Second Edition

ENCYCLOPEDIA OF ARTIFICIAL INTELLIGENCE

Volume 2
M-Z

Awarded

American Library Association's

Outstanding Reference Source

Association of American Publishers Award

Best New Professional and Scholarly Publication

Stuart C. Shapiro
Editor-in-Chief

Robert Van Gulick, PHILOSOPHICAL QUESTIONS, 1137-1147

From Shapiro, Stuart C., Editor-in-Chief, Encyclopedia of Artificial Intelligence, 2nd Ed., John Wiley & Sons, Inc., NY, 1992.

PHILOSOPHICAL QUESTIONS

The interests of philosophers and workers in AI intersect and overlap in many ways. Some philosophers have tried to use the resources of AI to shed new light on long-standing philosophical problems, and others have been vocal critics of the philosophical claims made by AI researchers. There has also been convergence. In carrying out specific projects, AI researchers have frequently been led into areas traditionally discussed and investigated by philosophers, providing new opportunities for collaborative exploration.

The philosophical impact of AI has been greatest on the philosophy of mind. AI has suggested new answers to long-standing questions about the nature of mind, led to the reformulation of traditional problems, and given birth to new controversies of its own. The mind–body problem, the mechanism of free-will debate, and disputes regarding the nature of understanding, intentionality, and intelligence have all been transformed in substantial ways by the advent of AI.

There are also important connections between AI and other areas of philosophy as diverse as the philosophy of science, the philosophy of language, metaphysics, and epistemology (qv). (The relevance of logic is almost too obvious to mention.) AI issues in these other areas have not generated the sort of emotion-laden controversy associated with issues in the philosophy of mind, but in the opinion of some philosophers they may turn out to be of greater importance in the long run (Glymour, 1985).

PHILOSOPHY OF MIND

At least since the seventeenth century and the rise of modern mechanistic theories of the physical world, philosophers have debated the place of mental phenomena within the mechanistic scheme. The development of modern computers and AI programs gave new impetus to these debates. For the first time it seemed possible to actually construct machines that were both undeniably mechanistic in their operations and possessed characteristics and abilities uniquely associated with minds. The possibility of computing machines able to play chess, prove theorems, and perhaps engage in conversation gave the mind–mechanism dispute a timely urgency.

The Turing Test

The mechanism question has been posed in a variety of related but independent forms, including "Can machines think?" "Do computers have minds?" "Is artificial intelligence really intelligence?" In his seminal 1950 paper "Computing Machinery and Intelligence," Turing (1950) considered the first of these questions and found it too

meaningless to deserve discussion. He proposed, therefore, to replace it with another question that was more precise and decidable but that captured the essential issues raised by the more familiar popular formulation.

Turing proposed an imitation game to be played by a human interrogator and two unseen participants X and Y, one a human and the other a machine (see TURING TEST). The interrogator is able to address any questions to X and Y and they are to respond by typewritten messages. Both the machine and the unseen human have the same objective in the game: each tries to convince the interrogator that it is the human respondent. Turing's replacement question was posed in terms of this game, "Are there imaginable digital computers which would do well in the imitation game?" He answered the question in the affirmative and predicted that within 50 years digital computers would be able to play the imitation game well enough that an interrogator would have no more than a 70% success rate in identification of the machine respondent after a five-minute conversation.

Objections to Test's Criteria. The truth value of Turing's prediction about machine success at the imitation game is obviously an empirical and not a philosophical matter. However, philosophers have challenged its adequacy as a substitute for the original "Can machines think?" query. Criticism has focused on the behavioral evidence that forms the basis of the text. The sample of behavior may seem too limited; a 5-minute conversation appears an inadequate basis for a judgment of mentality. Such objections are not really very serious. Although the conversation is brief, interrogators are not limited in their range of discussion topics. They may make inquiries about poetry, sports, music, or cuisine, and in all these areas the machine must produce plausible humanlike responses. The test could also be modified to allow for longer conversations without altering its rationale.

The behavioral evidence has been challenged in a more substantial way by philosophers who have argued that behavioral criteria alone cannot suffice for the applicability of mental predicates. They argue that having a mind is not merely a matter of exhibiting certain patterns of verbal or nonverbal behavior but also requires that the right sort of internal processes produce the relevant behavior. Sometimes the objection is raised as a denial that the machine does exhibit the same behavior as a human (Gunderson, 1964). The machine and the human respondent may produce the same end result, a given sequence of words on the printer's output, such as "I do not have much taste for Mexican cooking," but it does not follow that they are exhibiting the same behavior. In the human case, causing those words to be printed is the making of an assertion. It is a significant linguistic act produced as the result of a prior communicative intention. The machine respondent's printing of those same words would count as an assertion only if it were also produced by a similar communicative intention: Since interrogators have access only to the physically indistinguishable end products, they cannot determine which instances of seemingly communicative behavior are genuine assertions. Their inability to identify genuine assertions on the basis of limited available evidence in no way implies that the two respondents are both engaging in the same sorts of behaviors.

A General Antibehaviorist Objection to the Test. Another way of making the antibehaviorist point is by imagining a system that simulates human behavior well enough to pass the Turing test but does so as the result of internal processes that obviously involve no genuine thought or intelligence. One such example has been provided by Block (1981). Block imagines a device that consists primarily of an enormous memory that stores a very large but finite list of all English-language conversational exchanges up to a given length (with a further limit on the length of each utterance in the exchange). The list must be ordered in some way that allows rapid access. Confronted with an interrogator in a Turing test, the machine searches its memory to find a conversation whose first segment corresponds to the interrogator's initial question. It then prints the next utterance from that conversation. Following the interrogator's reply, it again searches for a conversation whose first three segments correspond to those in its present dialogue and prints the fourth utterance from that conversation. It continues this process up to the length of the test, which is no greater than the length of its stored conversations.

One may object that such a machine is wildly impractical and argue that no such machine could ever be built, but to do so would be to miss the philosophical point of the example. The machine is in practice an impossibility; indeed, the number of items it would have to store if the conversation length were increased might soon outstrip the number of elementary particles in the universe (Churchland and Churchland, 1981). As a thought experiment, the example is not intended as a practical suggestion for building conversation machines. It is intended to make a conceptual point about the notions of intelligence and having a mind by showing that they require more than the satisfaction of the sorts of behavioral criteria employed by the Turing test. Being intelligent or having a mind is also a matter of the internal processes that produce behavior. Such conceptual connections may be obscured by the fact that in normal human intercourse, judgments about another person's mental state are normally made on purely behavioral evidence. Someone who can paraphrase a story and answer a suitable range of questions about it is considered to have understood that story. But in such cases the other person's status as a rational thinking understanding agent is not in question, only that person's mastery of the particular story at hand. As philosophers have noted, the criteria that suffice in these specific cases cannot simply be extended to other cases where the basic issue of having a mind at all is in question (Scriven, 1953).

