary words as an analogy. Imagine that a doctor asks a nurse "Is the patient eating?" If they are deciding whether to perform an examination, the request might be paraphrased "Is she eating at this moment?" If the patient is in the hospital for anorexia and the doctor is checking the efficacy of the treatment, it might be more like "Has the patient eaten some minimal amount in the past day?" If the patient has recently undergone surgery, it might mean "Has the patient taken any nutrition by mouth," and so on. In responding, a person interprets the sentence as having relevance in the current situation, and will typically respond appropriately without conscious choosing among meanings.

In order to build a successful symbol system, decontextualized meaning is necessary – terms must be stripped of open-ended ambiguities and shadings. A medical expert system might have a rule of the form: "IF Eating(x) THEN . . . ", which is to be applied only if the patient is eating, along with others of the form "IF . . . THEN Eating (x)" which determine when that condition holds. Unless everyone who writes or reads a rule interprets the primitive term "Eating" in the same way, the rules have no consistent interpretation and the results are unpredictable.

In response to this, one can try to refine the vocabulary. "Currently-Dining" and "Taking-Solids" could replace the more generic term, or we could add construal rules, such as "in a context of immediate action, take 'Eating' to mean 'Currently-Dining'." Such approaches work for the cases that programmers anticipate, but of course are subject to the infinite regress of trying to decontextualize context. The new terms or rules themselves depend on interpretation that is not represented in the system.

5.3 Restriction of the domain

A consequence of decontextualized representation is the difficulty of creating AI programs in any but the most carefully restricted domains, where almost all of the knowledge required to perform the task is special to that domain (i.e., little common-sense knowledge is required). One can find specialized tasks for which appropriate limitations can be achieved, but these do not include the majority of work in commerce, medicine, law, or the other professions demanding expertise.

Holt characterized the situation: "A brilliant chess move while the room is filling with smoke because the house is burning down does not show intelligence. If the capacity for brilliant chess moves without regard to life circumstances deserves a name, I would naturally call it 'artificial intelligence.'"[3]

The brilliance of a move is with respect to a well-defined domain: the rules of chess. But acting as an expert doctor, attorney, or engineer takes the other kind of intelligence: knowing what makes sense in a situation. The most successful artificial intelligence programs have operated in the detached puzzle-like domains of board games and technical analysis, not those demanding understanding of human lives, motivations, and social interaction. Attempts to cross into these difficult territories, such as a program said to "understand tales involving friendship and adultery" (see discussion of BORIS program in Winograd and Flores, 1986), proceed by replacing the real situation with a cartoon-like caricature, governed by simplistic rules whose inadequacy is immediately obvious (even to the creators, who argue that they simply need further elaboration).

This reformulation of a domain to a narrower, more precise one can lead to systems that give correct answers to irrelevant problems. This is of concern not only when actions are based directly on the output of the computer system (as in one controlling weapons systems), but also when, for example, medical expert systems are used to evaluate the work of physicians (Athanasiou, 1987). Since the system is based on a reduced representation of the situation, it systematically (if invisibly) values some aspects of care while remaining blind to others. Doctors whose salaries, promotions, or accreditations depend on the review of their actions by such a program will find their practice being subtly shaped to its mold. The attempt to encode "the world's knowledge" inevitably leads to this kind of simplification. Every explicit representation of knowledge bears within it a background of cultural orientation that does not appear as explicit claims, but is manifest in the very terms in which the 'facts' are expressed and in the judgment of what constitutes a fact. An encyclopedia is not a compendium of "refined knowledge," but a statement within a tradition and a culture. By calling an electronic encyclopedia a 'knowledge base' we mystify its source and its grounding in a tradition and background.

6 The bureaucracy of mind

Many observers have noted the natural affinity between computers and bureaucracy. Lee argues that "bureaucracies are the most ubiquitous form of artificial intelligence . . . Just as scientific management found its idealization in automation and programmable production robots, one might consider an artificially intelligent knowledge-based system as the ideal bureaucrat" (Lee, 1985). Lee's stated goal is "improved bureaucratic software engineering," but his analogy suggests more.

