ENCYCLOPEDIA OF ARTIFICIAL INTELLIGENCE SECOND EDITION

VOLUME 1

Stuart C. Shapiro, *Editor-in-chief*

A Wiley-Interscience Publication

John Wiley & Sons, Inc.

New York / Chichester / Brisbane / Toronto / Singapore

Italy, Morgan-Kaufmann, San Mateo, Calif., 1987, pp. 366–372.

M. Ben-Bassat, R. W. Carlson, V. K. Puri, E. Lipnick, L. D. Portigal, and M. H. Weil, "Pattern-based Interactive Diagnosis of Multiple Disorders: The MEDAS System," *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-2,** 148–160 (1980).

G. F. Cooper, *NESTOR: A Computer-based Medical Diagnostic Aid That Integrates Causal and Probabilistic Knowledge*, Report No. STAN-CS-84-1031, Stanford University, Stanford, Calif., November 1984.

R. P. Goldman, "A Probabilistic Approach to Language Understanding," Ph.D. dissertation, Brown University, Providence, Rhode Island, 1990.

D. Heckerman, E. Horvitz, and B. Nathwani, "Update on the Pathfinder Project." *Proceedings of the Thirteenth Symposium on Computer Applications in Medical Care*, Washington, D.C., IEEE Computer Society Press, Silver Spring, M.D., 1989, pp. 203–207.

J. Kim and J. Pearl, "*CONVINCE: A CONV*ersational *IN*ference Consolidation *E*ngine," *IEEE Trans. Systems Man Cybernetics*, **SMC-17**(2), 120–132 (1987).

D. J. Spiegelhalter and R. P. Knill-Jones, "Statistical and Knowledge-based Approaches to Clinical Decision-Support Systems, With an Application to Gastroenterology," *J. Roy Stat. Soc. A*(147), 35–77 (1984).

Quasi-Bayesian Systems

R. O. Duda, P. E. Hart, P. Barnett, J. Gaschnig, K. Konolige, R. Reboh, and J. Slocum, "Development of the PROSPECTOR Consultant System for Mineral Exploration, Final Report for SRI Projects 5821 and 6915," Artificial Intelligence Center, SRI International, Menlo Park, Calif., 1978.

C. Kulikowski and S. Weiss, "Representation of Expert Knowledge for Consultation: The CASNET and EXPERT Projects," in P. Szolovitz, ed., *Artificial Intelligence in Medicine*, Westview Press, Boulder, Colo., 1982, pp. 21–55.

R. A. Miller, H. E. Pople, and J. P. Myers, "INTERNIST-1, An Experimental Computer-based Diagnostic Consultant for General Internal Medicine," *N. Engl. J. Med.* **307**(8), 468–470 (1982).

J. R. Quinlan, "INFERNO: A Cautious Approach to Uncertain Inference," Rand Note N-1898-RC, September 1982.

E. H. Shortliffe, *Computer-Based Medical Consultation: MYCIN*, Elsevier, New York, 1976.

J. PEARL
UCLA

This work was supported in part by the National Science Foundation, Grant DSR 83-13875.

BEAM SEARCH. See SEARCH, BEAM.

BELIEF REPRESENTATION SYSTEMS

A belief system may be understood as a set of beliefs together with a set of implicit or explicit procedures for acquiring new beliefs. The computational study of belief systems has focused on building computer sytems for representing or expressing beliefs or knowledge and for reasoning with or about beliefs or knowledge. Such a system is often expressed in terms of a formal theory of the syntax and semantics of belief and knowledge sentences.

Reasons for Studying Such Systems

There are several distinct, yet overlapping, motivations for studying such systems. As McCarthy and Hayes (1969), two of the earliest contributors to this field, have explained,

> A computer program capable of acting intelligently in the world must have a general representation of the world. . . . [This] requires commitments about what knowledge is and how it is obtained. . . . This requires formalizing concepts of causality, ability, and knowledge.

Thus, one motivation is as a problem in knowledge representation (see KNOWLEDGE REPRESENTATION). In the present context this might less confusingly be referred to as "information representation" since not only knowledge but also beliefs are represented. A second motivation is as a component of computational studies of action. Subcategories of the latter include planning systems (eg, Moore, 1977), systems for planning speech acts (eg, Cohen and Perrault, 1979), and systems for planning with multiple agents (eg, Appelt, 1980). These systems frequently involve representing and reasoning about other notions as well (such as can, want, etc).

A third motivation is the construction of AI systems that can interact with human users, other interacting AI systems, or even itself (eg, Konolige and Nilsson, 1980; McCarthy, 1977). Among the subcategories here are the study of user models for determining appropriate output (eg, Rich, 1979a,b) and the prediction of others' behavior and expectations on the basis of their beliefs (McCarthy, 1979). A fourth motivation is directly related to such interaction: the study of AI systems that can converse in natural language (eg, Wilks and Bien, 1979), either with users or with a "knowledge base" (eg, Levesque, 1984). A fifth motivation is the study of reasoning: how a particular individual reasons (Abelson and Reich, 1969), or how reasoning can be carried out with incomplete knowledge (eg, Halpern and McAllester, 1984), or in the face of resource limitations (eg, Konolige, 1983). Finally, there is the ever-present motivation of modeling a mind (eg, Abelson, 1973; Maida and Shapiro, 1982) or providing computational theories of human reasoning about beliefs (eg, Creary, 1979; Maida, 1983).

Types of Theories

There are four overlapping types of theories identifiable by research topics or by research methodologies. One is belief revision, which is concerned with the problem of revising a system's database in light of new, possibly conflicting information (cf Martins and Shapiro, 1988); such theories are dealt with in another entry. The other types of theory can be usefully categorized [by augmenting the scheme of McCarthy and Hayes (1969)] as (a) epistemological theories, concerned primarily with representational issues [eg, McCarthy (1979)]; (b) formal heuristic theories,

concerned primarily with the logic of belief and knowledge, that is, with reasoning in terms of a formal representation [eg, Moore (1977)]; and (c) psychological heuristic theories, also concerned with reasoning but using techniques that make some explicit claim to psychological adequacy—such theories typically are not concerned with representational issues *per se* [eg, Colby and Smith (1969); Wilks and Bien (1983)].

PHILOSOPHICAL BACKGROUND

Much of the data, problems, and theories underlying AI research on formal belief systems has come from philosophy, in particular, epistemology, philosophy of language, and logic (especially modal and intensional logics).

Philosophical Issues

There are several philosophical issues—logical, semantic, and ontological—that have been faced by AI researchers working on belief systems.

1. The problem of the relationship between knowledge and belief. This problem, dating back to Plato's *Theaetetus*, is usually resolved by explicating knowledge as justified true belief (see Gettier, 1963 for the standard critique of this view and Fetzer, 1985 for a discussion in the context of AI).

2. The problem of the nature of the objects of belief, knowledge, and other intentional (ie, cognitive) attitudes: are such objects extensional (eg, sentences, physical objects in the external world) or intensional (ie, nonextensional; eg, propositions, concepts, mental entities)?

3. Problems of referential opacity: the failure of substitutability of co-referential terms and phrases in intentional contexts. This can best be illustrated as a problem in deduction. From

> Susan believes that the Morning Star is a planet

and

> The Morning Star is a planet if and only if the Evening Star is a planet,

it does not logically follow that

> Susan believes that the Evening Star is a planet.

Nor from

> Ruth believes that Venus is a planet

and

> Venus = the Evening Star

does it logically follow that

> Ruth believes that the Evening Star is a planet.

4. The problem of quantifying in (ie, into intentional contexts). From

> Carol believes that the unicorn in my garden is white,

it does not logically follow that

> There is a unicorn in my garden such that Carol believes that it is white.

5. Problems of logical form (or semantic interpretation, or "knowledge representation" in the sense of AI): how should the following kinds of sentences be understood, and what are their relationships with simpler cases of belief and knowledge?

> Margot knows whether Ben's phone number is the same as Ariana's.
>
> Mike knows who Sally is.
>
> Jan believes that Stu believes that he is a philosopher.
>
> Harriet and Frank mutually believe that the movie at Loew's starts at 9 p.m.

6. The problem of the distinction between *de re* and *de dicto* beliefs: When a belief is a cause of a person's actions, one is not only interested in what the person believes, but also in how the person believes it. That is, one is not only interested in a third-person characterization of the agent's beliefs, but also in the agent's *own* characterization of those beliefs. Suppose that Ralph sees the person whom he knows to be the janitor stealing some government documents, and suppose—unknown to Ralph—that the janitor has just won the lottery. Then Ralph believes *de dicto* that the janitor is a spy, and he believes *de re* that the lottery winner is a spy. That is, if asked, Ralph would assent to the proposition "The janitor is a spy"; but he merely believes of the man known to the hearer as the lottery winner that he is a spy—Ralph would not assent to "The lottery winner is a spy." Traditionally viewed, a belief *de dicto* is a referentially opaque context, whereas a belief *de re* is referentially transparent. Thus, the inference

> Ralph believes [*de dicto*] that the janitor is a spy.
> The janitor = the lottery winner.
> ---
> Ralph believes [*de dicto*] that the lottery winner is a spy.

is invalid. Moreover, its conclusion not only presents false information but it also represents a loss of information, namely, of the information about the propositional "content" of Ralph's belief. On the other hand,

> Ralph believes [*de re*] of the janitor that he is a spy.
> The janitor = the lottery winner.
> ---
> Ralph believes [*de re*] of the lottery winner that he is a spy.

is valid. But the conclusion conveys just as little information about Ralph's actual belief *de dicto* as does the first premise. An AI system that is capable of explaining or recommending behavior must be able to distinguish between these two kinds of belief reports by having two distinct means of representing them.

