3-1990

A Postmodern Critique of Artificial Intelligence

John N. Hooker
Carnegie Mellon University, john@hooker.tepper.cmu.edu

A Postmodern Critique of Artificial Intelligence

J. N. Hooker

Revised March 1990

Working Paper 56-88-89

Graduate School of Industrial Administration
Carnegie Mellon University
Pittsburgh, PA 15213
Bitnet jh38@andrew.cmu.edu

Abstract

I argue that the "postmodern" understanding of language that has developed over the last few decades in Anglo-American philosophy provides the basis for a useful critique of artificial intelligence. This postmodern view corrects an error in the traditional Western conception of language that has led many researchers in AI and cognitive science into taking a rule-based or information-processing approach. Wittgenstein's view that language does not receive its meaning through definition, and Quine's view that neither words nor sentences but only discourse as a whole is the proper unit of meaning, argue against an attempt to formulate rules for understanding language, which is an essential part of "strong" AI. AI researchers are already beginning to correct this mistake, but an understanding of its true extent and depth can lead to the sort of radical rethinking that is necessary.

The field of artificial intelligence (AI) aims to reproduce human intelligence by artificial means. The AI community generally concedes that achieving this goal has been harder than anticipated and awaits significant breakthroughs. But it insists that there are no barriers in principle. The obstacles to artificial intelligence are technical in nature, and a vigorous application of human intelligence can eventually overcome them.

There are grounds to believe, however, that AI as customarily practiced rests on a fundamental mistake of principle. Hubert Dreyfus, one

of the loudest of AI's critics, makes such a claim in his book, What Computers Can't Do (1979), and elaborates it in subsequent articles (1984; Dreyfus and Dreyfus, 1984, 1986, 1988). His critique actually has two prongs. He castigates the leading figures of AI for overestimating the prospects of their field and consistently failing to deliver. His shrill polemic along these lines has generated controversy that can obscure the second, more general prong of his critique. It is that AI rests on the mistaken assumption that human intelligence can (roughly speaking) be understood at the information processing level. A philosopher trained in the Continental tradition, he draws on the work of Heidegger to identify the error and show how it surfaces in AI.

It is not my purpose to explain or defend Dreyfus' philosophical critique of AI. Rather I want to show that a related critique can be mounted on the "postmodern" understanding of language that has developed in the Anglo-American philosophical tradition, particularly in the work of Ludwig Wittgenstein and Willard Quine.

I do not claim that AI is impossible. On the contrary, AI's philosophical underpinnings, as I will identify them, have a complex and indirect relationship to its potential for technical success. But if I am right, AI practitioners have typically operated on a set of assumptions that are not only incorrect but uninformed by Western thought's self-criticism of the last few decades. My critique also applies to the science of cognitive psychology, to the extent that it presupposes that human intelligence can be understood on a purely information processing level.

In the five following sections I state my thesis, briefly paint the Western intellectual background that supplies AI with its present "modern" (as opposed to the "postmodern") worldview, explain why Wittgenstein and Quine think this worldview is mistaken, and then outline what I take to be the implications of this mistake for AI.

1. The Thesis.

I take no exception whatever to the view that one can, in principle, create intelligence artificially. For that matter, I can think of no reason in principle why one could not fabricate a flesh-and-blood human being in a laboratory, granting that the technical problem would be formidable. Furthermore, I have no objection to the idea that an intelligent digital computer can in principle be built. If nothing else,

one could use the computer to simulate the chemical and biological processes that occur inside a human body. Perhaps a robot would measure the intensity of physical disturbances (sound, light, pressure) around its surface. Transducers would convert the readings to binary signals and convey them to a computer that would, say, solve a massive system of differential equations so as to simulate the reaction of a human body to these stimuli. Again the technical problem is mind-boggling, and for all I know it may require computation speeds inconsistent with physical law. In any case I do not wish to make a case against the possibility of such a simulation. I elaborate further on this point elsewhere (Hooker, 1989).

My thesis is a conceptual one, that intelligence cannot be understood as information processing. By information processing I mean the activity of a device that embodies a formal system. The device reads in discrete pieces of information, perhaps in binary form (zeros and ones), manipulates them according to a finite list of precisely defined rules, and generates output. A key property of such a device is that it is totally characterized by the formal system it embodies. It may be a contraption of rods and wheels--such as Charles Babbage's original "analytical engine" (Morrison & Morrison, 1961)--or electronic circuitry. The input data may be indicated by positions of a cogwheel or by electric charges in a capacitor. All that matters is the set of governing rules (programming instructions). Babbage's analytical engine would be no less intelligent than a modern supercomputer if it implemented the same formal system (and if we overlook speed differences).