Stated simply, *the techniques of artificial intelligence are to the mind what bureaucracy is to human social interaction.*

In today's popular discussion, bureaucracy is seen as an evil – a pathology of large organizations and repressive governments. But in his classic work on bureaucracy, Weber argued its great advantages over earlier, less formalized systems, calling it the "unambiguous yardstick for the modernization of the state." He notes that "bureaucracy has a 'rational' character, with rules, means-ends calculus, and matter-of-factness predominating," (Weber, 1968), and that it succeeds in "eliminating from official business love, hatred, and all purely personal, irrational, and emotional elements which escape calculation" (Weber, 1968).

> The decisive reason for the advance of bureaucratic organization has always been its purely *technical* superiority over any other form of organization. The fully developed bureaucratic apparatus compares with other organizations exactly as does the machine with the non-mechanical modes of production. Precision, speed, unambiguity, knowledge of the files, continuity, discretion, unity, strict subordination, reduction of friction and of material and personal costs – these are raised to the optimum point in the strictly bureaucratic administration. [emphasis in original] (*Weber, 1968*)

The benefits of bureaucracy follow from the reduction of judgment to the systematic application of explicitly articulated rules. Bureaucracy achieves a predictability and manageability that is missing in earlier forms of organization. There are striking similarities here with the arguments given for the benefits of expert systems, and equally striking analogies with the shortcomings as pointed out, for example, by March and Simon:

> The reduction in personalized relationships, the increased internalization of rules, and the decreased search for alternatives combine to make the behavior of members of the organization highly predictable; i.e., they result in an increase in the *rigidity of behavior* of participants [which] increases the *amount of difficulty with clients* of the organization and complicates the achievement of client satisfaction. [emphasis in original] (*March and Simon, 1958*)

Given Simon's role in artificial intelligence, it is ironic that he notes these weaknesses of human-embodied rule systems, but sees the behavior of rule-based physical symbol systems as "adaptive to the demands of the environment." Indeed, systems based on symbol manipulation exhibit the rigidities of bureaucracies, and are most problematic in dealing with "client satisfaction" – the mismatch between the decontextualized application of rules and the human interpretation of the symbols that appear in them. Bureaucracy is most successful in a world that is stable and repetitive – where the rules can be followed without interpretive judgments. Expert systems are best in just the same situations. Their successes have been in stable and precise technical areas, where exceptions are not the rule.

Michie's claim that expert systems can encode "the rules of thumb, the hunches, the intuition and capacity for judgement . . ." is wrong in the same way that it is wrong to seek a full account of an organization in its formal rules and procedures. Modern sociologists have gone beyond Weber's analysis, pointing to the informal organization and tacit knowledge that make organizations work effectively. This closely parallels the importance of tacit knowledge in individual expertise. Without it we get rigidity and occasional but irreparable failure.

The depersonalization of knowledge in expert systems also has obvious parallels with bureaucracy. When a person views his or her job as the correct application of a set of rules (whether human-invoked or computer-based), there is a loss of personal responsibility or commitment. The "I just follow the rules" of the bureaucratic clerk has its direct analog in "That's what the knowledge base says." The individual is not committed to appropriate results (as judged in some larger human context), but to faithful application of the procedures. This forgetfulness of individual commitment is perhaps the most subtle and dangerous consequence of patchwork rationality. The person who puts rules into a knowledge base cannot be committed to the consequences of applying them in a situation he or she cannot foresee. The person who applies them cannot be committed to their formulation or to the mechanics by which they produce an answer. The result belongs to no-one. When we speak here of "commitment," we mean something more general than the kind of accountability that is argued in court. There is a deep sense in which every use of language is a reflection of commitment, as we will see in the following section.

7 Alternatives

We began with the question of thinking machines – devices that mechanically reproduce human capacities of thought and language. We have seen how this question has been reformulated in the pursuit of artificial intelligence, to reflect a particular design based on patchwork rationalism. We have argued that the current direction will be inadequate to explain or construct real intelligence.