Epistemic Logic

Of central importance from the point of view of AI have been the logics of belief and knowledge proposed by Hintikka (1962). The propositional fragment of Hintikka's logic of knowledge (propositional epistemic logic) can be axiomatized as a notational variant of the modal logic **S4** (see LOGIC, MODAL), replacing the necessity operator by a family of proposition-forming operators $\mathbf{K}_a$, for each individual a ($\mathbf{K}_a p$ is to be read "a knows that p"). The axioms are

(A1) If p is a tautology, then $\vdash p$.

(A2) If $\vdash p$ and $\vdash (p \rightarrow q)$, then $\vdash q$.

(A3) If $\vdash p$, then $\vdash \mathbf{K}_a p$.

(A4) $\vdash (\mathbf{K}_a p \rightarrow p)$.

(A5) $\vdash (\mathbf{K}_a p \rightarrow \mathbf{K}_a \mathbf{K}_a p)$.

(A6) $\vdash ([\mathbf{K}_a p \wedge \mathbf{K}_a (p \rightarrow q)] \rightarrow \mathbf{K}_a q)$

Roughly, (A3) says that a knows all theorems, (A4) says that what is known must be true (recall that knowledge is generally considered to be justified true belief), (A5) says that what is known is known to be known, and (A6) says that what is known to follow logically from what is known is itself known. A (propositional) logic of belief (a propositional doxastic logic) can be obtained by using operators $\mathbf{B}_a$ and deleting (A4); other epistemic and doxastic logics can be obtained by taking similar variants of other modal logics.

Possible-worlds semantics for epistemic and doxastic logics can be provided as in ordinary modal logics by interpreting the accessibility relation between possible worlds as a relation of epistemic or doxastic alternativeness. Thus, for example,

> $\mathbf{K}_a p$ is true in possible world w if and only if p is true in possible world w' for all w' that are epistemic alternatives to w.

Intuitively, a knows that p if and only if p is compatible with everything that a knows [see Hintikka (1962, 1969) for details]. Various restrictions on the alternativeness (or accessibility) relation yield correspondingly different systems. Thus, **S4** can be characterized semantically by requiring the relation to be only reflexive and transitive. If symmetry is allowed, the semantics characterizes the stronger system $\mathbf{S5} = \mathbf{S4} + \vdash \neg \mathbf{K}_a p \rightarrow \mathbf{K}_a \neg \mathbf{K}_a p$. (Roughly, what is unknown is known to be unknown.)

Note that none of these systems is psychologically plausible. For example, no one knows or believes all tautologies or all logical consequences of one's knowledge or beliefs as suggested by (A6). Nor is it clear how to interpret (A5)—is the consequent to be read as "a knows that a knows that p" or as "a knows that he (or she) knows that p"?—nor whether it is plausible. Indeed, some philosophers feel that there are no axioms that characterize a psychologically plausible theory of belief. There is a large philosophical literature discussing these issues (eg, Castañeda, 1964, the special issues of *Noûs* **1** (1967), and *Synthèse* **21** (1970)]. Other formalizations of epistemic logics that are of relevance to AI are to be found in Sato (1976) and McCarthy and co-workers (1978). Further discussion of the philosophical issues may be found in Linsky (1977), Edwards (1967), and through *The Philosopher's Index*. Interesting recent work on semantics of belief sentences dealing with linguistics and computational issues may be found in Moore and Hendrix (1982), Moravcsik (1973), and Partee (1967, 1973).

SURVEY OF THEORIES AND SYSTEMS

In this section the major published writings on belief systems are surveyed following the three-part categorization of types of theories and by lines within the types. The reader is reminded that the categorization is highly arbitrary and that virtually all of the research falls into more than one category.

Epistemological Theories

Early Work. One of the earliest works on AI belief systems, by McCarthy and Hayes (1969), begins by considering a system of interacting automata whose states at a given time are determined by their states at previous times and by incoming signals from the external world (including other automata). A person p is considered to be a subautomaton of such a system. Belief is represented by a predicate B, where $B_p(s, w)$ is true if p is to be regarded as believing proposition w when in state s. Four sufficient conditions for a "reasonable" theory of belief are given:

1. p's beliefs are consistent and correct.
2. New beliefs can arise from reasoning on the basis of other beliefs.
3. New beliefs can arise from observations.
4. If p believes that it ought to do something, then it does it.

However, criterion 1 is psychologically implausible and seems to better characterize knowledge; criterion 4 is similarly too strong. Knowledge is represented by a version of Hintikka's system (1962): The alternativeness relation, $shrug(p, s_1, s_2)$, is true if and only if: if p is in fact in situation s_2, then for all he knows he might be in situation s_1. (A "situation" is a complete, actual or hypothetical state of the universe.) $K_p q$ is true (presumably at s) if and only if $\forall t[shrug(p, t, s) \rightarrow q(t)]$, where $q(t)$ is a "fluent"—a Boolean-valued function of situations—that "translates" q, and where *shrug* is reflexive and transitive. Although this paper is significant for its introduction of philosophical concepts into AI, it discusses only a minimal representation of knowledge and belief.

A more detailed representation is offered by McCarthy (1977, 1979) in which individual concepts—that is, intensional entities somewhat like Fregean senses—are admitted as entities on a par with extensional objects, to allow for first-order expression of modal notions without problems of referential opacity. Notationally, capitalized terms stand for concepts, lowercase terms for objects.

Thus, *know*(*p*, *X*) is a Boolean-valued (extensional) function of a person *p* (an extensional entity) and a concept *X* (an intensional entity), meaning "*p* knows the value of *X*," defined as *true Know*(*P*, *X*), where *true* is a Boolean-valued function of propositions, and where *Know*(*P*, *X*) is a proposition-valued (ie, concept-valued) function of a person = concept *p* and a concept *X*. Nested knowledge is handled by *Know* rather than *know*; thus, "John knows whether Mary knows the value of *X*" is *Know*(John, *Know*(Mary, *X*)). The Hintikka-style knowledge ("knowledge-that") is represented by a function *K*(*P*, *Q*), defined as (*Q And Know*(*P*, *Q*)); thus, "John knows that Mary knows the value of *X*" is *K*(John, *Know*(Mary, *X*)). A denotation function maps intensional concepts to extensional objects, and a denotation relation, *denotes*, is introduced for concepts that lack corresponding objects. An existence predicate can be defined in terms of the latter: *true Exists X* if and only if ∃*x*[*denotes*(*X*, *x*)]. Belief is not treated in nearly as much detail. Functions *Believe* and *believe* are introduced, though so are functions *believespy* and *notbelievespy* (to handle a celebrated puzzle of referential opacity concerning spies; see Linsky, 1977), yet no axioms are provided to relate them to each other or to the ordinary belief functions. [A similar theory in the philosophical literature was described in Rapaport (1978).]

Creary (1979) extended McCarthy's theory to handle concepts of concepts. According to Creary, McCarthy's notation cannot represent three distinct readings of

Pat believes that Mike wants to meet Jim's wife

(generated by the *de re*/*de dicto* distinction) because it does not allow for the full hierarchy of Fregean senses (Frege, 1892). The three readings are

believes(pat, Wants{Mike, Meet${Mike$, Wife$ Jim$}})

believes (pat, Exist P$.Wants{Mike, Meet${Mike$, P$}} And Conceptof{P$, Wife Jim})

∃P$ P.believes(pat, Wants{Mike, Meet${Mike$, P$}}) ∧ conceptof(P$, P) ∧ conceptof(P, wife jim)

Here, if mike is the name of a person whose concept is: Mike, then Mike is the name of that concept and its concept is: Mike$, etc. It is not clear, however, that such a hierarchy is needed at all (cf Parsons, 1981) nor whether McCarthy's notation is indeed incapable of representing the ambiguity. Creary does, however, discuss reasoning about propositional attitudes of other agents by "simulating" them using "contexts"—temporary databases consisting of the agent's beliefs plus common beliefs and used only for reasoning, not for representation [thus escaping certain objections to "database approaches" raised by Moore (1977)]. Creary's system was subjected to criticism and refinement by Barnden (1983).

Barnden has revised and extended his own theory to solve a problem that he has identified as "incorrect imputation" (1986, 1989). For instance, the first reading above appears to "impute" to Pat a theory of concepts of concepts (ie, a theory of second-order concepts) that Pat might never have thought about. This seems, however, to be nothing more than the familiar *de re*/*de dicto* distinction.

Belief Spaces. The problems of nested beliefs and of the *de re*/*de dicto* distinction suggest that databases containing representations of beliefs should be partitioned into units (often called "contexts," "spaces," or "views") for each believer. One of the earliest discussions of these issues in a computational framework was by Moore (1973), who developed a LISP-like language, D-SCRIPT, that evaluates objects of belief in different environments (see also Bien, 1975). Another early use of such units was Hendrix's (1979) partitioning of semantic networks into "spaces" and "vistas": The former can be used to represent the propositions that a given agent believes; the latter are unions of such spaces. Similarly, Schneider (1980) introduced "contexts" to represent different views of a knowledge base, and Covington and Schubert (1980) used "subnets" to represent an individual's conception of the world. Filman and co-workers (1983) treat a context as a theory of some domain, such as an agent's beliefs, with the ability to reason with the agent's beliefs in the context and about them by treating the context as an object in a metacontext.