Some may object that the Block (1981) example shows only that the internal processes underlying genuine intelligence must satisfy real-time constraints. However, Block's example is intended to make a stronger point and is taken by many to have done so. The existence of such a machine operating in real time is at least a logical (if not

an engineering) possibility, and there is a strong intuition that such a device would have no genuine understanding of language.

Functionalism and AI

Philosophical dissatisfaction with the Turing test reflected a general rejection of behaviorism as a theory of mind. Few philosophers any longer believe that mental predicates can be defined or explicated in purely behavioral terms. However, the functionalist theory of mind (Block, 1980; Dennett, 1978) that dominates present philosophical thinking, preserves some elements of the earlier behaviorism, especially in its claim that many commonsense or folk-psychological mental concepts are to be explicated at least partially in terms of their behavior-causing roles. The functionalist program has been strongly influenced by analogies drawn from computer science and AI, both in its general outlook and in several of its specific applications to problems about the nature of mind.

Functionalism as a distinctive position was developed in the 1960s (Fodor, 1968; Putnam, 1967; Lewis, 1972) as an attempt to avoid the shortcomings of the two then most popular philosophical views of mind, behaviorism and physicalist identity theory, while retaining the strengths of each. Its central idea is that psychological states, such as desiring to be famous, believing that it will rain tomorrow, feeling a pain, or being angry, are type-individuated on the basis of their causal functional roles in mediating an organism's or system's interaction with its environment. Being in pain or having a belief that chalk is white is a matter of bearing appropriate causal relations to sensory inputs, behavioral outputs, and other internal states mediating the system's connections between perception and action. Items with radically different intrinsic natures can all count as instances of the same psychological kind as long as they play the same causal roles within their respective containing systems.

Functionalism thus rejects the Cartesian intuition that a mental state's psychological kind is fixed by its intrinsic, directly introspectible properties. For the functionalist it is the state's causal relations within the system that are relevant. Functionalism differs from behaviorism primarily in two respects (Block, 1978). First, it treats mental states as genuine causes, as actual internal states that play a role in the production of behavior. Many behaviorists regarded mental predicates simply as abbreviated ways of talking about behavioral patterns or regularities. Attributing to someone a desire to be wealthy was, for the behaviorist, only a way of describing the agent's behavior, not explaining the production of that behavior by reference to an inner cause. Functionalism's realism about inner causes also accounts for its second difference from behaviorism. Functional states are defined by their relations not only to input and output but also to one another. Functionalism can thus deal with the holism of the mental and the fact that mental states normally produce behavior only in joint operation. A desire for a cold beer will produce little behavior in the absence of suitable beliefs, and some mental states, such as a belief in a law of logic, will have functional roles that almost exclusively concern their influence on internal processes, such as patterns of inferential reasoning.

Functionalism and Physicalism. Functionalism departs from its other philosophic ancestor, type-identity theory, in its emphasis on function as opposed to structure. Type-identity theorists had proposed to identify mental kinds with specific physical kinds (normally neurophysiological kinds) empirically found to be correlated with their occurrence (Smart, 1959). The property of being in pain might, for example, be identified with the property of having one's C fibers firing. Functionalists, with the aid of insights drawn from computer science, have argued that even if as a matter of fact a given functional role is normally filled by a specific physical structure, other physical structures might in other contexts fill that same causal role (Putnam, 1967). Type-identity theorists have erred in identifying mental states with the narrow range of physical states that fill the relevant causal roles in human brains, thus unreasonably excluding organisms with different physiologies as well as robots and AI devices from the realm of the mental (Block, 1980). This functionalist critique of the identity theory, which is known as the multiple-realization argument, was directly inspired by computer and AI analogies. The philosophical presentations of the argument allude to the fact that the same algorithm may be carried out on a wide range of physically dissimilar devices (Putnam, 1967). The claim is also often explained by appeal to the software–hardware distinction, where again multiple realizations are possible and common.

Intentional Stance. AI has also had a strong influence in leading functionalists to distinguish among various levels of abstraction at which organized systems can be described and explained. Perhaps best known is Dennett's (1987) scheme, which is introduced in application to a chess-playing computer and includes three stances from which one may attempt to explain its behavior: the physical stance, the design stance, and the intentional stance. They involve descriptions and predictions based, respectively, on structure and physical causation, teleological function, and rational belief–desire explanation. They correspond in a rough way to what in AI might be called hardware, software, and knowledge-level descriptions. Dennett's account of the intentional stance should be of interest to AI researchers as well as philosophers since it separates the notion of having intentional states such as beliefs and desires from any metaphysical commitments about the system's underlying substance (spiritual, organic, or electronic) and provides a practical method for determining when descriptions of a system's (or subcomponent's) behavior involve implicit attributions of rationality and mentality. Although Dennett's intentional-systems theory has been subjected to much philosophical criticism (Stich, 1981), it is a clear improvement over the casual and unregimented use of intentional terminology in AI by which it is partly inspired.

Dennett's work also provides the best example of another major application of AI resources to the functionalist program in dealing with the problem of hidden theoretical homunculi (Dennett, 1975). Mentalistic psychology has often been faulted for implicitly relying on covert internal agents who account for regularities in external behavior by reproducing within a subpersonal component the cognitive abilities of the person that are supposedly being explained. The threat of vacuity or vicious infinite regress looms large when explanations of visual perception rely on a mind's eye to perceive an internal object or explanations of rational action refer to an inner decider who ranks alternative courses of action. Dennett has drawn directly on work in AI to resolve this centuries-old problem. The first part of his solution is to apply the AI strategy of decomposing a complex task or function into a set of organized subtasks and then subjecting those second-level tasks to the same sort of decomposition, repeating the process until the whole organized hierarchy comes to rest on interacting components whose behavior is straightforwardly mechanizable. Using the intentional system's method for keeping track of implicit attributions of rationality and mentality, the descriptions at each level become progressively less mental as one descends the hierarchy. Dennett (1975) describes the procedure as decomposing homunculi at each level into a committee of individually dumber homunculi at the next level down until the process terminates at an "army of idiots." Dennett's use of AI techniques thus answers philosophical criticism of homunculi by showing that their theoretical use need involve neither vacuity nor infinite regress.

Computational Theory of Mind

A third application of AI to functionalism involves the computational theory of mind (CTM). Functionalist philosophers have carried the analogy between AI programs and the organization of the mind one step further in suggesting not only that the mind decomposes into a hierarchical series of levels but also that the operations of the subcomponents at the underlying levels consist entirely in the computational manipulation of representational structures or formal symbols. The operations are computational in that they are governed by rules that can be completely and explicitly formulated in terms of the formal and syntactic properties of the representations (Fodor, 1980). As representations, such structures also have semantic content, but the underlying processors that manipulate them do so solely on the basis of their forms and syntax. The CTM is intended to resolve (or dissolve) the philosophical problem of explaining how intentional content can have a causal impact on the physical world. According to the CTM, content can have causal consequences only insofar as it is mirrored in formal structure. Differences in representational content that are not reflected in formal differences detectable by internal processors can have no impact on behavior. The CTM thus employs an entirely syntactic taxonomy of internal representations and individuates psychological states solely on the basis of the formal objects to which they are related. Since facts about the social, cultural, historical, or physical environment play no direct role in the determination of computational content, the CTM is said to be methodologically solipsistic in its approach to psychology (Fodor, 1980).