But, one might ask, does that mean that no machine could exhibit intelligence? Is artificial intelligence inherently impossible, or is it just fiendishly difficult? To answer sensibly we must first ask what we mean by "machine." There is a simple a priori proof that machines can be intelligent if we accept that our own brains are (in Minsky's provocative words) nothing but "meat machines." If we take "machine" to stand for any physically constituted device subject to the causal laws of nature, then the question reduces to one of materialism, and is not to be resolved through computer research. If, on the other hand, we take machine to mean "physical symbol system" then there is ground for a strong skepticism. This skepticism has become visible among practitioners of artificial intelligence as well as the critics.

7.1 Emergent intelligence

The innovative ideas of cybernetics a few decades ago led to two contrasting research programmes. One, which we have examined here, took the course of symbol processing. The other was based on modelling neural activity and led to the work on "perceptrons," a research line that was discounted for many years as fruitless and is now being rehabilitated in "connectionist" theories, based on "massively parallel distributed processing." In this work, each computing element (analogous to a neuron) operates on simple general principles, and intelligence emerges from the evolving patterns of interaction.[4]

Connectionism is one manifestation of what Turkle calls "emergent AI" (Turkle, 1987). The fundamental intuition guiding this work is that cognitive structure in organisms emerges through learning and experience, not through explicit representation and programming. The problems of blindness and domain limitation described above need not apply to a system that has developed through situated experience.

It is not yet clear whether we will see a turn back towards the heritage of cybernetics or simply a "massively parallel" variant of current cognitive theory and symbol processing design. Although the new connectionism may breathe new life into cognitive modeling research, it suffers an uneasy balance between symbolic and physiological description. Its spirit harks back to the cybernetic concern with real biological systems, but the detailed models typically assume a simplistic representational base much closer to traditional artificial intelligence. Connectionism, like its parent cognitive theory, must be placed in the category of brash unproved hypotheses, which have not really begun to deal with the complexities of mind, and whose current explanatory power is extremely limited.

In one of the earliest critiques of artificial intelligence, Dreyfus compared

it to alchemy (Dreyfus, 1965). Seekers after the glitter of intelligence are misguided in trying to cast it from the base metal of computing. There is an amusing epilogue to this analogy: in fact, the alchemists were right. Lead can be converted into gold by a particle accelerator hurling appropriate beams at lead targets. The AI visionaries may be right in the same way, and they are likely to be wrong in the same way. There is no reason but hubris to believe that we are any closer to understanding intelligence than the alchemists were to the secrets of nuclear physics. The ability to create a glistening simulacrum should not fool us either into thinking the rest is "just a matter of encoding a sufficient part of the world's knowledge" or into a quest for the philosopher's stone of "massively parallel processing."

7.2 Hermeneutic constructivism

Discussions of the problems and dangers of computers often leave the impression that on the whole we would be better-off if we could return to the pre-computer era. In a similar vein one might decry the advent of written language, which created many new problems. For example, Weber attributes the emergence of bureaucracy to the spread of writing and literacy, which made it possible to create and maintain systems of rules. Indeed, the written word made bureaucracy possible, but that is far from a full account of its relevance to human society.

The computer is a physical embodiment of the symbolic calculations envisaged by Hobbes and Leibniz. As such, it is not really a thinking machine, but a language machine. The very notion of "symbol system" is inherently linguistic and what we duplicate in our programs with their rules and propositions is really a form of verbal argument, not the workings of mind. It is tempting – but ultimately misleading – to project the image of rational discourse (and its reflection in conscious introspection) onto the design of embodied intelligence. In taking inner discourse as a model for the activity of Minsky's tiny agents, or of productions that determine what token to process next, artificial intelligence has operated with the faith that mind is linguistic down to the microscopic level.

But the utility of the technology need not depend on this faith. The computer, like writing, is fundamentally a communication medium – one that is unique in its ability to perform complex manipulations on the linguistic objects it stores and transmits. We can reinterpret the technology of artificial intelligence in a new background, with new consequences. In doing so we draw on an alternative philosophical grounding, which I will call hermeneutic constructivism.