Fully Intensional Theories. The notions of intensional entities and belief spaces come together in the work of Shapiro and his associates. Maida and Shapiro (1982) go a step beyond the approach of McCarthy by dropping extensional entities altogether. Their representational scheme, SNePS (qv), uses a fully intensional semantic network in which all nodes represent distinct concepts, all represented concepts are represented by distinct nodes, and arcs represent binary relations between nodes but cannot be quantified over (they are "nonconceptual"). The entire network is considered to model the belief system of an intelligent agent: asserted propositional nodes represent the agent's beliefs, and "base" nodes represent individual concepts. [Similar philosophical theories are those of Meinong (1904) and Castañeda (1972); see Rapaport (1985).] Two versions of 'know' are treated (both via agent–verb–object case frames): *know1* for "knows that" and *know2* for "knows by acquaintance." There are corresponding versions of 'believe' (though it is not clear what *believe2* is); the fundamental principle connecting knowledge and belief is that the system believes1 that an agent knows1 that *p* only if the system believes1 both that the agent believes1 that *p* and that the agent believes1 that *p* for the right reasons. Unlike other belief systems, their system can handle questions, as queries about truth values (which are represented by nodes). Thus, whereas most systems represent "John knows whether *p*" as "John knows that *p* or John knows that ¬*p*," Maida and Shapiro (1982) consider these to be merely logically equivalent but not intensionally identical; instead, they represent it as "John knows2 the truth value of *p*." Among the consequences of the fully intensional approach are (*1*) the ability to represent nested beliefs without a type hierarchy [see Maida (1983)], (*2*) the need for a mechanism of coreferentiality (actually, their "*a* EQUIV *b*" represents that *the system believes that a* and *b* are coreferential), (*3*) the dynamic introduction of new nodes, through user interaction, in the order they are needed (which sometimes requires node merging by means of EQUIV arcs), and (*4*)

the treatment of all transitive verbs as referentially opaque unless there is an explicit rule to the contrary.

Rapaport (1986) [see also Rapaport and Shapiro (1984)] makes essential use of the notion of a "belief space" to represent the distinctions between *de re* and *de dicto* beliefs. In dynamically constructing the system's belief space, he follows the principle that if there is no prior knowledge of coreferentiality of concepts in the belief spaces of agents whose beliefs are being modeled by the system, then those concepts must be represented separately. This has the effect of reintroducing a kind of hierarchy [see the discussion of Creary (1979), above], but there is a mechanism for "merging" such entities later as new information warrants. Thus, the conjunctive *de dicto* proposition "John believes that Mary is rich and Mary believes that Lucy is rich" requires four individuals: the system's John, the system's John's Mary, the system's Mary, and the system's Mary's Lucy. But the *de re* proposition "John believes of Mary that she is not rich" only requires two: the system's John and the system's Mary. This technique is used to represent quasi-indicators (Castañeda, 1967; Sells, 1987): virtually all other systems fail to distinguish between "John believes that he* is rich" and "John believes that John is rich" [although Moore, 1980 and Smith (1986) briefly discuss this]; the starred, quasi-indexical occurrence of "he" is the *system's* way of depicting *John's* use of 'I' in John's statement, "I am rich." This is represented as a *de dicto* proposition requiring two individuals: the system's John and the system's John's representation of himself (which is distinct from the system's John's John). This theory has been extended by Wiebe (Wiebe and Rapaport, 1986).

Other Theories. Among other theories that may be classified as epistemological (though some have considerable overlap with formal heuristic theories) are the important early work of Konolige (1982), a series of papers by Kobsa (1984a–c) and Kobsa and Trost (1984), Xiwen and Weide (1983), and Soulhi (1984).

Konolige. Konolige (1982) is concerned with the other side of the coin of knowledge: ignorance. In order to prove ignorance based on knowledge limitations ["circumscriptive ignorance"; see McCarthy (1980)], he uses a representation scheme based on a logic called *KI*4, an extension of the work of Sato (1976). *KI*4 has two families of modal operators: knowledge operators, $[S]$, for each agent S, and (what might be called "context") operators, $[\alpha]$, for each proposition α; and it has an agent 0 ("fool"), where $[0]\alpha$ means "α is common knowledge." The axioms and rules of *KI*4 include analogs of (A1)–(A6)(system $K4$), plus:

(A7) $\vdash [0]\alpha \rightarrow [0][S]\alpha$

(A8) If $\alpha \vdash_{K4} \beta$, then $\vdash_{KI4} [\alpha]\beta$

(A9) If not-$(\alpha \vdash_{K4} \beta)$, then $\vdash_{KI4} \neg[\alpha]\beta$

Roughly, (A7) says that if α is common knowledge, then it is common knowledge that S knows it; (A8) says that if β follows from α in $K4$, then β is true in the context of α in *KI*4; and (A9) says that if β does not follow from α in $K4$, then it is not true in the context of α in *KI*4. The context operator may be explained as follows: If $\alpha = [S]q$, then $[\alpha]$ identifies S's theory whose axiom is q. Thus, "all S knows about p is that q_1 or q_2" can be represented as: $[\alpha][S]p$, where $\alpha = [S]q_1 \vee [S]q_2$.

Kobsa and Trost. Kobsa and Trost (1984) use the KL-ONE knowledge representation system, augmented by their version of partitions: "contexts"—collections of "nexus" nodes linked to "concept" nodes, representing that the agent modeled by the context containing the nexus nodes believes propositions about the concepts. There is a system context and separate contexts for each agent whose beliefs are modeled, with explicit (co-referential-like) links between isomorphic structures in the different contexts (instead of structure sharing or pattern matching). Of particular interest is their use of "embedded" (ie, nested) beliefs to represent recursive beliefs (the special case of nesting where a lower level context models a higher level one, as in the system's beliefs about John's beliefs about the system's beliefs) and mutual beliefs (by linking the context for one agent embedded in the context for another with the embedding context).

Formal Heuristic Theories

Moore. One of the most influential of the formal theories (both epistemological and heuristic) has been that of Moore (1977, 1980, 1981). His was the first AI theory to offer both a representational scheme and a logic and to show how they can interact with other notions to reason about action. For his representation, Moore uses a first-order axiomatization of the possible-worlds semantics of Hintikka's **S4** [rather than the modal axiomatic version; it should be noted that Moore (1977) erroneously added the **S5** rule]. Specifically, he introduces a predicate $T(w, p)$ to represent that the object language formula p is true in possible world w, and the predicate $K(A, w1, w2)$ to represent that $w2$ is possible according to what A knows in $w1$. "A knows that p" is then represented by $Know(A, p)$, which satisfies the axiom: $T(w1, Know(a1, p1)) \equiv \forall w2(K(a1, w1, w2) \rightarrow T(w2, p1))$. Since Moore is concerned with using knowledge to reason about actions, he formulates a logic of actions, where complex actions are built out of sequences, conditionals (defined in terms of *Know*), and loops, and a logic for "can," understood as "knowing how to do." The criticisms one can offer of Moore's work are both two-sided: (*1*) its psychological inadequacy (primarily due to his reliance on Hintikka's system)—but, of course, this is shared by most other formal theories—and (*2*) its similarity to much work that had been going on in philosophy during the 1960s and 1970s, but here it must be noted that one advantage of (some) AI theories over (some) philosophical theories is the former's attention to detail, which can often indicate crucial gaps in the latter. (Moore's critique of the database approach is discussed below.) Moore's line of research has been extended, most recently, by Morgenstern (1986).

Konolige. Konolige and Nilsson (1980) consider, from a formal point of view, a planning system involving cooper-

ating agents. Each agent is represented by a first-order language, a "simulation structure" (a partial model of the language), a set of facts (expressed in the language and including descriptions of other agents), a "goal structure" (consisting of goals and plans), a deduction system, and a planning system. An agent uses a formal metalanguage to describe the languages of other agents and can use its representation of other agents (or itself–but not quasi-indexically) to reason by simulation about their plans and facts in order to take them into account when making its own plans. Belief, rather than knowledge, is taken as the appropriate cognitive attitude, to allow for the possibility of error [not allowed by axiom (A4), above], and "agent A0 believes that agent A1 believes that agent A0 is holding object B" is represented by FACT(A1, 'HOLDING(A0, B)') appearing in A0's FACT-list. Although an analog of axiom (A5) is taken as an axiom here, the analog of (A6) is not, since (*1*) their system allows different agents to have different deduction systems and (*2*) the deductive capabilities of the agents are considered to be limited.

This theory was made more rigorous in Konolige (1983, 1984). Here, a planning system with multiple agents has a "belief subsystem" consisting of (*1*) a list of "base" sentences (about a situation) expressed in a formal language with a modal belief operator and a Tarski-like truth value semantics; (*2*) a set of deduction processes (or deduction rules) that are sound, effectively computable, have "bounded" input, and are, therefore, monotonic; and (*3*) a control strategy (for applying the rules to sentences). Belief derivation is "total"; that is, all queries are answered in a bounded amount of time. The system is deductively consistent (ie, a sentence and its negation are not simultaneously believed), but it is not logically consistent (ie, there might not be a possible world in which all beliefs are true). Thus, some measure of psychological plausibility is obtained. A system can be deductively though not logically consistent if there are resource limitations on deductions; that is, the deductive processes might be incomplete because of either weak rules or a control strategy that does not perform all deductions. Konolige uses the former (though his sample of a weak rule—*modus ponens* weakened by conjoining a "derivation depth" to each sentence—seems to require a nonstandard conjunction in order to prevent ordinary *modus ponens* from being derivable). The system satisfies two properties: *closure* (sentences derived in the system are closed under the deduction rules; ie, all deductions are made) and *recursion* (the belief operator $[S]$ is interpreted as another belief system). Thus, $[S]\alpha$ means that α is derivable in S's belief system. A "view" [similar to Hendrix's "vista" (1979)] is a belief system as "perceived through a chain of agents"; for example ν = *John, Sue* is John's perception of Sue's beliefs. To bound the recursive reasoning processes, the more deeply nested a system is, the weaker are its rules. Konolige presents a Gentzen-style propositional doxastic logic **B** consisting of: the axioms and rules of propositional logic; a set of rules for each view ν; and, for each ν, (*1*) a rule *Cut** (essentially *modus ponens*) that implements closure, (*2*) a rule B_5 that formalizes agent i's deductive system in view ν (roughly, the rule is that if a sentence δ from some set of sentences Δ can be inferred using the rules of the view ν, i from a set of sentences Γ that are believed by S_i, then $[S_i]\Delta$ can be inferred using the rules of ν from $[S_i]\Gamma$), and (*3*) a rule B_c that says that anything can be derived from logically inconsistent beliefs. **B** is stronger than might be desired, since, if the ν rules are complete and recursion is unbounded, **B** is equivalent to **S5** − (A4). Konolige points out, however, that it can be weakened to **S4** − (A4).