Anticomputationalist Critique. To those philosophers who accept the CTM, it represents perhaps the most important application of AI theory to the philosophy of mind. However, it has provoked other philosophers to the strongest and most widely discussed criticisms of recent work in AI. The outstanding critic in this regard has been Searle, who in his influential article "Minds, Brains and Programs" (1980) attempts to refute the claims made by Roger Schank and others that their script-based story-understanding programs literally "understand" the stories on which they comment and answer questions (see SCRIPTS; STORY ANALYSIS). Searle has a larger goal than merely challenging some perhaps exaggerated or premature claims about the level of present AI success. His aim is to refute the CTM and all work in AI that relies on it. His examples and arguments are meant to demonstrate that understanding (or having any other intentional state) can never be simply a matter of having the right sort of internal formal structure or being an instantiation of the right sort of computer program.

The central focus of Searle's argument is a thought experiment that has come to be known as the "Chinese room." Searle imagines himself locked in a room with three batches of Chinese writing and some sets of instructions for manipulating Chinese symbols. The rules are given in English, and Searle can carry them out although he does not understand Chinese since they specify the operations to be carried out purely in virtue of the shapes of the Chinese symbols and do not allude to their meanings. Searle carries out the instructions and passes back strings of Chinese symbols produced as the result of his operations. The reader is asked to imagine that the three batches of Chinese symbols correspond to what Schank and his colleagues would call a script, a story and questions, that the instructions correspond to a program, and that the strings of symbols Searle produces represent conversationally appropriate answers to the questions about the story. Searle's contention is that in such a case he would satisfy all the conditions on the basis of which a computer running Schank's program is said to understand Chinese, and yet it is intuitively obvious that in such a case he (Searle) would not understand a word of Chinese; thus, he infers that Schank's computer is equally lacking in understanding of Chinese. Although Searle's argument bears some similarities to earlier criticisms of the Turing test, it has relevance to a far wider class of theories. It is directed not only against behaviorist views, which equate understanding with performance, but also against the CTM and all those versions of functionalism and AI that attempt to account for the mental production of behavior purely in terms of formal structures and computational rules.

Replies to Anticomputationalist Argument. Searle's argument has provoked a great deal of vigorous criticism, but he has displayed considerable dialectical skill in re-

plying to his critics. It has been argued that although Searle alone would not understand Chinese, the ensemble of Searle plus instructions and batches of symbols does understand Chinese. In reply, Searle has offered a modified example in which he memorizes the batches of symbols and rules, although continuing to treat the symbols as mere uninterpreted shapes. In such a case every part of the ensemble would be internal to Searle, and yet intuitively he would not seem to understand Chinese. Some critics have conceded genuine understanding requires more than the formal ability to manipulate symbols since understanding a symbol's meaning requires the ability to relate the symbol to the nonsymbolic external world. Thus, genuine understanding would require a robot capable of perception and action as well as of merely "conversational" performance.

Such AI proposals are in keeping with the functionalist view that mental kinds are determined by the causal role that a state plays in regulating purposeful interaction with the environment. Searle has also varied his example to answer these robot proposals, albeit perhaps with less convincing success. He imagines himself in the robot's control room, where in addition to his earlier symbolic inputs, various formal symbols appear on a screen. Those unknown to Searle are perceptual inputs; to him they are just further uninterpreted shapes to be manipulated according to rules governing only formal operations. Unknown to him, his activity produces appropriate external responses by the robot. Searle argues that he still would not understand Chinese although the robot's behavior might seem to show an understanding of how Chinese symbols relate to real-world items. Many functionalists do not share Searle's intuitions about the robot case, especially if the entire organized robot, rather than merely its Searle "component," is considered as the potential understander of Chinese. Searle's intuitions seem to rest on the absence of any subjective or experiential elements in the robot's internal processes, but it is a theoretically open question whether such processes are essential for genuine understanding. Thus, the controversy surrounding Searle's argument turns in the end on a basic conflict of fundamental intuitions.

Semantics and the Problem of Original Intentionality. The problem raised by Searle is closely related to what Haugeland has called the problem of original intentionality (Haugeland, 1985). Some symbols, such as the words of a natural public language like English, may borrow or derive their meaning from other symbols, such as ideas or other inner mental symbols, through a process of association. But if a vicious and infinite regress is to be avoided, there must be some symbols that have their meaning in a nonderivative or primary sense. The problem of original meaning is that of explaining how such nonderivative symbols come to have the semantics or meanings that they do. Haugeland's own solution is a form of interpretational semantics; he requires that a way of interpreting the symbols of an automatic symbol manipulating system (whether a brain or a computer) be found so that its inputs and outputs make sense, a notion that he admittedly leaves less than precisely defined. However, anticomputationalists like Searle would say that such an interpretational theory can provide at best an account of what it is to treat something as if it had semantics or meaning, but not an account of what it is to have original meaning in the literal sense.

A wide variety of other theories have been offered to solve the problem of original semantics. Following Cummins (1989), these theories can be classified into five main categories which respectively account for semantics or representational content in terms of similarity (Locke, 1959), covariance (Fodor, 1987; Dretske, 1981), adaptational role (Millikan, 1984), functional role (Block, 1986), or interpretational semantics (Haugeland, 1985). Each position has supporters and merits along with critics and weaknesses, and as yet no consensus view has emerged. The problem remains one of the most active and exciting areas of current philosophical research.

Connectionism and the Philosophy of Mind. The advent of connectionism (qv) and the development of parallel distributed processing architectures for AI tasks has provoked considerable discussion. Some philosophers have appealed to connectionism to attack more traditional versions of the computational theory of mind (Churchland, 1986); in particular they have tried to diminish the plausibility of the language-of-thought hypothesis, which attempts to explain mental states in terms of computational relations to inner symbolic structures with both syntactic and semantic properties (Smolensky, 1988). The impressive initial successes of connectionist architectures in simulating at least some cognitive abilities has lent support to the claim that a computational explanation of mentality need involve no commitment to inner sentences or any type of computationally manipulated inner symbols.