We begin with some fundamental questions about what language is and how it works. In this we draw on work in hermeneutics (the study of interpretation) and phenomenology, as developed by Heidegger and Gadamer, along with the concepts of language action developed from the later works of Wittgenstein and through the speech act philosophy of Austin, Searle, and Habermas (see chapter 5 of Winograd and Flores, 1986).

Two guiding principles emerge:

1 People create their world through language.
2 Language is always interpreted in a tacitly understood background.

Austin pointed out that "performative" sentences do not convey information about the world, but act to change that world. "You're hired," when uttered in appropriate conditions, creates – not describes – a situation of employment. Searle applied this insight to mundane language actions such as asking questions and agreeing to do something. Habermas extended it further, showing how sentences we would naively consider statements of fact have force by virtue of an act of commitment by the speaker.

> The essential presupposition for the success of [a language] act consists in the speaker's entering into a specific engagement, so that the hearer can rely on him. An utterance can count as a promise, assertion, request, question, or avowal, if and only if the speaker makes an offer that he is ready to make good insofar as it is accepted by the hearer. The speaker must engage himself, that is, indicate that in certain situations he will draw certain consequences for action. *(Habermas, 1979)*

Descartes' descendants in the rationalistic tradition take the language of mathematics as their ideal. Terms are either primitive or can be fully defined; the grammar is unambiguous; and precise truth conditions can be established through formal techniques. But even in apparently simple and straightforward situations, human language is metaphorical, ambiguous and undefinable. What we can take as fundamental is the engagement – the commitment to make good what cannot be fully made precise.

This grounding is especially evident for statements of the kind that Roszak characterizes as "ideas rather than information" (Roszak, 1986). "All men are created equal" cannot be judged as a true or false description of the objective world. Its force resides in the commitments it carries for further characterization and further action. But it is critical to recognize that this social grounding of language applies equally to the mundane statements of everyday life. "The patient is eating" cannot be held up to any specific set of truth conditions across situations in which it may be uttered. The speaker is not reporting an objectively delineated state of affairs, but indicating the "engagement" to enter sincerely into a dialogue of articulation of the relevant background.

This unavoidable dependence of interpretation on unspoken background is the fundamental insight of the hermeneutic phenomenologists, such as Gadamer. It applies not just to ordinary language, but to every symbolic representation as well. We all recognize that in "reducing things to numbers" we lose the potential for interpretation in a background. But this is equally true of "reducing them to symbol structures."

Whenever a computer program is intended to guide or take action in a human domain, it inevitably imports basic values and assumptions. The basic nature of patchwork rationalism obscures the underlying constitutive "ideas" with a mosaic of fragmentary bits of "information." The social and political agenda concealed behind these patches of decontextualized and depersonalized belief is dangerous in its invisibility.

7.3 Language machines

Symbol structures are ultimately created by people and interpreted by people. The computer, as a language machine, manipulates symbols without respect to their interpretation. To the extent that relations among the meanings can be adequately reflected in precise rules, the computational manipulations make sense. The error is in assuming that these manipulations capture, rather than augment or reify parts of the meaning. If an expert system prints out "Give the patient penicillin" or "Fire the missiles now," room for interpretation is limited and meaning is lost. But instead we can see the computer as a way of organizing, searching, and manipulating texts that are created by people, in a context, and ultimately intended for human interpretation.

We are already beginning to see a movement away from the early vision of computers replacing human experts. For example, the medical diagnostic system described above is being converted from "Internist" (a doctor specializing in internal medicine) to an "advisory system" called "QMR" (for "Quick Medical Reference") (Newquist, 1987). The rules can be thought of as constituting an automated textbook, which can access and logically combine entries that are relevant to a particular case. The goal is to suggest and justify possibilities a doctor might not otherwise have considered. The program need not respond with an evaluation or plan for action, but is successful through providing relevant material for interpretation by an expert. Similarly, in areas of real-time control (like a nuclear power plant), an advisory system can monitor conditions and provide warnings, reports, and summaries for human review. In a similar vein, an interactive computer-based encyclopedia need not cover all of human knowledge or provide general purpose deduction in order to take advantage of the obvious computer capacities of speed, volume, and sophisticated inferential indexing.