The distinction of AI based on information processing from AI in general is, in its essentials, hardly new. Not only does Dreyfus make it, but so does John Haugeland (1985), who refers to information-processing-based AI as "Good Old Fashioned Artificial Intelligence." John Searle (1980) uses instead the term "strong AI.".

An example may clarify the precise distinction I have in mind. Suppose I want to build a chess-playing computer. The traditional AI approach is to try to build an information processing machine. I would encode the position of the chess pieces as, say, binary data, and let the machine manipulate the data according to a finite and precisely defined set of rules so as to generate output data that indicate the next move. There is another approach, however, that is equally suitable for a digital computer. I can, for instance, try to build a computer that simulates the chemical and biological processes inside a human chess player, as described above. This may seem to be just another embodiment of a formal system and therefore an information processing machine, since it, too, processes

binary data according to a finite set of rules. But there is an important difference in how the input is interpreted. In the information processing machine, the input data encode information according to certain prespecified rules. The data can be represented by any type of physical state, mechanical or electrical, since they are nothing more than tokens to be manipulated in a formal system. In the simulation machine, the input is itself a particular physical phenomenon, generated by sensors. Rather than being part of a language that encodes the position of the chess pieces, the input is a causal effect of the position of the chess pieces themselves. To extract information from the data, one need not "understand" the code as he would understand a language. He can simply trace the chain of causes and effects back to the chess pieces.

In fact, the strategy of my argument has to do with how an intelligent device interprets the input data. In a simulation model there is no problem in principle, since the computer need not understand the input data as one would understand a language. It is true that certain sequences of zeros and ones are likely to correspond to certain configurations of pieces, and one could formulate these correspondences as rules for "interpreting" the data. But the data do not obtain their meaning from such a rulebook. They indicate chess piece positions by virtue of their role in a chain of causes and effects, not by virtue of meanings assigned to them a priori.

In the information processing model, however, the input data represent the information directly. The data would be useless and meaningless were it not for prearranged conventions as to how they encode the positions of the chess pieces. Therefore, a digital intelligence must embody a finite list of rules for discerning their meaning. But it is, I believe, impossible in principle to do this for human language. The popular belief to the contrary is based on a mistaken understanding of what meaning is -- a misunderstanding that has been debunked by Wittgenstein, Quine and others.

I realize that the meaning and importance of my thesis may be foggy at this point. They should become clearer, however, as I present my argument and its implications. It is typical of philosophical claims that one must live with them for a while -- so as to see how they are defended and what implications they have -- before their meaning begins to sink in (a phenomenon, incidentally, that is predicted by the theory of language I present in this paper). The reader should therefore be patient and allow his understanding to grow in time.

2. Historical Background.

In the 1930's and 40's Wittgenstein worked out a radical critique of the traditional Western understanding of how language obtains meaning. Quine, influenced by logical positivists (e.g., Otto Neurath) and the American pragmatic tradition (Charles Sanders Peirce, William James, John Dewey), propounded a related critique in the 1950's and 60's based on his philosophy of science. To understand the significance of these critiques it is necessary to understand the view of language they received from the past.

To a large extent Western thought is the product of two lines of development, rationalism and empiricism, and their interaction. Since empiricism is now ascendant, it is difficult for us to grasp the older, rationalistic view that was the gift of the ancient Greeks and remains a silent influence on our thought. Perhaps it is easiest to begin with Pythagoras. That he is the namesake of the famous theorem, perhaps the first ever proved, is no accident. The Greeks believed that the world is reasonable and explicable, that one can understand why everything is the way it is, and that the natural order is therefore beautiful and harmonious; their very word for the universe, cosmos, is cognate with the Greek word for beauty. Pythagoras reflected this remarkable faith in the intelligibility of the world in his belief that it is fundamentally mathematical. A mathematical world consists of ideal objects like triangles and circles whose every property can be known by thought alone. The length of a right triangle's hypotenuse relative to that of its legs need not be observed; it can be deduced. Pythagoras believed that the world, properly understood, consists of mathematical objects and is therefore transparent to a sufficiently astute intellect. Therefore, the faculty that is suited to perceive the world is one's mind or intellect (psyche) and not his sense organs. He believed that this will become clear to us when the psyche (often translated "soul") is released at the time of death and is unencumbered by its bodily "tomb." Pythagoras' views sound exotic and in fact have oriental roots, but they deeply permeate the Western mind. They are evident in our penchant for mathematical models in science and our notion of the immortal soul, borrowed from the Greeks by Christianity.