The proponents of the language-of-thought hypothesis have defended their position vigorously and argued that connectionist models are incapable of accounting for certain features of our psychological makeup, such as the systematicity of mental representation, ie, the fact that one cannot, for example, acquire the ability to believe that Tom loves Mary without also acquiring the ability to believe that Mary loves Tom (Fodor and Pylyshyn, 1988). A third group of philosophers has suggested that the traditional language-of-thought (or rules and representations) view is in fact compatible with connectionism and the two models need not be regarded as competitive or mutually exclusive models of the mind (Bechtel, 1987). A further issue has concerned the compatibility of connectionism with our common sense or folk theory of mind. Some supporters of connectionism have argued that if connectionism turns out to be the correct computational model of human cognition, it will follow that humans do not actually have beliefs, desires, or any of the other common mental states referred to by our fold psychology (Ramsey and co-workers, 1990). At this point connectionism is too new and developing too rapidly for there to be much clarity about its philosophical implications.

The Problem of Machine Consciousness. A number of philosophers have argued that problem of subjective consciousness is the real core of the mind–machine question.

They argue that when concerned about whether a computer can have a mind, the real question is, in Thomas Nagel's words, "Whether there is anything that it is like to be such a computer?" in the sense that there is something that it is like to be a human being, a dog, or a bat (Nagel, 1974). Nagel has claimed that most current theories of mind are inadequate because of their neglect of the subjective aspect of mentality and that current attempts to understand the mind by analogy with computers will eventually be recognized as a gigantic waste of time (1986). Searle has also argued for the thesis that the notion of mentality is essentially connected with that of consciousness and that there can be no such thing as a genuinely mental and intentional state that is in principle inaccessible to consciousness (Searle, 1990).

Defenders of the computational theory of mind have responded in a twofold way. They have attempted to reconstruct some aspects of the traditional notion of consciousness within a computational or functionalist framework, and rejected those aspects of the traditional notion of consciousness that resist such treatment as the confused and reactionary residue of discredited Cartesianism (Van Gulick, 1988).

The debate remains undecided. The arguments offered in support of the centrality of consciousness seem intuitively appealing but less than logically compelling. On the other hand, the computational theories of consciousness offered to date are far from adequate, and the success of future theory construction remains an open question.

Antimechanism and the Gödel Argument

The possibility of AI and the computational modeling of mind has also been attacked with philosophical arguments of quite a different sort based on rigorous results in mathematical logic. It has been argued that Kurt Gödel's theorems concerning the imcompleteness of arithmetic and the limits of formal systems show that no machine or computational device can be a completely adequate model of the human mind and that human minds are fundamentally different from machines (see COMPLETENESS). Although these arguments have generated extensive philosophical discussion and debate, many of the central issues remain obscure, and it is often difficult to understand just what claims are being made or denied. The logical results are quite clear, precise, and beyond question. However, their application to questions about mechanism and the human mind remain far from obvious.

In his seminal article "Minds, Machines and Gödel" Lucas (1961) appealed to the Gödel results in an attempt to show that no machine can duplicate the abilities of the human mind. Given any machine that might be thought to do so, Lucas argued that there will always be some sentence the machine cannot show to be true that he (Lucas) can recognize and show to be true. The Gödel results figure in the following way. Gödel's first theorem states that any consistent formal system S containing an adequate axiomatization of arithmetic will be incomplete, ie, there will be a sentence G of S such that neither it nor its negation is a theorem of S. Gödel proved this by showing that if S is an adequate axiomatization of arithmetic, S is adequate to express arithmetic statements that encode statements about its own syntax and proof relation. Thus, it is possible to construct within S a sentence G that encodes the statement that it, G, is not a theorem of S, ie, that it is not provable in S. He also showed that if a sentence K is a theorem of S, there is another sentence of S that says or encodes that K is provable in S, and that the latter sentence will also be provable in S. Thus, if G were provable in S, the sentence that says G is provable in S would also be a theorem of S. But that sentence would be logically equivalent to the negation of G. So both G and its negation would be theorems of S, and S would not be consistent. So if S is consistent, G cannot be proved in S. It is this Gödel sentence G that plays the crucial role in Lucas's argument. He claims that any machine M_i is the concrete instantiation of a formal system S_i such that the sequence of states through which M_i passes in producing a sentence F as output corresponds to a proof of F in S_i. Thus, he argues that for any machine M_i proposed as a candidate model of the human mind there will be a Gödel sentence G_i that M_i cannot produce as true. This is the sentence that says of itself that it is not provable by M_i (or not provable in the formal system S_i that M_i instantiates). But Lucas asserts that he, standing outside of M_i, can see that G_i is true. Thus, there is something he can do that M_i is not able to do, and M_i has failed to duplicate his abilities. Since the argument presented is fully general and applies to any machine, Lucas concludes that no machine can duplicate the abilities of the human mind.

Replies to Gödel's Argument. The replies to Lucas's argument have been numerous and diverse. His claim that any machine is the instantiation of a formal system has been questioned since it is not clear just what counts as a machine in Lucas's sense. A more precise Lucas-style argument can be given if the well-defined notion of a Turing machine (qv) is substituted for Lucas's inexact intuitive notion of a machine. The revised claim is that no Turing machine can be an adequate model of the human mind. It is easy to link formal systems with Turing machines since for any given formal system S_i there is a Turing machine T_i whose output consists exactly of the theorems of S_i. For any such Turing machine T_i there will be a Gödel sentence G_i that it cannot prove, a sentence that says of itself that it is not provable by T_i. Although shifting to a claim about Turing machines provides a clear connection with Gödel's results regarding axiomatized formal systems, it leaves unclear the implications concerning actual concrete machines.

A Turing machine is an abstract device exhaustively specified by its machine table, which is a function from ordered pairs of machine state and input to ordered pairs of subsequent machine state and output. Any actual concrete physical device will be an instantiation of many different Turing machines. As Dennett (1978, pp. 256–266) has argued, the limits that apply to a concrete device under one of its Turing-machine descriptions need not restrict absolutely what it can do as a physical machine or under one of its other descriptions. If an actual device X is an instantiation of Turing machine T_i, it cannot under that description prove G_i, the Gödel sentence of T_i, but X

may well do so under one of its other descriptions. Thus, the Turing-machine version of Lucas's claim might well prove inadequate to establish his larger antimechanist conclusion. Nonetheless, if the Turing-machine claim could be established, it might be sufficient to refute the computational theory of mind and undermine AI hopes to duplicate or fully explain human mental abilities since any modern digital computer is in principle equivalent to some Turing machine (leaving aside the fact that the computer as physical device might never produce its total output).

However, the Turing-machine version of Lucas's argument is also open to objection. Benacerraf (1967) has pressed the question of just what it is that Lucas supposes himself able to do in besting an alleged Turing machine duplicate T_i. As Benacerraf notes, it cannot be proving T_i's Gödel sentence G_i as a theorem of S_i, that is, proving G_i using S_i's axioms and inference rules. Lucas is no more able to do this than is T_i. But if the sense of "prove" is left vague and informal, it no longer remains certain that T_i cannot in this informal sense "prove" G_i.