Levesque. A very different approach was taken by Levesque in a series of papers (1981, 1984a,b) on knowledge bases. The problem he confronts is that of treating a knowledge base that is incomplete (ie, that lacks some information needed to answer queries) as an abstract data type. However, his use of epistemic logic is not as a representation device within the knowledge base but as a query language. He defines a first-order language $\mathcal{L}$ that has its singular terms partitioned by means of a relation v into equivalence classes of coreferential terms; the classes are referred to by numerical "parameters" (for the knowledge base to be able to answer wh-questions). $\mathcal{L}$ has a truth value semantics based on a set s of "primitive" (true) sentences, and $\mathcal{L}$ is said to describe a "world structure" $\langle s, v\rangle$. Levesque argues that although $\mathcal{L}$ may be sufficient to query the knowledge base about the world, it is not sufficient to query it about itself. For this, $\mathcal{L}$ is extended to a language $\mathcal{KL}$, containing a knowledge operator $\boldsymbol{K}$ and satisfying two principles: (*1*) "every logical consequence of what is known is also known," but not everything is known (ie, the knowledge base is "an incomplete picture of a" possible world); and (*2*) "a pure sentence (ie, one that is about only the knowledge base) is true exactly when it is known" (ie, the knowledge base is an accurate picture of itself). The operator $\boldsymbol{K}$ satisfies slightly modified axioms for $\mathcal{L}$ (which are like those for a typical first-order logic), plus:

$$\text{If } \vdash_{\mathcal{L}}\alpha, \text{ then } \vdash_{\mathcal{KL}}\boldsymbol{K}\alpha.$$

$$\vdash_{\mathcal{KL}}((\boldsymbol{K}\alpha \wedge \boldsymbol{K}(\alpha \rightarrow \beta)) \rightarrow \boldsymbol{K}\beta).$$

$$\vdash_{\mathcal{KL}}(\forall x\boldsymbol{K}\alpha \rightarrow \boldsymbol{K}\forall x\alpha).$$

$$\text{If } \alpha \text{ is pure, then } \vdash_{\mathcal{KL}}(\alpha \equiv \boldsymbol{K}\alpha).$$

The first of these says, roughly, that if α is provable in $\mathcal{L}$, then "α is known" is provable in $\mathcal{KL}$; the second is similar to (A6); the third says, roughly, that if everything is such that α is known to hold of it, then it is known that everything is such that α holds of it; and the fourth says, roughly, that the $\boldsymbol{K}$ operator is redundant in pure sentences. Semantically, if k is a set of world structures (ie, those compatible with the knowledge base), then $\boldsymbol{K}\alpha$ is true on s, v, k, if and only if α is true on all $\langle s', v'\rangle$ in k. It should be observed that $\boldsymbol{K}$ is more like a *belief* operator since $\boldsymbol{K}\alpha \rightarrow \alpha$ is *not* a theorem, whereas $\boldsymbol{K}(\boldsymbol{K}\alpha \rightarrow \alpha)$ *is*. Two operations on an abstract data type KB can then be defined roughly as follows: (I) ASK: KB × $\mathcal{KL}$ → {yes, no, unknown}, where ASK = yes if $\boldsymbol{K}\alpha$ is true in KB; ASK = no if $\boldsymbol{K}\neg\alpha$ is true in KB; and ASK is unknown otherwise. (II) TELL: KB × $\mathcal{KL}$ → KB, where TELL = the intersection of KB with the set of all world structures on which the

query is true. Although the query language is epistemic, Levesque proves a representation theorem stating that the knowledge in KB is representable using $\mathcal{L}$[essentially by trading in $K\alpha$ for $\vdash_{\mathcal{L}}(k \rightarrow \alpha)$, where k may be thought of as the conjunction of sentences in KB].

In Levesque (1984b), principle 1 is weakened, for several psychologically interesting reasons: (a) it ignores resource limitations; (b) it requires belief of all valid sentences; (c) it ignores differences between logically equivalent, yet distinct, sentences; and (d) it requires belief of all sentences if inconsistent ones are believed. To achieve an interpretation sensitive to these, two belief operators are used: $B\alpha$ for "α is explicitly (or actively) believed" and $L\alpha$ for "α is implicit in what is believed." To distinguish (A) situations in which only α and $\alpha \rightarrow \beta$ are believed from (B) those in which they are believed together with β—without being forced to distinguish (C) situations in which only $\alpha \vee \beta$ is believed from (D) those in which only $\beta \vee \alpha$ is believed—Levesque uses "partial possible worlds," in which not all sentences get truth values. A formal logic is defined in which L is logically "omniscient" (much like Levesque's earlier K), but B is not. More precisely: (i) $B\alpha \rightarrow L\alpha$ is valid, but its converse is not; (ii) B is not closed under $\rightarrow$; (iii) B need not apply to all valid sentences or to both of two logically equivalent ones; and (iv) B allows inconsistent beliefs. Of great philosophical interest is a theorem that $B\alpha \rightarrow B\beta$ if and only if α *entails*β, where *entails* comes from relevance logic (Anderson and Belnap, 1975). Levesque has summarized his most recent work (1986a,b).

Other Theories. Most of the recent research on formal heuristic theories has been collected in the proceedings of the Conferences on Theoretical Aspects of Reasoning about Knowledge (eg, Halpern, 1986a) (cf Rapaport, 1988). One application of Kripke-style possible-worlds semantics for propositional epistemic logic for m agents is in the analysis of distributed systems (Halpern, 1986b). The abstract notion of a possible world can be interpreted as a global state of a distributed system (ie, as a description of each processor's state), and the accessibility relation for agent i can be interpreted as the relation between two global states s and t such that processor i has the same state in s and t. Thus, processor i "knows" proposition ϕ if and only if ϕ is true in all global states consistent with i's current state, where ϕ expresses information about processors' states or the values of their variables, for example. (Computational interpretations such as this of the abstract paraphernalia of possible-worlds semantics for modal logics are among the clearest, most revealing, and least metaphysically suspect.)

Another major topic is the problem of "logical omniscience"—that all agents "know all valid formulas and all logical consequences of their knowledge" (Halpern, 1986b, p. 7). There are three approaches to the solution of this problem. First, there is Kurt Konolige's syntactic approach, which employs incomplete sets of deduction rules (1986). Second, there is Levesque's semantic approach, discussed above. Finally, there is the combined syntactic-semantic approach of Ronald Fagin and Halpern's "logic of general awareness", which "adds to each state [of a Kripke structure] a set of formulas that the agent is 'aware' of at that state" (Halpern, 1986b, p. 8). On this view, implicit knowledge is the same as the standard epistemic-logic concept of knowledge, and an agent a explicitly knows ϕ if and only if a implicitly knows ϕ and ϕ is in a's awareness set. It is of some, perhaps sociological, interest that the most serious attention to the problem of logical omniscience has been paid, not by pure philosophers of mind or of language, but by computer scientists. [Cf also the work of Vardi (1986).]

Psychological Heuristic Theories

This category of research, which attempts to be more psychologically realistic than either of the preceding two, may be further subdivided along a spectrum ranging from the more formal to the more psychological.

More Formal than Psychological. There are two major, and related, topics investigated under this heading: speech act theory and mutual belief.

Speech Act Theory. Speech act theory, developed by the philosophers Austin, Grice, and Searle considers the basic unit of linguistic communication to be the rule-governed production of a token of a sentence (or word) in the performance of an illocutionary speech act (such as the act of making a statement or asking a question). According to Grice's version of this theory, meaning must be understood in terms of intending: a speaker S means something by his or her utterance U addressed to hearer H if and only if, roughly, S intended the utterance of U to produce a certain effect in H by means of the recognition of this intention (see references and further details in Searle, 1965).