The rationalist worldview received its modern expression in the work of Renw Descartes, Baruch Spinoza and Gottfried Wilhelm Freiherr von Leibniz. Descartes believed that intellection rather than sensation is the only sure route to knowledge. Spinoza undertook to exposit his entire

philosophy in the form of a mathematical text, complete with theorems, proofs and corollaries. Leibniz believed that that since the world is an intelligible realm, no less so than mathematics, one can in principle formulate a precise language (characteristica universitas) within which one can deduce all truths from self-evident axioms. Furthermore, since calculation in mathematics is nothing more than automated deduction, one can in principle obtain all truths by calculation. Leibniz lacked the technical wherewithal to formulate his calculus of reasoning (calculus ratiocanator), but George Boole, now regarded as one of the fathers of computer science, did not. He showed how deduction in propositional logic can be rendered as calculation in a kind of binary arithmetic remarkably similar to that employed by computers a century later (Boole, 1951, undated). But Boolean calculation is symbolic, not arithmetical, and since symbolic calculation is one of the key concepts of AI, Boole can be regarded as a precursor of AI as well as computer science. My point is that Boole's ideas were a direct result of the rationalist vision that received its purest expression in the work of Leibniz.

The modern empiricist worldview was perhaps first clearly articulated by Francis Bacon (1960), but for us it is more instructive to examine its implications for the philosophy of language already evident in the writings of John Locke (1967). Locke believed that the mind at birth is a blank tablet that over the years becomes impressed by "ideas," or bits of sensation. He concluded that knowledge is of two kinds: empirical knowledge, which one acquires by collecting sense impressions or "ideas," and rational knowledge, which one acquires by comparing ideas. It is a short step to a philosophy of language whereby words obtain meaning in two ways: by standing directly for sensations, as the word "red" might stand for a red sense-datum, and by being defined as a logical combination of other words. Words like "red" are said to be "ostensively defined," whereas words of the second kind are analytically defined.

Ironically, this modern view of language was carried to its logical extreme by a young Wittgenstein, who later repudiated it along with an entire philosophical tradition. Influenced by Bertrand Russell and other logicians, Wittgenstein says in his early Tractatus Logico-Philosophicus (1961) that a sentence has meaning because it pictures or mirrors a possible state of affairs: it is ultimately analyzable as ostensively defined words that bear a certain logical relation to each other. Thus language can in principle be made precise because it can be analyzed in terms of well-defined logical relations between primitive terms. For the early Wittgenstein the only fundamental difference between mathematics and discourse in general is that in mathematics the primitive terms may be left

undefined, whereas in general discourse they are defined ostensively. Therefore, reasoning is in principle no less susceptible to calculation in general discourse than it is in mathematics. Empiricism, as well as rationalism, leads to the view that meaning can be precisely defined and reasoning therefore carried out automatically.

It is natural, then, that one steeped in the Western worldview, shaped by the intertwining of rationalist and empiricist ideas, would take it for granted that the meaning of a stream of binary data can be precisely determined according to a finite set of rules.

3. Wittgenstein's Critique.

After writing the Tractatus Wittgenstein retired from doing philosophy, since he believed he had said all that could be said about the field. But soon a tiny crack appeared in his edifice. He became nagged by the "color exclusion problem," which asks: why is it a contradiction in terms to say that an object cannot be both red and blue at the same spot at the same time? If "red" and "blue" are defined simply by ostension, then neither excludes the other by definition. But if one defines "red" to mean, in part, "red and not blue," then the "red" being defined must have a different sense than the "red" in the definition, else the definition is circular. But the "red" in the definition surely excludes blue no less than the "red" being defined, and an infinite regress of definitions develops. In short, neither analysis nor ostension seems to account for what "red" means.