Another major line of criticism has focused on the fact that the Gödel result applies only to consistent formal systems. Thus, in order to establish the truth of G_i for any Turing machine T_i, Lucas must be able to show that T_i or its corresponding formal system S_i is consistent. How can Lucas know this? Given the dialectical manner in which he presents his imagined contest with the mechanist, Lucas (1961) does have a response that he can and does make on this point. Lucas assumes that humans are consistent. Thus, if a candidate machine is inconsistent, it cannot be an adequate model of the human mind. If it is consistent, the Gödel result applies to it, and it can be shown inadequate by the original argument. The assumption that humans are consistent, on which Lucas's response depends, would seem to be falsified by ordinary facts and observations. Human beings are far from perfectly consistent. Lucas has tried to deny the relevance of familiar human inconsistency by distinguishing between different ways of being inconsistent: as the result of malfunctioning misuse of consistent principles and as the result of the proper use of inconsistent principles. However, Lucas's claim that humans are inconsistent only in the former sense and thus consistent in the sense needed for his argument has been found less than convincing and been regarded as one of his argument's weakest links (Benacerraf, 1967). The consistency assumption is especially problematic since it is not independent of the central point at issue: if humans are Turing machines, it follows by Gödel's second theorem that they cannot prove their own consistency (at least in the sense of proving it within the formal systems corresponding to the machines they instantiate). Thus, for Lucas to assume that he can prove his own consistency might seem to beg the question against his mechanist opponent.

Although Lucas's arguments fail to establish his antimechanist conclusions, the relevance of the Gödel results to AI remains an unresolved but intriguing issue of potentially great philosophical importance. Benacerraf has suggested that perhaps the Gödel theorems show that if humans are Turing machines, they cannot know which machines they are (Benacerraf, 1967), and Hofstadter (1979) has conjectured that they may provide a fundamental key to understanding consciousness and the nature of mind.

EPISTEMOLOGY AND AI

The nature of knowledge has been a central question of philosophical investigation since the birth of philosophy. The problem originally posed by Plato of how to distinguish knowledge (episteme) from true opinion (doxa) remains a subject of debate. Although it is generally accepted today that knowledge cannot be analyzed merely as justified true belief (Gettier, 1963), most philosophers do accept the necessity for including some sort of justification condition in the analysis of knowledge. Knowledge differs from mere true belief at least in part in the rational justification the knower has for his belief. Thus, the theory of knowledge has a greater interest in the nature of human reasoning. Its concerns here naturally overlap with those of AI researchers attempting to formalize systems of rational inference for use in cognitive problem-solving programs and knowledge engineering. Philosophical assessments of the prospects for machine rationality have differed widely and tended to reflect the affinities or divergences between competing philosophical views of rationality and the methods employed by AI. On the whole, philosophers of a phenomenological orientation have been less sympathetic and more pessimistic about AI attempts at programming rationality, and philosophers in the logical empiricist tradition have been more optimistic.

Phenomenological Critique

Critique of Early AI. The phenomenological critique of AI's attempt to formalize human reasoning can be best understood through the work of its most prominent proponent, Dreyfus (see PHENOMENOLOGY). In the two editions of his book "What Computers Can't Do," Dreyfus (1979) aimed to expose the weaknesses and shortcomings of a variety of then existing AI programs. His larger intent was to show that these defects were not merely incidental to the programs he considered nor remediable by further use of the basic techniques originally employed. Dreyfus argued that the programs were flawed in principle and failed to take account of essential features of creative human problem solving and relied on fundamentally mistaken underlying assumptions. Dreyfus examined early AI work in game playing, language comprehension, problem solving, and pattern recognition and found that the programs in each area lacked important abilities possessed by humans.

The chess-playing programs Dreyfus considered rely on heuristically guided searches and examine a large number of possible moves. Human expert players are able to zero in on a small number of promising moves as the result of fringe consciousness, the phenomenon by which implicit background understanding focuses attention and transforms the object of attention. In the chess-playing case (see COMPUTER CHESS AND SEARCH) the fringe consciousness embodies the expert player's implicit understanding of global patterns of board organization acquired through

experience. In problem-solving (qv) programs Dreyfus focused on human insight and the ability to discriminate the essential from the nonessential features of a task situation. Understanding the deep structure of the problem is a necessary first step in human problem solving. Dreyfus noted that in many AI programs this process was carried out not by the program but by the programmer in setting up the problem task and in the choice of factors from which the program was required to fashion a solution.

A similar point applies to computer learning in which the Piercean process of abduction is performed largely by the programmer. The human ability to learn is as much a matter of figuring out what factors are likely to yield regularities and how they must be categorized in order to reveal them as it is of inductively discovering the connections among those variables. However, in the AI learning programs considered by Dreyfus, such as Winston's (1976) arch-learning program, the range of potentially relevant factors was already greatly constrained and categorized by the programmer. Thus, such learning programs at best simulate one component of the human ability to learn, and perhaps the less interesting component. In the area of language comprehension Dreyfus emphasized the human ability to tolerate a high degree of ambiguity in linguistic expressions and to disambiguate them in context on the basis of extralinguistic information relevant to the communication situation. Moreover, he argued that this human ability did not depend on the use of any underlying fully determinate rules. In contrast, the early language-comprehension programs he considered relied on determinate rules sensitive to only a much narrower range of variables actually present in the text.

Context and Holism. Two major themes run through these specific objections, both of which are inspired by the history of the phenomenological movement and the later work of Wittgenstein. One is the global nature of human understanding and the crucial role played by context in comprehension. Facts are not understood in isolation but always as parts of a large-scale structure of meaning. To grasp the significance of any given event, problem, or assertion, one must be able to appreciate its place within a larger context of meanings. Using a notion employed by Husserl (1931), the founder of the phenomenological movement, every object and event is perceived within an "outer horizon." It is the outer horizon that, although not itself explicitly perceived, structures and organizes that of which one is aware. In chess it is the human player's overall understanding of the game or board position that implicitly provides the outer horizon within which the player perceives a given piece or move.

The second major theme is the claim that the implicit organizing background cannot be made explicit. In particular, it cannot be articulated by a system of relations between context-free elements or as an exhaustive set of determinate rules. This view derives not from the work of Husserl but from that of the later phenomenologists Heidegger (1962) and Merleau-Ponty (1962) as well as Wittgenstein's *Philosophical Investigations* (1953). Husserl, who set out to make the organizing framework of meaning explicit, found the project incompletable as the outer horizon of meaning always receded at his approach. Heidegger went further and argued that Husserl had failed to complete his program, not because he had undertaken an enormous or infinite task, but because the very conception of the project was mistaken. Heidegger stressed that the surrounding context of meaning did not consist just of further beliefs, expectations, and rules. Rather, it included the context of social and cultural practices, physical artifacts, tools, and equipment as well as the physical, biological, and historical situation within which human beings live, perceive, and use language. According to Heidegger, it is the actual situations within which human beings live that provide the background for the structure of human meaning, and these situations are not to be confused with a set of beliefs about one's situation or an internalized representation of the situation. It is the situations themselves that provide the boundaries, the limits, and the organizing structure for meaningful behavior. On this view it is simply mistaken to suppose that all of this structure must somehow be internalized in the mind of the agent in order for that person's action to fit within and derive its significance from the larger context.