Another opportunity for design is in the regularities of the structure of language use. As a simple example, a request is normally followed in coherent conversation by an acceptanee, a rejection, or a request to modify the conditions. These in turn are followed by other language acts in a logic of "conversation for action" oriented towards completion (a state in which neither party is awaiting further action by the other). The theory of such conversations has been developed as the basis for a computer program called The Coordinator which is used for facilitating and organizing computer-message conversations in an organization (see Flores, 1982; Winograd and Flores, 1986; Winograd, 1987/88). It emphasizes the role of commitment by the speaker in each speech act and provides the basis for timely and effective action.

Howard has studied the use of computer systems by professionals evaluating loan applications for the World Bank. He argues that their use of computers while on field missions increases the "transparency" of their decision-making process, hence increasing their accountability and enhancing opportunities for meaningful negotiation. The computer serves as a medium of discourse in which different commitments and their consequences can be jointly explored.

> As a result, the dialogue between them [the bankers and their clients] suddenly becomes less about the final results – "the numbers" – and more about the assumptions behind the numbers, the criteria on which decisions are themselves based . . .[quoting a bank professional] "Instead of just saying, 'I don't believe you, my opinion is X,' we explore it. We say, 'let's see what the consequences of that are.' And, sometimes, we end up changing our assumptions." (*Howard, 1986*)

Current expert systems methodologies are not well suited to this kind of dialogue. They separate the construction of the knowledge base from the use of its "expertise."

The experts (with the help of knowledge engineers) enter the knowledge in the laboratory, and the users apply it in the field to get results. But we might instead use the computer to support the discourse that creates the reality – as a tool for the cooperative articulation of the characterizations and rules that will be applied. Rather than seeing the computer as working with objectified refined knowledge, it can serve as a way of keeping track of how the representations emerge from interpretations: who created them in what context, and where to look for clarification.

8 Conclusion

The question of our title demands interpretation in a context. As developed in the paper, it might be formulated more precisely "Are we machines of the kind that researchers are building as 'thinking machines'?" In asking this kind of question we engage in a kind of projection – understanding humanity by projecting an image of ourself onto the machine and the image of the machine back onto ourselves. In the tradition of artificial intelligence, we project an image of our language activity onto the symbolic manipulations of the machine, then project that back onto the full human mind.

But these projections are like the geometric projection of a three-dimensional world onto a two-dimensional plane. We systematically eliminate dimensions, thereby both simplifying and distorting. The particular dimensions we eliminate or preserve in this exercise are not idiosyncratic accidents. They reflect a philosophy that precedes them and which they serve to amplify and extend. In projecting language as a rule-governed manipulation of symbols, we all too easily dismiss the concerns of human meaning that make up the humanities, and indeed of any socially grounded understanding of human language and action. In projecting language back as the model for thought, we lose sight of the tacit embodied understanding that undergirds our intelligence. Through a broader understanding, we can recapture our view of these lost dimensions, and in the process better understand both ourselves and our machines.

Notes

I thank Gary Chapman, Brad Hartfield and especially Carol Winograd for insightful critical readings of early drafts. I am also grateful for my continuing conversation with Fernando Flores, in which my understanding has been generated.

1 As described by Russell (1952) in *A History of Western Philosophy*.
2 One large-scale and quite controversial example was the Massachusetts Institute of Technology/Club of Rome simulation of the world social and economic future (The Limits of Growth).
3 Remarks made by Anatol Holt at the Advanced Research Project's Agency Principal Investigator's Conference, Los Angeles, February 6–8, 1974 (unpublished manuscript).
4 For a historical account and analysis of the current debates, see H. Dreyfus and S. Dreyfus, *Making a Mind vs. Modeling the Brain* (1988). For a technical view, see Rumelhart and McClelland (1986a), *Parallel Distributed Processing*. Maturana and Varela, in *The Tree of Knowledge* (1987), offer a broad philosophy of cognition on this base.