Peter Hacker (1972) points out that at about this time Wittgenstein attended a lecture by the famous intuitionistic mathematician L. E. J. Brouwer. Brouwer said roughly that we use a mathematical theory because it is useful in explaining the world, even if we cannot say what its primitive terms refer to. It is difficult, for instance, to say what imaginary numbers refer to, even though they occur in a very useful algebra. Yet the algebra as a whole is meaningful because of the role it plays in science, and imaginary numbers acquire meaning from the role they play in the algebra.

Brouwer's lecture may have been instrumental in widening the crack in Wittgenstein's "picture" theory of language, since mathematics already contains a clue to the revolutionary theory he would later propound in his Philosophical Investigations (1958). Russell, in formulating his notion of "contextual definition," had already recognized that a primitive term in

mathematics can acquire some degree of meaning simply from the way it is juxtaposed with other terms. An imaginary number, for instance, derives meaning from the fact that its square is a certain real number. Although we cannot say what an imaginary number "is," we know how the term "imaginary number" interacts with "square" and "real number" in a mathematical system, and this alone permits us to know, after a fashion, what the term means. Wittgenstein may have remarked to himself: perhaps there is nothing more to knowing what a mathematical term means than knowing how to use it -- not only when working within the mathematical system, but when applying the system to the world as well. Furthermore, perhaps something similar is true for language in general.

Wittgenstein eventually developed the view that the meaning of an utterance consists entirely in the role it plays in linguistic behavior, or as he put it, in a language game. For games offer another clue to how words obtain meaning. How could one define "strike out" or "foul ball" except in the context of a baseball game? We cannot define "strike out" by saying it refers to a particular action, since one who thrice swings a bat at a ball and misses, in a country where no one has ever heard of baseball, does not strike out. He must be playing baseball before he can strike out, because "strike out" derives it meaning from its role in a baseball game. It is the same for language in general, says Wittgenstein. Even the phrase "swings a bat" is not defined by its reference to a certain kind of action, but by its role in a broader language game.

The issue is clearest for ostensive definition. Suppose I try to define "red" for a small child by pointing to a red ball. What will the child do? He will look at my finger. Even if he looks at the ball, he must discern whether I mean the ball, its surface, the paint on its surface, its shape, or (perchance) its color, and even then whether I mean its precise hue or some class of similar hues. Actually the act of ostensive "definition" is itself part of a language game one must know how to play before he knows what is meant by the "definition."

It is tempting to suppose that, since a word's meaning is given by its use, one can define it analytically by describing its use. One problem with this supposition, but not ultimately the most serious problem, is that the description may be very long. The baseball analogy is misleading in that it suggests that someone has laid down a manageable list of well-defined rules for language. Grammarians and dictionarians may try to impose rules on linguistic behavior but end up describing the way some people in fact talk. They may distinguish certain ways of talking as "standard" or "accepted," but this does not show that talk acquires its

meaning from rules of accepted speech promulgated in dictionaries and stylebooks. Rather, the creation of dictionaries and stylebooks is itself a type of linguistic behavior, and it helps to determine the meaning of words by being part of the overall language game, not by laying down rules for the game. Thus a language game is a practice constituted by everyone's disposition to speak in a certain way, not by a set of imposed rules. The meaning of a word depends ultimately on the constantly-changing biology and experiences of speakers, since it is these that determine their dispositions to speak the way they do.

Therefore, to define a word by describing its role in a language game, one must describe every speaker's linguistic dispositions, a seemingly hopeless task. But the mere difficulty of this sort of definition does not show that it cannot in principle account for how words get meaning. There is a more basic problem. If it is to account for how words obtain meaning, it must do the same for the words in the definition, and for the words in their definitions, and so on. Eventually some words must ultimately be defined in terms of themselves, or we have an infinite regress, and nothing is defined.

The task of writing precise rules for understanding language, then, incurs both a theoretical and a practical difficulty. The theoretical problem is that if such rules could precisely capture the meaning of language, they would constitute a set of definitions. But language does not and cannot receive meaning through definitions. Its meaning is inherent in the shifting and complex patterns of linguistic behavior that people actually exhibit. The practical problem is that even if one were content to approximate the meaning of language by trying to write rules for appropriate linguistic behavior, he undertakes a herculean task. I cannot say outright that it is impossible, and I will consider this issue presently. But it is a mistake to suppose that it is possible because words acquire meaning through definitions that can be formulated as rules for understanding language.