This second theme is related as well to Wittgenstein's (1953) claim that any attempt to analyze meaningful behavior as acting in accordance with a rule must always require sooner or later the existence of a meaning-giving context that is not itself to be explicated in terms of rule following. Wittgenstein argued that a behavior could be counted as following a given rule only relative to a context. The physical actions of pointing and uttering a sound count as actions of demonstrative reference and naming only relative to a social context that connects these behaviors with a variety of others. And using a word according to a rule to refer on each occasion to the same sort of thing can only make sense relative to a context that determines what is to count as the same kind of thing. If this context were itself to be analyzed in terms of rule following, it would, in turn, presuppose yet a further context within which the social behaviors would count as rule following. If an infinite and vicious regress is to be avoided, at some point there must be a meaning giving social context that is not itself a matter of following rules.

Critique of More Recent AI

It is these two major claims about the global role of context and the impossibility of making the context fully explicit as a system of beliefs or rules that constitute the basis for the continuing phenomenological critique of AI. Given the rapidly changing nature of AI research, Dreyfus's criticisms of early AI programs would be only of historical interest. Indeed, more recent work in AI has sought to remedy many of the defects Dreyfus noted. Chess-playing programs employ descriptions of larger organizational patterns, and language-comprehension programs include a large store of information about the nonlinguistic world. AI programmers have become especially aware of the need to include a great deal of background information about ordinary commonsense matters in their programs, as in Minsky's (1975) Frames (qv) or Schank's Scripts (qv) (Schank and Abelson, 1977). But phenomeno-

logical critics like Dreyfus argue that although these more recent attempts are improvements over early AI programs, they are nonetheless certain to fall short of achieving genuine intelligence or understanding. These critics see the attempt to program the background context of everyday knowledge as a repetition of Husserl's unsuccessful program of phenomenology and fated to fail for the same reasons.

Current AI researchers could be said to have responded to one of the two themes of the phenomenological critique (the importance of context) but to have not accepted the other theme: the claim that the global context should not be thought of as a structure of beliefs, rules, or representations but rather as an actual situation or form of life within which the understander lives. To accept this second claim would be to abandon the project of AI, at least in anything like its current form, and so it is not likely to be acknowledged as was the role of context. As the phenomenological critics admit, the claim that meaning and understanding presuppose a nonrepresentational context is not the sort of claim that can be proved by demonstrative argument. Its force derives rather from an overall picture of meaningful human behavior. In the absence of conclusive argument, AI researchers are not likely to abandon their own competing picture, which treats the background or context of meaning as capable of explicit and determinate formalization. The conflict between AI and its phenomenological critics will have to be settled on the basis of AI's subsequent success or failure rather than on the basis of *a priori* arguments.

Logical Empiricism and AI

Many of the problems faced by AI were also addressed by the philosophical movement known as logical positivism and its successor logical empiricism (Ayer, 1959). Since the methodological assumptions of the positivists were much closer to those of present-day workers in AI than were those of the phenomenologists, philosophers sympathetic to the positivist program are more likely than are phenomenologists to be optimistic about the prospects of AI. The philosopher of science, Glymour (1985), has argued that AI, or at least those areas of AI concerned with machine learning, should be viewed as continuous with the logical empiricist philosophy of science in their goals and approach despite their minimal direct historical connection. The positivist and logical empiricist programs were aimed at a rational reconstruction of scientific method. The result was intended not only to be descriptive in rendering explicit the methods and reasoning that underlay the success of modern science but also to provide an idealized account of scientific method, which might diverge from and improve on some of the practices of actual scientists. Glymour (1985) has pointed out that the goals of the positivist program were similar in many respects to current work in AI. The positivists aimed to formulate precise rules to specify such central scientific notions as the empirical or observational content of a theory, the degree of confirmation of a hypothesis by a body of evidence, and the explanation of an event by the appeal to scientific laws. In each case the positivists demanded a high degree of specificity in any rules that were to count as solutions to these or other problems in their program.

As Glymour has stressed, it is this demand for specificity as much as the problems addressed that marks the similarity between logical positivism and AI. The AI demand for precise and fully determinate rules, which has drawn criticism from phenomenologists, is what most closely links AI with positivism. The positivist's demand grew out of their epistemological commitment to a strong form of foundationalism; semantic properties were attributed only to a limited class of statements concerning the simple immediate objects of sense experience, sense data. All other meaningful statements were to be constructed in terms of logical relations defined over sense-data statements. In consequence, the method allowed for virtually no undefined semantic relations. One could not take as understood such notions as being a positive instance of a hypothesis or being an observation consequence of a theory. All such relations had to be spelled out by precise and largely syntactic rules. Even ordinary commonsense notions such as being a physical object or remaining rigid in motion had to be defined precisely by reference to the restricted class of sense-data statements. AI's demand for specificity has a different source. AI rules must be computable if they are to be implemented, and they must not rely on undefined semantic notions since the machine has no prior knowledge of what has not been programmed into it. Thus, Glymour argues, the different methodological constraints of AI and logical positivism impose a common demand for spelling out crucial semantic notions by specific precise syntactic rules. The formal tools employed also differ since logical positivists relied primarily on the predicate calculus, but the goal of formalization remains the same.

Logical Empiricism and Machine Learning. The fact that logical positivism is generally regarded as having failed to achieve its goals, need not, Glymour contends, reflect unfavorably on AI. The weakness in the positivistic program lay in its commitment to a sense-data foundationalism, which is not shared by AI. One can hope that positivist goals with respect to such problems as formalizing the process of hypothesis confirmation will still be solved by AI. Glymour argues that greater communication between those working on machine learning and those who wish to continue the logical empiricist program would be mutually beneficial. He notes that the AI attempt by Shapiro (1983) to model a logic of confirmation on Popper's logic of scientific discovery (1959) makes clear just what needs to be done to employ Popper's method of conjecture and refutation. In particular, machine implementation requires specific rules to determine which hypothesis to blame or refute when a negative result is obtained with respect to the prediction entailed by a conjunction of hypotheses. Without solving the problem of assessing blame, the general method cannot be programmed. In the absence of the demands for specificity imposed by machine implementation, the importance of the problem can be overlooked in philosophical discussion. Conversely, as the AI application of Popper's method illustrates, AI has much to gain

from research done in the philosophy of science on topics that have naturally and independently arisen in AI.