Cohen and Perrault. Cohen and Perrault (1979) attempt to provide "a theory that formally models the possible intentions underlying speech acts . . . by treating intentions as plans" involving "the communication of beliefs." Plans are treated as prespecified sequences of "action" operators, which consist of preconditions, bodies, and effects and are evaluated relative to the planner's world model (including models of the planner's interlocutor's beliefs). When the action operator is a speech act, it takes beliefs and goals and returns plans for the appropriate speech act. Their criteria of adequacy for a theory of beliefs are that it must (*1*) distinguish agent AGT1's beliefs from AGT1's beliefs about AGT2's beliefs and (*2*) allow AGT1 to represent (a) that AGT2 knows whether P without AGT1 having to know which of P and $\neg$P AGT2 believes and (b) that AGT2 believes that R*ab* and that $\exists x$ R*ax* and that AGT2 knows what the x such that R*ax* is *without* AGT1 knowing what AGT2 thinks the x such that R*ax* is. Their logic of belief takes BELIEVE as a relation (though they call it a modal operator) between an agent and a proposition, satisfying the following axioms (for each agent a):

(B1) If P is an axiom of first-order logic, then $\vdash$aBELIEVE(P)

(B2) ⊢aBELIEVE(P) → aBELIEVE(aBELIEVE)(P))

(B3) ⊢aBELIEVE(P) ∨ aBELIEVE(Q) → aBELIEVE(P ∨ Q)

(B4) ⊢aBELIEVE(P & Q) → aBELIEVE(P) & aBELIEVE(Q)

(B5) ⊢aBELIEVE(P) → ¬aBELIEVE(¬P)

(B6) ⊢aBELIEVE(P → Q) → (aBELIEVE(P) → aBELIEVE(Q))

(B7) ⊢∃x[aBELIEVE(P(x))] → aBELIEVE(∃xP(x))

(B8) ⊢ All agents believe that all agents believe (B1)–(B7)

They admit that this is too strong to be psychologically plausible. Agents' wants are also represented but not axiomatized.

Cohen and Levesque. Cohen and Levesque (1980) claim that illocutionary act definitions can be derived from statements describing the recognition of shared plans and that this requires a definition of mutual beliefs. They offer perhaps the most honest, if not psychologically plausible, representation of belief:

(BEL *x p*) is true if and only if *p* follows from what *x* believes

(KNOW *x p*) is defined as (AND *p*(BEL *x p*)) and (KNOWIF *x p*) as (OR (KNOW *x p*)(KNOW *x* (NOT *p*))). The latter is used to define an if–then–else rule, along the lines of Moore (1977). Mutual belief (discussed in more detail below) is characterized by two axioms:

If ⊢*p*, then ⊢(MB *x y p*).

⊢(MB *x y p*) = (BEL *x* (AND *p* (MB *y x p*))).

A "plan" for an agent x to achieve goal q is defined as an action a and formulas $p_0, \ldots, p_k, q_0, \ldots, q_k = q$ such that (roughly) x believes that p_0 implies that the result of x doing a is q_0 and that p_i implies that x's making q_{i-1} true thereby makes q_i true (for $i = 1, \ldots, k$). Various illocutionary operators are characterized using notions such as these. This line of research has been extended by Cohen and Levesque (1990).

Allen and Perrault. This research program was continued by Allen and Perrault (1980) in order to model "helpful" linguistic behavior, that is, appropriate responses by a hearer (much in the manner of user modeling; see below). They offer a simple example (stated in the first person), which is presented here in more generality (in order to illustrate some of the complications that virtually all theories have ignored; compare the discussion of quasi-indicators, above): For S to inform H that he* (S) is tired, there must be two preconditions: that S believe that he* is tired and that he (S) intend that H believe that he* (S) is tired, and there should be the effect that H believe that S is tired. Their methodology is as follows: (*1*) There are planning rules; for example, if an agent wants to achieve P and does not know whether P is true, then the agent may want to achieve "agent knows whether P is true." (*2*) Figuring out another agent's plans depends on the observer's knowledge of planning and his or her beliefs about the agent's goals. (*3*) There are inference rules for inferring actions; for example corresponding to the planning rule above, if S believes that A has a goal of knowing whether P is true, then S may believe that A has a goal of achieving P or S may believe that A has a goal of achieving $\neg P$. Their logic of belief and knowledge is based on Hintikka (1962). For instance, there is an axiom schema of the form (though in different notation) $(B_A(P \rightarrow Q) \wedge B_A P) \rightarrow B_A Q$, although their commentary suggests that such schemata are really of the form $B_S(B_A(P \rightarrow Q) \wedge B_A P) \rightarrow B_S B_A Q$. Knowledge is defined as true belief: $K_A P = (P \wedge B_A P)$, interpreted as $B_S K_A P$ if and only if B_S(S and A agree that P). Knowing-whether and knowing-who are defined as follows:

$$\text{KNOWIF}_A P = (P \wedge B_A P) \vee (\neg P \wedge B_A \neg P).$$

$$\text{KNOWREF}_A P = \exists y[y = \text{the } x \text{ such that } D(x) \wedge B_A(y = \text{the } x \text{ such that } D(x))].$$

There are also numerous rules relating these forms of belief and knowledge to wants and actions.

Other theories include those of Allen, Sidner, and Israel. Allen (1984) continued this line of research, embedding it in a theory of action and time; here, BELIEVES(A, p, T_p, T_b) is taken to mean that A believes during time interval T_b that p holds during time interval T_p. Sidner and Israel (1981) and Sidner (1983) attack similar problems, treating the "intended meaning" of utterance U by speaker S for hearer H as a set of pairs of propositional attitudes (beliefs, wants, intentions, etc.) and propositional "contents" that are such that S wants H to hold the attitude toward the content by means of U.

Mutual Belief. The problems of mutual belief and mutual knowledge, notions generally accepted to be essential to research programs such as these, are most clearly stated by Clark and Marshall (1981). They raise a paradox of mutual knowledge: To answer a successful definite reference by speaker S to hearer H that term t refers to referent R, a doubly infinite sequence of conditions must be satisfied: $K_S(t \text{ is } R)$, $K_S K_H(t \text{ is } R)$, $K_S K_H K_S(t \text{ is } R)$, . . . , and $K_H(t \text{ is } R)$, $K_H K_S(t \text{ is } R)$, ... But each condition takes a finite amount of time to check, yet successful reference does not require an infinite time. Their solution is to replace the infinite sequences by mutual knowledge defined in terms of "copresence": S and H mutually know that t is R if and only if there is a state of affairs G such that S and H have reason to believe that G holds, G indicates to them that they have such reason, and G indicates to them that t is R. Typically, G will be either (*1*) community membership (ie, shared world knowledge), for example, when t is a proper name; (*2*) physical copresence (ie, a shared environment), for example, where t is an indexical; or (*3*) linguistic copresence (ie, a shared discourse), for example, where t is anaphoric (see Perrault and Cohen (1981) for a critique).

Mutual knowledge has been further investigated by Appelt (1980, 1982) and Nadathur and Joshi (1983). Appelt's planning system is an intellectual descendant of the work of Allen, Cohen, Perrault, and Moore. It reasons about A's and B's mutual knowledge by reasoning about the knowledge of a (virtual) agent—the "kernel"—whose knowledge is characterized by the union of sets of possible worlds that are consistent with A's and B's knowledge. Nadathur and Joshi replace Clark and Marshall's (1981) requirement of mutual knowledge for successful reference by a weaker criterion: if S knows or believes that H knows or believes that t is R, and if there is no reason to doubt that this is mutual knowledge, then S conjectures that it is mutual knowledge. This is made precise by using Konolige's *KI*4 to formulate a sufficient condition for S's using t to refer to R.

Other Theories. Other formal psychological heuristic work has been done by Taylor and Whitehill (1981) on deception and by Airenti and co-workers (1982) on the interaction of belief with conceptual and episodic knowledge.

More Psychological than Formal

Wilks and Colleagues. The various logics of nested beliefs in general and of mutual beliefs in particular each face the threat of infinite nestings or combinatorial explosions of nestings. Wilks and Bien (1979, 1983) have attempted to deal with this threat by using what might be called psychological heuristics. Their work is based on Bien's (1975) approach of treating natural-language utterances as programs to be run in "multiple environments" (one of the earliest forms of belief spaces): a global environment would represent a person P, and local environments would represent P's models of his or her interlocutors. The choice of which environment within which to evaluate a speaker's utterance U depends on P's attitude toward the discourse: if P believes the speaker, then U would be evaluated in P's environment, else in P's environments for the speaker and hearer. Wilks and Bien use this technique to provide an algorithm for constructing nested beliefs, given the psychological reality of processing limitations. They offer two general strategies for creating environments: (*1*) "Presentation" strategies determine how deeply nested an environment should be to represent information about someone. The "minimal" presentation strategy, for simple cases, constructs a level only for the subject of the information but none for the speaker; the "standard" presentation strategy constructs levels for both speaker and subject; and "reflexive" presentation strategies construct more complex nestings. (*2*) "Insertional" strategies determine where to store the speaker's information about the subject; for example, the "scatter gun" insertion strategy would be to store it in all relevant environments. A local environment is represented as a list of statements indexed by their behavior and nested within a relatively global environment: $A^{\{B\}}$ represents A's beliefs about B, $A^{\{B\{C\}\}}$ represents A's beliefs about B's beliefs about C. Suppose a *USER* informs the *SYSTEM* about person A. To interpret the *USER*'s utterance, a nested environment within which to run it is constructed, only temporarily, as follows: $SYSTEM^{\{A\}}$ and $SYSTEM^{\{USER\}}$ are constructed, and the former is "pushed down into" the latter to produce $SYSTEM^{\{USER\{A\}\}}$. Pushing is done according to several heuristics: (*1*) "Contradiction" heuristics: The *SYSTEM*'s beliefs about the *USER*'s beliefs about A are assumed to be the *SYSTEM*'s beliefs about A unless there is explicit evidence to the contrary. (*2*) Pragmatic inference rules change some of the *SYSTEM*'s beliefs about A into the *SYSTEM*'s beliefs about A's beliefs about A. (*3*) "Relevance" heuristics: Those of the *SYSTEM*'s beliefs about the *USER*'s beliefs that explicitly mention or describe A become part of the *SYSTEM*'s beliefs about A. (*4*) "Percolation" heuristics: Beliefs in $SYSTEM^{\{USER\{A\}\}}$ that are not contradicted remain in $SYSTEM^{\{A\}}$ when the temporary nested environment is no longer needed for evaluation purposes. Thus, percolation seems to be a form of learning by means of trustworthiness, though there is no memory of the source of the new beliefs in $SYSTEM^{\{A\}}$ after percolation has occurred; that is, the *SYSTEM* changes its beliefs about A by merely contemplating its beliefs about the *USER*'s beliefs. Other difficulties concern "self-embedded" beliefs: In $SYSTEM^{\{SYSTEM\}}$, there are no beliefs that the *SYSTEM* has about the *SYSTEM* that are not its own beliefs, but surely a *SYSTEM* might believe things that it does not believe that it believes; and there are potential problems about quasi-indicators when $SYSTEM^{\{A\}}$ is pushed down into itself to produce $SYSTEM^{\{A\{A\}\}}$. Wilks has extended this line of research (Wilks, 1986; Wilks and co-workers, 1989).