4. Quine's Critique.

Quine's project in analyzing language is primarily to correct the verificationist theory meaning handed down by logical positivists and others. The theory makes a distinction, similar to Locke's, between analytic and synthetic statements. An analytic statement, such as, "Bachelors are unmarried," is true solely by virtue of the meanings of its

words (Locke would say by virtue of comparing ideas). Mathematics, for instance, can prove its propositions with a degree of certainty rare in this life, because they are all analytic. But there is a catch: analytic propositions say nothing about the world but only, so to speak, define their own terms. A synthetic proposition, such as, "Bachelors are lonely," makes a substantive claim about the world. On the verificationist view, its meaning is given by the experiences that would confirm or refute it. When a physicist says that the speed of light is the same in all inertial frames of reference, he means, on this view, that if one conducts an experiment like that of Michaelson and Morley (as well as an indefinite number of others), certain results will ensue. But if a theologian says that the godhead is omnipotent and omniscient, it is unclear what experiences could conceivably confirm or refute this claim. It is therefore unclear, on the verificationist theory, that it has any meaning at all. In fact the verificationist criterion traces to Immanuel Kant (1929), who used it to reject wooly metaphysics with no grounding in experience.

Ironically, one can detect the seeds of Wittgenstein's later view in the verificationist theory. It says in effect that the meaning of a statement has to do with the set of situations in which one would be disposed to assent to it (i.e., in which it is "confirmed") or dissent from it. The meaning therefore derives from its role in the "language game" of confirmation and refutation, which reaches its greatest refinement in science. Wittgenstein's insight is that there are a great variety of language games that can equally well endow language with meaning. Verificationists eventually softened their view by accepting the teaching of "ordinary language" philosophers that meaningful utterances, such as questions or promises, may be neither true nor false (and therefore unsusceptible to confirmation or refutation). But they insisted that any "cognitive content" must pass the verificationist test, even if this means that ethical and religious discourse is either noncognitive or well-nigh meaningless.

Quine's theory of language is in one way narrower than Wittgenstein's, since it continues to ground meaning in a language game of refutation and confirmation. But in another way it is broader, because it explicitly recognizes that only an entire discourse, and not the words and sentences in it, can be properly said to have meaning. Perhaps disciplined by his work in mathematical logic, Quine also favors us with a very detailed and closely-reasoned exposition of his views.

Quine begins by repudiating the analytic/synthetic distinction, in his

essay, "Two Dogmas of Empiricism" (1961). He argues that no statement is analytic, since we might deny even such canons of logic as the law of the excluded middle (i.e., "either A or not-A") if science found it convenient to do so. Rather, statements to which we assent arrange themselves in a "web of belief" that meets experience on its periphery (Quine, 1970). Statements around the edge are closely tied to observations and can be easily denied if experience warrants. An example might be, "The population of the United States is under 230 million." Statements near the center, such as the law of the excluded middle, are more dearly held and are rejected only if the web is radically restructured.

One might counter that if we deny the law of the excluded middle, we in effect alter the meaning of "or" and "not." Quine wholeheartedly agrees, for he asserts that the meaning of statements changes when our dispositions to assert and deny them change. Every time we revise our opinions, we alter the meaning of the very words we use to state those opinions. A sentence denied has a meaning different from the same sentence asserted.

This is less paradoxical when we realize that, on Quine's view, the unit of meaning is, strictly speaking, neither a word nor a sentence, but our entire body of opinions about the world. The reason is no one statement need be rejected in the face of recalcitrant experience if adjustments are made elsewhere in the theory. Johannes Kepler, for instance, could continue to maintain that planetary motion is circular, despite observation to the contrary, by allowing that it can be the compounded circular motion of cycloids. It is therefore not individual statements that are tied to experience (and thus to meaning), but our entire theory of the world. The meaning of this theory consists in our dispositions to retain or revise it in the face of changing experience. Thus neither a word nor a sentence can receive meaning through definition, because it has no meaning in isolation. Moreover it makes no sense to try to define the meaning of an entire worldview, since the definition would need to be cast in terms that are meaningful only within that worldview.