OTHER AREAS OF PHILOSOPHY

Although epistemology and the philosophy of mind are the two subfields of philosophy with the most direct connection to AI, other areas such as metaphysics, the philosophy of action, and the philosophy of language also have relevance to AI. As McCarthy and Hayes (1969) have noted, this is especially true when one wishes to program general understanding, which requires building into the program a great deal of background knowledge about such common but metaphysically central notions as probability, causality, action, intention, and personhood. Such concepts have long been the subjects of intense philosophical investigation within diverse philosophical traditions and with varying degrees of formal rigor. Although no general philosophical consensus has developed on most of these issues, progress has been made and many promising suggestions, which might tempt uninformed workers in AI, have been explored by philosophers and found unacceptable. Among the many particular topics on which the philosophical literature might be of interest to AI researchers would be causation (see REASONING, CAUSAL) (Sosa, 1975), counterfactuals (Lewis, 1973), conditionals (Harper, 1980), practical reasoning (Jeffrey, 1965), intentional action (Goldman, 1970), the structure of events (Bennett, 1988), speech acts (qv) (Searle, 1969), and conversational implications (Grice, 1975).

BIBLIOGRAPHY

A. J. Ayer, ed., *Logical Positivism,* Macmillan, New York, 1959.

W. Bechtel, "Connectionism and the Philosophy of Mind," *South. J. Philos.* supplement 17–41 (1987).

P. Benacerraf, "God, the Devil, and Gödel," *Monist* **51,** 9–32 (1967).

J. Bennett, *Events and their Names,* Hackett, Indianapolis, Ind., 1988.

N. Block, "Troubles with Functionalism," in C. W. Savage, ed., *Perception and Cognition, Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science,* Vol. 9, University of Minnesota Press, Minneapolis, Minn., 1978, pp. 261–325.

N. Block, "What is Functionalism?" in N. Block, ed., *Readings in the Philosophy of Psychology,* Vol. 1, Harvard University Press, Cambridge, Mass., 1980, pp. 171–184.

N. Block, "Psychologism and Behaviorism," *Philos. Rev.* **90,** 5–43 (1981).

N. Block, "Advertisement for a Semantics for Psychology," in *Midwest Studies in Philosophy,* Vol. 10, University of Minnesota Press, Minneapolis, Minn., 1986.

P. Churchland, *Neurophilosophy,* MIT Press, Cambridge, Mass., 1986.

P. M. Churchland and P. S. Churchland, "Functionalism, Qualia, and Intentionality," *Philos. Top.* **12,** 121–145 (1981).

R. Cummins, *Meaning and Mental Representation,* MIT Press, Cambridge, Mass., 1989.

D. C. Dennett, "Why the Law of Effect Won't Go Away," *J. Theor. Soc. Behav.* **2,** 169–187 (1975); reprinted in Dennett, 1978.

D. C. Dennett, *Brainstorms,* MIT Press, Cambridge, Mass., 1978.

D. C. Dennett, *The Intentional Stance,* MIT Press, Cambridge, Mass., 1987.

F. Dretske, *Knowledge and the Flow of Information,* MIT Press, Cambridge, Mass., 1981.

H. Dreyfus, *What Computers Can't Do,* 2nd ed., Harper & Row, New York, 1979.

J. Fodor, *Psychological Explanation,* Random House, New York, 1968.

J. Fodor, "Methodological Solipsism Considered as a Research Strategy for Cognitive Science," *Behav. Brain Sci.* **3,** 63–109 (1980).

J. Fodor, *Psychosemantics,* MIT Press, Cambridge, Mass., 1987.

J. Fodor and Z. Pylyshyn, "Connectionism and Cognitive Architecture: A Critical Analysis," *Cognition* **28,** 2–71 (1988).

E. Gettier, "Is Justified True Belief Knowledge," *Analysis* **23,** 121–123 (1963).

C. Glymour, "Android Epistemology: Reflections on Artificial Intelligence and the Philosophy of Science," paper presented at Pacific Division Meeting of the American Philosophical Association, San Francisco, Calif., Mar. 1985.

Goldman, *A Theory of Human Action,* Prentice-Hall, Englewood Cliffs, N.J., 1970.

H. P. Grice, "Logic and Conversation," in D. Davidson and G. Harman, eds., *The Logic of Grammar,* Dickenson, Encino, Calif., pp. 64–74, 1975.

K. Gunderson, "The Imitation Game," *Mind* **73,** 234–245 (1964).

W. Harper, *Ifs: Conditionals, Belief, Decision, Chance and Time,* Reidel, Dordrecht, The Netherlands, 1980.

J. Haugeland, *AI, The Very Idea,* MIT Press, Cambridge, Mass., 1985.

M. H. Heidegger, in J. Macquarrie and E. Robinson, eds., *Being and Time,* p. 45; reprinted in M. Merleau-Ponty, *Phenomenology of Perception,* Routledge and Kegan Paul, London, 1962.

D. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid,* Basic Books, New York, 1979.

E. Husserl, *Ideas General Introduction to Pure Phenomenology,* Macmillan, New York, 1931.

D. Jeffrey, *The Logic of Decision,* McGraw-Hill, New York, 1965.

D. Lewis, "Psychophysical and Theoretical Identification," *Austral. J. Philos.* **50,** 249–258 (1972).

D. Lewis, *Counterfactuals,* Harvard University Press, Cambridge, Mass., 1973.

J. Locke, *An Essay Concerning Human Understanding,* edition of A. Fraser, Dover, New York, 1959.

J. R. Lucas, "Minds, Machines and Gödel," *Philosophy* **36,** 112–127 (1961).

J. McCarthy and P. J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence," in B. Meltzer and D. Michie, eds., *Mach. Intell.,* Vol. 4, Halstead, New York, 1969, pp. 463–502.

M. Merleau-Ponty, *Phenomenology of Perception,* Routledge and Kegan Paul, London, 1962.

R. Millikan, *Language, Thought and Other Biological Categories,* MIT Press, Cambridge, Mass., 1984.

M. Minsky, "A Framework for Representing Knowledge," Memo 306, MIT Artificial Intelligence Laboratory, Cambridge, Mass., excerpts published in P. H. Winston, ed., *The Psychology of Computer Vision,* McGraw-Hill, New York, 1975, pp. 211–277.

T. Nagel, "What Is It Like to Be a Bat?" *Philos. Rev.* **83,** 435–450 (1974).

T. Nagel, *The View from Nowhere,* Oxford University Press, New York, 1986.

K. R. Popper, *The Logic of Scientific Discovery,* Hutchinson, London, 1959.

H. Putnam, "The Nature of Mental States," in W. H. Capitan and D. D. Merrill, eds., *Art, Mind, and Religion,* University of Pittsburgh Press, Pittsburgh, Pa., 1967, pp. 37–48.

W. Ramsey, S. Stich, and J. Garon, "Connectionism, Eliminativism, and the Future of Folk Psychology," in J. Tomberlin, ed., *Philosophical Perspectives 4: Action Theory and the Philosophy of Mind,* Ridgeview Publishing, Atascadero, Calif., 1990.