Colby. Although the work of Wilks and Bien has a certain formality to it, they are not especially concerned with the explicit logic of a belief operator, an accessibility relation, or a formal logic. The lack of concern with such issues may be taken to be the mark of the more psychological approaches. The pioneers of this approach were Colby and Abelson and their co-workers.

Colby and Smith (1969) constructed an "artificial belief system," ABS_1. ABS_1 had three modes of operation: During "talktime" a user would input sentences, questions, or rules; these would be entered on lists for that user (perhaps like a belief space; but see below). If the input were a question, ABS_1 would either search the user's statement list for an answer (taking the most recent if there were more than one answer), or deduce an answer from the statement list by the rules, or else generate an answer from other users' lists. During "questiontime" ABS_1 would search the user's statement list for similarities and ask the user questions about possible rules; the user's replies would enable ABS_1 to formulate new rules. ABS_1 would also ask the user's help in categorizing concepts. During "thinktime" ABS_1 would infer new facts (assigned to a "self"-list) and compute "credibility" weightings for the facts, rules, and user.

It should be noted that beliefs in this system are merely statements on a user's list, which makes this approach seem very much like the database approach criticized by Moore (1977). Moore's objections are as follows: (*1*) If the system does not know which of two propositions p or q a user believes, then it must set up two databases for the user, one containing p and one containing q, leading to

combinatorial explosion. (2) The system cannot represent that the user does not believe that p, since neither of the two database alternatives—omitting p or listing $\neg p$—is an adequate representation. Although these are serious problems, Colby and Smith's ABS_1 seems not to have them. First, ABS_1 only reasons about explicit beliefs; thus, it would never have to represent the problematic cases. Of course, a more psychologically adequate system would have to. Second, ABS_1 does not appear to reason about the fact that a user believes a statement but only about the statement and ABS_1's source for its believing the statement.

In Colby (1973) a belief is characterized as an individual's judgment of acceptance, rejection, or suspended judgment toward a conceptual structure consisting of concepts—representations of objects in space and time, together with their properties—and their interrelations. A statement to the effect that A believes that p is treated dispositionally (if not actually behavioristically) as equivalent to a series of conditionals asserting what A would say under certain circumstances. More precisely, "U Believe$_E C, t$" if and only if experimenter E takes the linguistic reaction (ie, judgment of credibility) of language user U to an assertion conceptualized as C as an indicator of U's belief in C during time T. Thus, what is represented are the objects of a user's beliefs, not the fact that they are believed. Various psychologically interesting types of belief systems (here understood as sets of interacting beliefs)—neurotic, paranoid, and so on—can then be investigated by "simulating" them. The most famous such system is Colby's PARRY (1971, 1972), which has been the focus of much controversy [see Colby (1981) and Weizenbaum's (1974) critique].

Abelson. A similar research program has been conducted by Abelson (1973) and with his co-workers (Abelson and Reich, 1969). Underlying their work is a theory of "implicational molecules," that is, sets of sentences that "psychologically" (ie, pragmatically) imply each other; for example, a "purposive-action" molecule might consist of the sentence forms "person A does action X," "X causes outcome Y," and "A wants Y." The key to their use in a belief system is what Abelson and Reich consider a Gestalt-like tendency for a person who has such a molecule to infer any one of its members from the others. Thus, a computer simulation of a particular type of belief system can be constructed by identifying appropriate molecules, letting the system's beliefs be sentences connected in those molecules (together with other structures, such as Schank's "scripts") and then having the system understand or explicate input sentences in terms of its belief system. A model of a right-wing politician was constructed in this manner [see also the discussions of Colby's as well as Abelson's work in Boden (1977)].

User Models. An extended, database type of belief system is exemplified by user models such as those investigated by Rich (1979a,b). Here, instead of the system being a model of a mind, the system must construct a model of the user's mind, yet many of the techniques are similar in both cases. A user model consists of properties of the user ("facts") ranked in terms of importance and by degree of certainty (or confidence) together with their justifications. The facts come from explicit user input and inferences based on these, on "stereotypes" (so that only minimal explicit user input is needed), and on the basis of the user's behavior (so that the model is not merely the user's self-model). The user model is built dynamically during interaction with the user. For further discussion, see Kobsa and Wahlster (1988).

DISCUSSION AND CONCLUSIONS

If there is any criticism to be leveled at the wide variety of current research, it is that the formal systems have not been sufficiently informed by psychology (and, hence, behave more like logicians than like ordinary people), and the psychological theories have not been flexible enough to handle some of the logical subtleties (which ordinary people, perhaps with some instruction, are certainly capable of). What is needed is a robust system whose input–output performance (if not the intervening algorithms) is psychologically plausible but whose underlying logic is competent, if needed, to handle the important (if often ignored) formal subtleties.

In spite of radically differing approaches and terminology, it seems clear that AI research into belief systems shares common issues and goals. This can be brought out by discussing Abelson's (1979) characterization of a belief system. For Abelson, a "system" is a "network of interrelated concepts and propositions" and rules, with procedures for accessing and manipulating them. Such a system is a "belief system" if:

1. The system's elements are not consensual.

This can be taken, perhaps, either as a rejection of $\mathbf{B}p \rightarrow p$ or as Wilks and Bien's heuristics. By contrast, a "knowledge system" *would* be consensual. Abelson urges that 1 be exploited by AI belief systems even though it makes them nongeneralizable.

2. The system is concerned with existence questions about certain conceptual objects.

The need to have a logic of the intensional objects of belief may be seen as a version of 2, even though 1 and 2 make it difficult to deal with beliefs that *are* held in common.

3. The system includes representations of "alternative worlds."

This desideratum may be taken as covering the notions of possible worlds and of nested and mutual beliefs.

4. The system relies on evaluative and affective components.
5. The system includes episodic material.

A "knowledge system" would rely more on general knowledge and principles. Clearly, though, a full system would need both.

6. The system's boundaries are vague.
7. The system's elements are held with different degrees of certitude.

Although these criteria are psychologically oriented, many of them are also applicable to formal approaches. In particular, 1–3 and 7 are relevant to logical issues; 4–7 are relevant to psychological issues.

Indeed, except for the choice of underlying logic, most of the systems discussed here seem compatible, their differences arising from differences in aim and focus. For instance, Abelson and Reich's implicational molecules could be among the ν rules in Konolige's system. Note that the rules do not have to be "logical" if they do not need to be consistent; moreover, as mentioned earlier, there might not be any (psychologically plausible) logic of belief. As a consequence, a psychologically plausible belief system, whether "formal" or not, must be able to deal with incompatible beliefs. This could be done by a belief revision mechanism or by representational or reasoning techniques that prevent the system from becoming "aware" of its inconsistencies (with, of course, occasional exceptions, as in real life). It is, thus, the general schemes for representation and reasoning that seem most important and upon which, as a foundation, specific psychological heuristics may be built.

In this way, too, it may be possible to overcome the computational complexity that is inevitably introduced when the underlying inference package is made to be as powerful as envisaged by, say, Konolige or when the underlying representational scheme is made to be as complete as proposed by, say, Shapiro and colleagues (Maida and Shapiro, 1982; Rapaport, 1985, 1986; Rapaport and Shapiro, 1984; Shapiro and Rapaport, 1987). A psychologically adequate "shell" that would be efficient at handling ordinary situations could be built on top of a logically adequate "core" that was capable of overriding the shell if necessary for correct interpretation.

The trade-offs between psychological and logical adequacy that have been made in most current systems can, in principle, be overcome. (They have, after all, been overcome in those humans who study the logic of belief yet have not been hindered from interacting in ordinary conversational situations.) Whether it is more feasible to make a formally adequate system psychologically adequate or to "teach" a psychologically adequate system to be logically subtle remains an interesting research issue.

BIBLIOGRAPHY

R. P. Abelson, "The Structure of Belief Systems," in R. C. Schank and K. M. Colby, eds., *Computer Models of Thought and Language*, W. H. Freeman, San Francisco, Calif., 1973, pp. 287–339.

R. P. Abelson, "Differences Between Belief and Knowledge Systems," *Cogn. Sci.* **3**, 355–366 (1979).

R. P. Abelson and C. M. Reich, "Implicational Molecules: A Method for Extracting Meaning from Input Sentences," *Proc. of the First IJCAI*, Washington, D.C., 1969, pp. 641–647.

G. Airenti, B. G. Bara, and M. Colombetti, "Knowledge and Belief as Logical Levels of Representation," *Proc. Cogn. Sci. Soc.* **4**, 212–214 (1982).