In Word and Object (1960) Quine further refines his view by exploring what it means to translate from one language to a radically different language. If an anthropologist wants to translate an utterance in an exotic language, he can begin by ascertaining its "stimulus meaning," which is the set of situations in which the natives are disposed to assent to it. Even stimulus meaning presents a problem, since as we noted earlier, these dispositions are determined by the changing biology and experiences of the speakers rather than by fixed rules. But the problem goes deeper, because

stimulus meaning does not uniquely determine translation. If a native says "gavagai" while pointing to a rabbit, does he mean "rabbitlike behavior," "rabbitlike appearance," "rabbit stage," "rabbithood," "four-legged while-tailed mammal," or "rabbit"? Quine argues carefully that translation requires "analytical hypotheses" that are underdetermined by linguistic behavior. It follows that two different and incompatible translations can be equally correct, since they may rest on incompatible analytical hypotheses. If sentences had well-defined meanings, this could not occur.

Quine's views on translation are related to Thomas Kuhn's famous argument that science undergoes revolutions or "paradigm shifts" in such a way that different paradigms are somehow "incommensurable." Paradigms are like different languages between which determinate translation is impossible. His views also relate to the recently popular notion that one cannot "step out of" the worldview of his historical epoch to evaluate it sub specie aeternitatis, a notion that has been advanced in various forms by Hans Georg Gadamer (1975), Jtrgen Habermas (1975), Richard Rorty (1979), Richard Bernstein (1983), and others. Unfortunately some philosophers infer that one cannot judge one worldview to be more rational than another, but I believe that one can accept Quine's insights on language and meaning without committing himself to this sort of relativism. The important lesson here is that, on Quine's view as well as Wittgenstein's, words and sentences cannot receive their meaning through definition. Whereas Wittgenstein emphasizes that meaning resides in linguistic behavior, Quine emphasizes that it properly belongs to an entire body of opinion, not words and sentences individually.

5. Consequences for Artificial Intelligence.

I said earlier that an information processing model of intelligence is beset with a theoretical and a practical problem. The theoretical problem has to do with what might lead one to believe that formal rules can be laid down for interpreting language. If one believes this on the ground that words or sentences receive their meaning through definition, so that one need only codify their definition as rules for interpreting language, then his belief rests on a mistake.

AI enthusiasts are of course free to deny that their optimism has rested on the modernist view of language I have criticized. The history of the field suggests that they have in fact become aware of some difficulties in the modernist view but have not, perhaps until recently, begun to to

sense the need for radical overhaul.

Some have dealt with the lack of precise definitions in natural language by generating an enormous literature on fuzzy logic and fuzzy set theory (beginning with Zadeh, 1965 and 1975). But the problem here is not merely that terms can only be roughly defined, and this vagueness must somehow be captured in logic. The problem is that language does not acquire meaning through definition in the first place.

The notion of "frames" and similar concepts (Minsky, 1975; Schank & Abelson, 1977) reflect an awareness that the meaning of language somehow depends on the particular context in which it is used. But it is not enough to try to write interpretation rules that are specific to a "frame" or context. One must realize that language derives its meaning entirely from the behavior of language users in such "frames," or (for Wittgenstein) "language games," and not from rules of usage laid down for such games.

Workers in natural language translation have gradually recognized that one cannot understand a statement without bringing to bear a good deal of knowledge about the world (Winograd, 1972; Weizenbaum, 1976). But it is not enough merely to conclude that translation rules must be supplemented with a knowledge base; we must, with Quine, take the radical view that meaning (or at least "cognitive" meaning) is determined by the contents of our knowledge base and our dispositions to revise it in the light of new experience.

The recent shift from rule-based AI to learning in neural networks (Rumelhart & McClelland, 1987 and 1988) may indicate that AI has weaned itself from the traditional view, but it may only reflect frustration with the failure the rule-based approach, or perhaps a new fascination with the analogy with the human brain.

The defender of rule-based AI can, however, concede that his field has perhaps uncritically accepted the modernist view of language, and still insist that its belief in the possibility of rule-based AI may be correct even if it has historically been held for the wrong reasons. He can concede that the AI community equated intelligent machinery with programmed machinery as long as it did because it was under the spell of a Lockean or Cartesian worldview, and point out that the impossibility of programming a natural language interpreter hardly follows. He can grant that an implicit faith in the possibility of something like Leibniz's characteristica universitas may have led some researchers to invest enormous energy in implementing formal systems, and maintain that this very faith may impel

them to achievements they would otherwise judge impossible. He can point out that Kepler might never have discovered the elliptical nature of planetary orbits had he not been motivated by a religious belief in the perfection of the heavens.