R. Schank and R. Abelson, *Scripts, Plans, Goals and Understanding,* Lawrence Erlbaum, Hillsdale, N.J., 1977.

M. Scriven, "The Mechanical Concept of Mind," *Mind* **62,** 230–240 (1953).

J. Searle, *Speech Acts,* Cambridge University Press, Cambridge, 1969.

J. Searle, "Minds, Brains and Programs," *Behav. Brain Sci.* **3,** 417–457 (1980).

J. Searle, "Consciousness, Explanatory Inversion, and Cognitive Science," *Behav. Brain Sci.* **13,** 585–642 (1990).

E. Shapiro, *Algorithmic Program Debugging,* MIT Press, Cambridge, Mass., 1983.

J. J. C. Smart, "Sensations are Brain Processes," *Philos. Rev.* **68,** 141–156 (1959).

P. Smolensky, "On the Proper Treatment of Connectionism," *Behav. Brain Sci.* **11,** 1–74 (1988).

E. Sosa, *Causation and Conditionals,* Oxford University Press, London, 1975.

S. Stich, "Dennett on Intentional Systems," *Philos. Top.* **12,** 39–62 (1981).

A. Turing, "Computing Machinery and Intelligence," *Mind* **59,** 433–460 (1950).

R. Van Gulick, "A Functionalist Plea for Self-Consciousness," *Philos. Rev.* **97,** 149–188 (1988).

P. Winston and staff of MIT AI Laboratory, Proposal to ARPA, MIT AI Memo 336, Cambridge, Mass., 1976.

L. Wittgenstein, *Philosophical Investigations,* Blackwell, Oxford, 1953.

General References

A. R. Anderson, *Minds and Machines,* Prentice-Hall, Englewood Cliffs, N.J., 1964 (an anthology of important articles from the early period of the mind–machine debate).

P. M. Churchland, *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind,* MIT Press, Cambridge, Mass., 1984.

H. Dreyfus, ed., *Husserl, Intentionality and Cognitive Science,* MIT Press, Cambridge, Mass., 1982 (a collection of essays bringing the phenomenological perspective to bear on work in AI and cognitive science).

C. Glymour, *Theory and Evidence,* Princeton University Press, Princeton, N.J., 1980 (a detailed account of confirmation).

J. Haugeland, *Mind Design Philosophy, Psychology, and Artificial Intelligence,* MIT Press, Cambridge, Mass., 1982 (an excellent anthology of recent philosophical and theoretical work on AI).

J. Lucas, *The Freedom of the Will,* Oxford University Press, Oxford, 1970 (includes further discussion of the Gödel argument with bibliography).

A. Newell, "Intellectual Issues in the History of Artificial Intelligence," in F. Machlup and U. Mansfield, eds., *The Study of Information Interdisciplinary Messages,* John Wiley and Sons, Inc., New York, 1983, pp. 187–228.

F. Suppe, ed., *The Structure of Scientific Theories,* 2nd ed., University of Illinois Press, Urbana, Ill., 1979 (includes a critical introduction that provides an overview of recent work in the philosophy of science that would be of relevance to AI).

R. Van Gulick
Syracuse University

PHRAN AND PHRED

A natural-language analyzer and a natural-language generator that have been developed around 1980 and used at the University of California at Berkeley in Robert Wilensky's research group, these two programs have served as the front-end and back-end, respectively, of planning systems like PAM (qv) and UC. PHRAN was written by Yigal Arens, and PHRED was written by Steve Upstill and Paul Jacobs (see R. Wilensky, *Planning and Understanding,* Addison-Wesley, Reading Mass., 1983, and F. Rose, *Into the Heart of the Mind,* Harper & Row, New York, 1984, pp. 98–115).

K. S. Arora
SUNY at Buffalo

PHYSICS, NAIVE

Naive physics is the body of knowledge that people have about the surrounding physical world. The main enterprises of naive physics are explaining, describing, and predicting changes in the physical world. There is an important distinction between classic physics and naive physics. Classical physics is based on the presupposition that there is a shared unstated common sense prephysical knowledge rooted in experience. Naive physics is this prephysical knowledge rooted in experience. It is important to notice that in classical physics, concepts such as state, law, cause equilibrium, oscillation, momentum, feedback, etc are qualitative in nature. However, they have been embedded in a complex framework established by the mathematics of real numbers and differential equations. The relationship between qualitative (commonsense) models and mathematical models can be shown as in Figure 1. Qualitative simulation captures less detail and, therefore, may produce partial behavioral descriptions. Also, the quantitative precision of these descriptions is reduced whereas crucial distinctions are retained.

A research directed at understanding and modeling naive physics reasoning concentrates mainly on deriving the qualitative concepts used in naive physics from formal models and identifying the core knowledge underlying physical intuition (Hobbs and Moore, 1985). The main conjectures guiding the research are that naive physics knowledge, to a large degree, can be derived from commonsense observation and that quantitative laws can be mapped to qualitative ones.

ENCYCLOPEDIA OF ARTIFICIAL INTELLIGENCE

This extensively revised and expanded *Second Edition* of the *Encyclopedia of Artificial Intelligence* defines the discipline by bringing together the core of knowledge from all fields encompassed by AI. It covers the latest developments in current AI topics such as neural networks, fuzzy logic, machine vision, natural language generation, and many more. Includes:

- Over 450 articles—all entries written expressly for the *Encyclopedia*
- Over 5,000 literature references; 454 illustrations and color photographs
- Over 50% new and revised material
- Exemplary indexing and cross-referencing for easy, complete information access to all topics

Praise for the *First Edition*...

"The *Encyclopedia* is a wonder of clarity and scope: surprisingly easy to read...the clarity is an especially pleasant surprise, considering the articles were all written by AI experts...It's a treasure house of easily accessible knowledge."
—*Language Technology*

"Excellent bibliographies are attached to most of the articles, and diagrams and sketches are clear and helpful. The indexing and cross-indexing are exemplary. As the editor points out, the reader will be led by the extensive cross-references to almost every other article..." —*Artificial Intelligence Reporter*

"The *Encyclopedia* is a first-class piece of work that will be an indispensable part of any AI library." —*Computing Reviews*

"...A tour de force...a truly fantastic encyclopedia which no one in the field of artificial intelligence can afford to be without."
—*Systems Research & Information*

WILEY-INTERSCIENCE
John Wiley & Sons, Inc.
Professional, Reference and Trade Group
605 Third Avenue, New York, N.Y. 10158-0012
New York • Chichester • Brisbane • Toronto • Singapore
ISBN 0 471-50307-X (Two-Volume Set)
ISBN 0 471-50305-3 (Vol. 1)
ISBN 0 471-50306-1 (Vol. 2)

ISBN 0-471-50306-1

9 780471 503064 90000