J. F. Allen, "Towards a General Theory of Action and Time," *Artif. Intell.* **23**, 123–154 (1984).

J. F. Allen and C. R. Perrault, "Analyzing Intention in Utterances," *Artif. Intell.* **15**, 143–178 (1980).

A. R. Anderson and N. D. Belnap, Jr., *Entailment: The Logic of Relevance and Necessity*, Princeton University Press, Princeton, N.J., 1975.

D. E. Appelt, "A Planner for Reasoning about Knowledge and Action," *Proceedings of the First National Conference on AI, Stanford, Calif.*, 1980, pp. 131–133.

D. E. Appelt, "Planning Natural-Language Utterances," *Proceedings of the Second National Conference on AI*, Pittsburgh, Penn., 1982, pp. 59–62.

J. A. Barnden, "Intensions as Such: An Outline," *Proceedings of the Eighth IJCAI*, Karlsruhe, FRG, 1983, pp. 280–286.

J. A. Barnden, "Imputations and Explications: Representational Problems in Treatments of Propositional Attitudes," *Cogn. Sci.* **10**, 319–364 (1986).

J. A. Barnden, "Towards a Paradigm Shift in Belief Representation Methodology," *Journal of Experimental and Theoretical Artificial Intelligence* **2**, 133–161 (1989).

J. S. Bień, "Towards a Multiple Environments Model of Natural Language," *Proc. of the Fourth IJCAI*, Tbilisi, Georgia, 1975, pp. 379–382.

M. Boden, *Artificial Intelligence and Natural Man*, Basic Books, New York, 1977.

B. C. Bruce, *Belief Systems and Language Understanding*, BBN Report No. 2973, 1975.

H.-N. Castañeda, review of Hintikka, 1962, *J. Symbolic Logic* **29**, 132–134 (1964).

H.-N. Castañeda, "Indicators and Quasi-Indicators," *American Philosophical Quarterly* **4**, 85–100 (1967).

H.-N. Castañeda, "Thinking and the Structure of the World," *Philosophia* **4**, 3–40 (1974). Originally written in 1972; reprinted in 1975 in *Critica* **6**, 43–86 (1972).

H. H. Clark and C. R. Marshall, "Definite Reference and Mutual Knowledge," in A. Joshi, B. Webber, and I. Sag, eds., *Elements of Discourse Understanding*, Cambridge University Press, Cambridge, U.K., 1981, pp. 10–63.

P. R. Cohen and H. J. Levesque, "Speech Acts and the Recognition of Shared Plans," *CSCSI* **3**, 263–271, 1980.

P. R. Cohen and H. J. Levesque, "Intention is Choice with Commitment," *Artif. Intell.* **42**, 213–261 (1990).

P. R. Cohen and C. R. Perrault, "Elements of a Plan-based Theory of Speech Acts," *Cogn. Sci.* **3**, 177–212 (1979); reprinted in B. L. Webber and N. J. Nilsson, eds., *Readings in Artificial Intelligence*, Tioga, Palo Alto, Calif., 1981, pp. 478–495.

K. M. Colby, "Simulations of Belief Systems," in R. C. Schank and K. M. Colby, eds., *Computer Models of Thought and Language*, W. H. Freeman, San Francisco, Calif., 1973, pp. 251–286.

K. M. Colby, "Modeling a Paranoid Mind," *Behav. Brain Sci.* **4**, 515–560 (1981).

K. M. Colby, F. D. Hilf, S. Weber, and H. C. Kraemer, "Turing-like Indistinguishability Tests for the Validation of a Computer Simulation of Paranoid Processes," *Artif. Intell.* **3**, 199–221 (1972).

K. M. Colby and D. C. Smith, "Dialogues Between Humans and an Artificial Belief System," *Proceedings of the First IJCAI*, Washington, D.C., 1969, pp. 319–324.

K. M. Colby, S. Weber, and F. Dennis Hilf, "Artificial Paranoia," *Artif. Intell.* **2**, 1–25 (1971).

A. R. Covington and L. K. Schubert, "Organization of Modally

Embedded Propositions and of Dependent Concepts," *Proc. CSCSI*, **3**, 87–94 (1980).

L. G. Creary, "Propositional Attitudes: Fregean Representation and Simulative Reasoning," *Proceedings of the Sixth IJCAI*, Tokyo, 1979, pp. 176–181.

P. Edwards, ed., *Encyclopedia of Philosophy*, Macmillan and Free Press, New York, 1967.

J. H. Fetzer, "On Defining 'Knowledge'," *AI Mag.* **6**, 19 (Spring 1985).

R. E. Filman, J. Lamping, and F. S. Montalvo, "Meta-language and Meta-reasoning," *Proceedings of the Eighth IJCAI*, Karlsruhe, FRG, 1983, pp. 365–369.

G. Frege, "On Sense and Reference" (1892), translated by M. Black in P. Geach and M. Black, eds., *Translations from the Philosophical Writings of Gottlob Frege*, Basil Blackwell, Oxford, U.K., 1970, pp. 56–78.

E. L. Gettier, "Is Justified True Belief Knowledge?," *Analysis* **23**, 121–123 (1963); reprinted in A. P. Griffiths, ed., *Knowledge and Belief*, Oxford University Press, Oxford, 1967.

J. Y. Halpern, ed., *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the 1986 Conference, Monterey, Calif.*, Morgan-Kaufmann, Los Altos, Calif., 1986a.

J. Y. Halpern, "Reasoning About Knowledge: An Overview," in J. Y. Halpern, ed., *Theoretical Aspects of Reasoning About Knowledge*, Morgan-Kaufmann, Los Altos, Calif., 1986b, pp. 1–17.

J. Y. Halpern and D. A. McAllester, "Likelihood, Probability, and Knowledge," IBM Research Report RJ 4313 (47141), 1984; shorter version in *Proceedings of the Fourth National Conference on AI*, Austin, Texas, 1984, pp. 137–141.

G. G. Hendrix, "Encoding Knowledge in Partitioned Networks," in N. V. Findler, ed., *Associative Networks*, Academic Press, New York, pp. 51–92, 1979.

J. Hintikka, *Knowledge and Belief: An Introduction to the Logic of the Two Notions*, Cornell University Press, Ithaca, N.Y., 1962.

J. Hintikka, "Semantics for Propositional Attitudes," in J. W. Davis and co-workers, eds., *Philosophical Logic*, D. Reidel, Dordrecht, 1969, pp. 21–45, reprinted in Linsky, 1977, pp. 145–167.

A. Kobsa, "VIE-DPM: A User Model in a Natural-Language Dialogue System," in *Proceedings of the 8th German Workshop on Artificial Intelligence*, Berlin, 1984a.

A. Kobsa, "Three Steps in Constructing Mutual Belief Models from User Assertions," in *Proceedings of the 6th European Conference on Artificial Intelligence*, Pisa, Italy, 1984b.

A. Kobsa, "Generating a User Model from Wh-Questions in the VIE-LANG System," in *Proceedings of GLDV Meeting on Trends in Linguistischer Datenverarbeitung*, 1984c.

A. Kobsa and H. Trost, "Representing belief models in semantic networks," *Cybern. Sys. Res.* **2**, 753–757 (1984).

A. Kobsa and W. Wahlster, eds., "User Modeling," special issue, *Computational Linguistics* **14**(3), 1988.

K. Konolige, "Circumscriptive Ignorance," *Proceedings of the Second National Conference on AI*, Pittsburgh, Penn., 1982, pp. 202–204.

K. Konolige, "A Deductive Model of Belief," *Proceedings of the Eighth IJCAI*, Karlsruhe, FRG, 1983, pp. 377–381.

K. Konolige, *Belief and Incompleteness*, CSLI Report No. CSLI-84-4, Stanford University, 1984.

K. Konolige, "What Awareness Isn't: A Sentential View of Implicit and Explicit Belief," in J. Y. Halpern, ed., *Theoretical Aspects of Reasoning About Knowledge*, Morgan-Kaufmann, San Mateo, Calif., 1986, pp. 241–250.

K. Konolige and N. J. Nilsson, "Multiple-Agent Planning Systems," *Proceedings of the First National Conference on AI*, Stanford, Calif., 1980, pp. 138–144.

H. J. Levesque, "The Interaction with Incomplete Knowledge Bases: A Formal Treatment," *Proceedings of the Seventh IJCAI*, Vancouver, Brit. Col., 1981, pp. 240–245.

H. J. Levesque, "Foundations of a Functional Approach to Knowledge Representation," *Artif. Intell.* **23**, 155–212 (1984a).

H. J. Levesque, "A Logic of Implicit and Explicit Belief," *Proceedings of the Fourth National Conference on AI*, Austin, TX, 1984b, pp. 198–202.

H. J. Levesque, "Making Believers Out of Computers," *Artif. Intell.* **30**, 81–108 (1986a).

H. J. Levesque, "Knowledge Representation and Reasoning," *Annual Review of Computer Science* **1**, 255–287 (1986b).

L. Linsky, ed., *Reference and Modality*, Oxford University Press, Oxford, 1977, corrected edition.

A. S. Maida and S. C. Shapiro, "Intensional Concepts in Propositional Semantic Networks," *Cogn. Sci.* **6**, 291–330 (1982).

A. S. Maida, "Knowing Intensional Individuals, and Reasoning about Knowing Intensional Individuals," *Proceedings of the Eighth IJCAI*, Karlsruhe, FRG, 1983, pp. 382–384.

J. Martins and S. Shapiro, "A Model for Belief Revision," *Artif. Intell.* **35**, 25–79 (1988).

J. McCarthy, "Epistemological Problems of Artificial Intelligence," *Proceedings of the Fifth IJCAI*, Cambridge, Mass., 1977, pp. 1038–1044.