The AI defender can even drop his defense of a rule-based approach and turn the tables on me. He can point out that if the meaning of language consists in how it is used, then a computer can understand language in the fullest sense of the word simply by learning how to use it. Since machine learning is, after all, one of the most vigorous areas of AI, the theory of meaning I propound can only counsel optimism for the prospects of AI.

I do not dispute these points. I said in the beginning that I take no issue with the idea that intelligence, and with it language comprehension, can be created artificially. But I am more reserved about an attempt to embody intelligence in a formal system, and this is where I encounter the practical, as opposed to the theoretical, problem of an information-processing approach. I cannot be certain that a clever researcher laboring under the Leibnizian view of language cannot, by dint of hard labor, put together a rule base that will approximate human intelligence. I grant readily that nothing about my own view of language and meaning rules this out a priori. My view does, however, suggest that a rule-based approach is not the natural one to try. The intricacy of the language games in which language acquires meaning suggests moreover that an adequate rule base would be incredibly complex. In any case I would not want the majority of the AI community to feel obliged to pursue an information-processing approach because it feels that it is the natural approach to take.

It is much better, I believe, to seize the insight that learning language is tantamount to learning language behavior, not to learning language rules. This implies a focus on machine learning, but we must be wary, because a good deal of work in machine learning is occupied with extracting rules that describe the sort of behavior we want to teach the machine, and using these rules to program the machine. In other words, it is occupied with induction, conceived as the process of inferring general rules from specific instances (Holland et al., 1986; Valiant, 1984). This again lapses into an information-processing model. A careful study of induction can be beneficial for expert systems and the like, but I believe it would be a mistake to focus primarily on induction because one believes it is the natural approach. The natural approach is to look for any learning mechanism, neural networks or otherwise, that can approximate human language behavior.

In fact the view of language I propose eases, in an interesting way, the task of building a machine that understands language. Let us suppose that intelligently talking machines become so prevalent that most people have occasion to converse with one nearly every day. Let us also suppose that their command of language is, as we might expect, imperfect. After all, no matter how many language lessons a machine may have been given, situations to which no lessons apply will arise. The machine must somehow interpolate (and extrapolate) his lessons to come up with an appropriate response. Human beings do this readily; with relatively little exposure to language, we find ourselves at no loss for words in an incredible variety of situations, ranging from the idle chatter of a party to a discussion of theoretical physics in a seminar. Moreover, we generally find others quite capable of understanding the utterances we come up with in novel situations. It is indeed as though human beings, or at least human beings within a given culture, are built to interpolate in similar ways, whether or not it be quite the way suggested by Noam Chomsky's work.

Machines, however, are likely to be wired differently than humans and to interpolate differently. They would be like persons from a radically different culture, or even alien beings from a different planet, trying to learn our language. On the traditional view of language, their departure from our speech habits would consign them forever to substandard usage. But on the view I advocate, they become part of our language games and therefore help to determine the meaning of language. As the machines become integrated into our conversations, as they adjust to us and we to them, new language games evolve, and language takes on new meanings. Eventually the distinction between well-spoken humans and inarticulate machines crumbles, as both develop a new, common patois. This phenomenon is prevalent in history when cultures mix, in such mild cases as American English or Canadian French, or in such more extreme cases as Pidgin English, the Cajun dialect of Louisiana, or the Creole dialect of Haiti. It is therefore not necessary to build machines that precisely mirror present human speech patterns in order to build machines that will one day be as articulate as humans.

Still, machines are more likely to resemble Martian invaders than Acadian migrants. They could bring an unprecedentedly radical shift in our language, for good or for evil. If the machines are, let us say, simple-minded, if they are prone to blur distinctions and overlook subtleties, our common language acould be debased. But if the machines are complex devices with a sophistication of their own, our new language may become all the richer for it. Since language profoundly influences

thought, machines could dull our faculties or lead to new and perceptive ways of thinking.

References.

Bacon, F. (1960). The Nova Organon and Related Writings, ed. by F. H. Anderson. New York: Liberal Arts Press.

Bernstein, R. (1983). Beyond Objectivism and Relativism: Science, Hermeneutics and Praxis. Philadelphia: Univ. of Pennsylvania Press.