J. McCarthy, M. Sato, T. Hayashi, and S. Igarashi, *On the Model Theory of Knowledge*, Stanford Artificial Intelligence Laboratory Memo AIM-312, Stanford University, 1978.

J. McCarthy, "First-order Theories of Individual Concepts and Propositions," in J. E. Hayes, D. Michie, and L. I. Mikulich, eds., *Machine Intelligence*, Vol. 9, Ellis Horwood, Chichester, UK, pp. 129–147, 1979.

J. McCarthy, "Circumscription—A Form of Non-monotonic Reasoning," *Artif. Intell.* **13**, 27–39 (1980).

J. McCarthy and P. J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence," in B. Meltzer and D. Michie, eds., *Machine Intelligence*, Vol. 4, Edinburgh University Press, Edinburgh, pp. 463–502, 1969, reprinted in B. L. Webber and N. J. Nilsson, eds., *Readings in Artificial Intelligence*, Tioga, Palo Alto, Calif., 1981, pp. 431–450.

A. Meinong, "Über Gegenstandstheorie" (1904), in R. Haller, ed., *Alexius Meinong Gesamtausgabe*, Vol. 2, Akademische Druck- u. Verlagsanstalt, Graz, 1971, pp. 481–535. English translation ("The Theory of Objects") by I. Levi and co-workers, in R. M. Chisholm, ed., *Realism and the Background of Phenomenology*, Free Press, New York, 1960, pp. 76–116.

J. Moravcsik, "Comments on Partee's paper," in K. J. J. Hintikka, J. M. E. Moravcsik, and P. Suppes, eds., *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, D. Reidel, Dordrecht, 1973, pp. 349–369.

R. C. Moore, "D-SCRIPT: A Computational Theory of Descriptions," *Proceedings of the Third IJCAI*, Stanford, Calif., 1973, pp. 223–229.

R. C. Moore, *Reasoning about Knowledge and Action*, Technical Note No. 191, SRI International, Menlo Park, Calif., 1980.

R. C. Moore, "Reasoning about Knowledge and Action," *Proceedings of the Fifth IJCAI*, Cambridge, Mass., 1977, pp. 223–227; reprinted in B. L. Webber and N. J. Nilsson, eds., *Readings in*

Artificial Intelligence, Tioga, Palo Alto, Calif., pp. 473–477, 1981.

R. C. Moore, "Problems in Logical Form," *Proc. ACL* **19,** 117–124 (1981).

R. C. Moore and G. G. Hendrix, "Computational models of belief and the semantics of belief sentences," in S. Peters and E. Saarinen, eds., *Processes, Beliefs, and Questions: Essays on Formal Semantics of Natural Language and Natural Language Processing*, D. Reidel, Dordrecht, pp. 107–127, 1982.

L. Morgenstern, "A First Order Theory of Planning," in J. Y. Halpern, ed., *Theoretical Aspects of Reasoning about Knowledge*, Morgan-Kaufmann, Los Altos, Calif., 1986, pp. 99–114.

G. Nadathur and A. K. Joshi, "Mutual Beliefs in Conversational Systems: Their Role in Referring Expressions," *Proceedings of the Eighth IJCAI*, Karlsruhe, FRG, 1983, pp. 603–605.

T. D. Parsons, "Frege's Hierarchies of Indirect Senses and the Paradox of Analysis," in P. A. French and co-workers, eds., *Midwest Studies in Philosophy* **6,** 3–57 (1981).

B. H. Partee, "The Semantics of Belief-Sentences," in K. J. J. Hintikka, J. M. E. Moravcsik, and P. Suppes, eds., *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, D. Reidel, Dordrecht, 1973, pp. 309–336.

B. H. Partee, "Belief-Sentences and the Limits of Semantics," in S. Peters and E. Saarinen, eds., *Processes, Beliefs, and Questions: Essays on Formal Semantics of Natural Language and Natural Language Processing*, D. Reidel, Dordrecht, 1982, pp. 87–106.

C. R. Perrault and P. R. Cohen, "It's for Your Own Good: A Note on Inaccurate Reference," in A. Joshi, B. Webber, and I. Sag, eds., *Elements of Discourse Understanding*, Cambridge University Press, Cambridge, U.K., 1981, pp. 217–230.

W. J. Rapaport, "Meinongian Theories and a Russellian Paradox," *Noûs* **12,** 153–180 (1978); errata, **13,** 125 (1979).

W. J. Rapaport, "Meinongian Semantics for Propositional Semantic Networks," *Proc. ACL* **23,** 43–48 (1985).

W. J. Rapaport, "Logical Foundations for Belief Representation," *Cogn. Sci.* **10,** 371–422 (1986).

W. J. Rapaport, Review of J. Y. Halpern, ed., *Theoretical Aspects of Reasoning about Knowledge*, in *J. Symbolic Logic* **53,** 660–670 (1988).

W. J. Rapaport and S. C. Shapiro, "Quasi-indexical Reference in Propositional Semantic Networks," *Proceedings of COLING-84*, 1984, pp. 65–70.

E. Rich, "Building and Exploiting User Models," *Proceedings of the Sixth IJCAI*, Tokyo, 1979a, pp. 720–722.

E. Rich, "User Modeling via Stereotypes," *Cog. Sci.* **3,** 329–354 (1979b).

M. Sato, *A Study of Kripke-Type Models for Some Modal Logics by Gentzen's Sequential Method*, Kyoto University Research Institute for Mathematical Sciences, Kyoto, 1976.

P. F. Schneider, "Contexts in PSN," *Proc. CSCSI*, **3,** 71–78 (1980).

J. R. Searle, "What is a Speech Act?," in M. Black, ed., *Philosophy in America*, Allen and Unwin, London, 1965, pp. 221–239; reprinted in J. R. Searle (ed.), *The Philosophy of Language*, Oxford University Press, Oxford, 1971, pp. 39–53.

P. Sells, "Aspects of Logophoricity," *Linguistic Inquiry* **18,** 445–479 (1987).

S. C. Shapiro and W. J. Rapaport, "SNePS Considered as a Fully Intensional Propositional Semantic Network," in N. Cercone and G. McCalla, eds., *The Knowledge Frontier: Essays in the Representation of Knowledge*, Springer-Verlag, New York, 1987, pp. 262–315.

C. L. Sidner, "What the Speaker Means: The Recognition of Speakers' Plans in Discourse," in N. Cercone, ed., *Computational Linguistics*, Pergamon Press, Oxford, 1983, pp. 71–82.

C. L. Sidner and D. J. Israel, "Recognizing Intended Meaning and Speaker's Plans," *Proceedings of the Seventh IJCAI*, Vancouver, Brit. Col., 1981, pp. 203–208.

B. C. Smith, "Varieties of Self-Reference," in J. Y. Halpern, ed., *Theoretical Aspects of Reasoning about Knowledge*, Morgan Kaufmann, Los Altos, Calif., 1986, pp. 19–43.

S. Soulhi, "Representing Knowledge about Knowledge and Mutual Knowledge," *Proceedings of COLING-84*, 1984, pp. 194–199.

G. B. Taylor and S. B. Whitehill, "A belief representation for understanding deception," *Proceedings of the Seventh IJCAI*, Vancouver, Brit. Col., 1981, pp. 388–393.

M. Y. Vardi, "On Epistemic Logic and Logical Omniscience," in J. Y. Halpern, ed., *Theoretical Aspects of Reasoning about Knowledge*, Morgan Kaufmann, Los Altos, Calif., 1986, pp. 293–305.

J. Weizenbaum, "Automating psychotherapy," *ACM Forum*, **17,** 543 (1974); reprinted with replies, *CACM* **26,** 28 (1983).

J. M. Wiebe and W. J. Rapaport, "Representing *De Re* and *De Dicto* Belief Reports in Discourse and Narrative," *Proc. IEEE* **74,** 1405–1413 (1986).

Y. Wilks, "Default Reasoning and Self-Knowledge," *Proc. IEEE* **74,** 1399–1404 (1986).

Y. Wilks and J. Bien, "Speech Acts and Multiple Environments," *Proceedings of the Sixth IJCAI*, Tokyo, 1979, pp. 968–970.

Y. Wilks and J. Bien, "Beliefs, Points of View, and Multiple Environments," *Cogn. Sci.* **7,** 95–116 (1983).

Y. Wilks, A. Ballim, and E. Dietrich, "Pronouns in Mind: Quasi-Indexicals and the 'Language of Thought'," *Computers and Artificial Intelligence* **8,** 493–503 (1989).

M. Xiwen and G. Weide, "W-JS: A Modal Logic of Knowledge," *Proceedings of the Eighth IJCAI*, Karlsruhe, FRG, 1983, pp. 398–401.

W. J. RAPAPORT
SUNY Buffalo

BELIEF REVISION

The ability to reason about and adapt to a changing environment is an important aspect of intelligent behavior. Most computer programs constructed by researchers in AI maintain a model of their environment (external and/or internal environment) that is updated to reflect the perceived changes in the environment. One reason for model updating is the detection of contradictory information about the environment. The conventional approach to handling contradictions consists of changing the most recent decision made (*chronological backtracking*) (see BACKTRACKING). An alternative solution, *dependency-directed backtracking*, consists of changing not the last choice made, but an assumption that provoked the unexpected condition. This second approach generated a great deal of research in one area of AI, which became loosely called belief revision.

Belief revision is an area of AI research concerned with the issues of revising sets of beliefs when new information