Boole, G. (1951). The Mathematical Analysis of Logic. Oxford: Basil Blackwell.

Boole, G. (undated). An Investigation of the Laws of Thought. New York: Dover.

Dreyfus, H. E. (1979). What Computers Can't Do: The Limits of Artificial Intelligence, revised ed. Harper, New York.

Dreyfus, H. E. (1984). What expert systems can't do, Raritan: A Quarterly Review, 3, 22-36.

Dreyfus, H. E., & Dreyfus, S. E. (1984). Mindless machines: computers don't think like experts, and never will, The Sciences (New York Academy of Sciences, Nov./Dec.) 18-22.

Dreyfus, H. E., and S. E. Dreyfus (1986). Why computers may never think like people, Technology Review (January) 43-61.

Dreyfus, H. E., and S. E. Dreyfus (1988). Making a mind vs. modeling the brain: artificial intelligence back at a branchpoint, Daedalus, 117, 15-43, reprinted in S. R. Graubard (Ed.), The Artificial Intelligence Debate: False Starts, Real Foundations. Cambridge, MA: MIT Press, 1988.

Gadamer, H.-G. (1975). Truth and Method, transl. and ed. by G. Barden and J. Cumming, 4th ed. New York: Seabury Press.

Habermas, J. (1975). Legitimation Crisis, transl. T. McCarthy. Boston: Beacon Press.

Hacker, P. M. S. (1972). Insight and Illusion: Wittgenstein on Philosophy and the Metaphysics of Experience. London: Oxford University Press.

Haugeland, J. (1987). Artificial Intelligence: The Very Idea. Cambridge, MA: MIT Press.

Holland, J., K. J. Holyoak, R. E. Nisbett, and P. R. Thagard (1986). Induction: Processes of Inference, Learning and Discovery. Cambridge, MA: MIT Press.

Hooker, J. N. (1989). Can I be transplanted into a computer?, Working paper 45-88-89, Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, PA 15213.

Kant, I. (1929). Critique of Pure Reason, transl. N. Kemp Smith. New York: St. Martins Press.

Locke, J. (1967). An Essay Concerning Human Understanding, 2 vols., ed. by J. W. Yolton. New York: Dutton.

McCulloch, W. S., and W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics, 5, 115-133.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), The Psychology of Computer Vision. New York: McGraw-Hill.

Morrison, P. & Morrison, E. (1961). Charles Babbage and His Calculating Engines. New York: Dover.

Quine, W. V. (1961). Two dogmas of empiricism. In From a Logical Point of View: Nine Logico-Philosophical Essays, 2nd ed. New York: Harper and Row.

Quine, W. V. (1960). Word and Object. Cambridge, MA: MIT Press.

Quine, W. V., & Ullian, J. S. (1970). The Web of Belief. New York: Random House.

Rorty, R. (1979). Philosophy and the Mirror of Nature. Princeton Univ. Press.

Rumelhart, D. E., & McClelland, J. L., Eds. (1987 and 1988). Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 3 volumes. Cambridge, MA: MIT Press.

Schank, R. C., & Abelson, R. P. (1977). Scripts, Plans, Goals and Understanding. Hillsdale, NJ: Lawrence Erlbaum.

Searle, J. R. (1980). Minds, brains and programs, The Behavioral and Brain Sciences, 3, 63-73, reprinted in J. Haugeland (Ed.), Mind Design: Philosophy, Psychology, Artificial Intelligence. Cambridge, MA: MIT Press, 1981.

Valiant, L. G. (1984). A theory of the learnable, Communications of the ACM, 27, 1134-1142.

Weizenbaum, J. (1976). Computer Power and Human Reason: From Judgment to Calculation. San Francisco: Freeman.

Winograd, T. (1972). Understanding Natural Language. New York: Academic Press.

Wittgenstein, L. (1958). Philosophical Investigations, 3rd ed., English transl. by G. E. M. Anscombe parallel to original German. New York: Macmillan.

Wittgenstein, L. (1961). Tractatus Logico-Philosophicus, transl. by D. F. Pears and B. F. McGuinness. London: Routledge and Kegan Paul.

Zadeh, L. A. (1965). Fuzzy sets, Information and Control, 8, 338-353.

Zadeh, L. A. (1975). Fuzzy logic and approximate reasoning, Synthese, 30, 407-428.