Ann. Rev. Comput. Sci. 1987. 2: 147–86

NONMONOTONIC REASONING

Raymond Reiter[1]

Department of Computer Science, University of Toronto, Toronto, Ontario M5S 1A4, Canada

1. INTRODUCTION

If Artificial Intelligence (AI) researchers can agree on anything, it is that an intelligent artifact must be capable of reasoning about the world it inhabits. The artifact must possess various forms of knowledge and beliefs about its world, and must use this information to infer further information about that world in order to make decisions, plan and carry out actions, respond to other agents, etc. The technical problem for AI is to characterize the patterns of reasoning required of such an intelligent artifact, and to realize them computationally. There is a wide range of such reasoning patterns necessary for intelligent behavior. Among these are:

- Probabilistic reasoning (e.g. Bundy 1985; Nilsson 1986), in which probabilities are associated with different items of information. Reasoning requires, in part, computing appropriate probabilities for inferred information, based upon the probabilities of the information used to support the inference.
- Fuzzy reasoning (e.g. Zadeh 1981), designed to characterize vague concepts like "tall" or "old" and to assign degrees of vagueness to conclusions inferred using such concepts.
- Inductive reasoning (e.g. Michalski 1983), which is concerned with determining plausible generalizations from a finite number of observations.
- Deductive reasoning, the concern of mathematical logic, which characterizes, among other things, the axiomatic method in mathematics.

This is far from a complete enumeration of human reasoning patterns. The most recent addition to this list is nonmonotonic reasoning, the study

[1] Fellow of the Canadian Institute for Advanced Research.

8756–7016/87/1115–0147$02.00

of which appears to be unique to AI. In order to convey an intuitive sense of what this is all about, it is first necessary to consider what has come to be known in AI as the knowledge representation problem.

Because an agent must reason *about* something (its knowledge, beliefs), any consideration of the nature of reasoning requires a concomitant concern with how the agent represents its knowledge and beliefs. The stance adopted by AI research on nonmonotonic reasoning is in agreement with the dominant view in AI on knowledge representation; the "knowledge content" of a reasoning program ought to be represented by data structures interpretable as logical formulas of some kind. As Levesque (1986) puts it:

> For the structures to represent knowledge, it must be possible to interpret them *propositionally*, that is, as expressions in a language with a *truth theory*. We should be able to point to one of them and say what the world would have to be like for it to be true.

The province of nonmonotonic reasoning is the derivation of plausible (but not infallible) conclusions from a knowledge base viewed abstractly as a set of formulas in a suitable logic. Any such conclusion is understood to be tentative; it may have to be retracted after new information has been added to the knowledge base.

In what follows, I assume the reader is logically literate, at least with respect to the fundamental ideas of first-order logic (with a smattering of second-order) and the familiar modal logics of necessity (e.g. *S4* and *S5*).

2. MOTIVATION

Nonmonotonic reasoning is a particular kind of plausible reasoning. Virtually every example in AI that calls upon such reasoning fits the following pattern:

Normally, A holds.

Several paraphrases of this pattern are commonly accepted:

Typically, A is the case.
Assume A by default.

The remainder of this section is devoted to a number of examples of this pattern as it arises in various settings of special concern to AI. The ubiquity of this pattern is remarkable. Once one learns to look for it, one discovers it virtually everywhere.

2.1 *The Canonical Example*

The standard example in AI of a nonmonotonic reasoning pattern has to do with flying birds. The sentence "Birds fly" is not synonymous with "All birds fly" because there are exceptions. In fact, there are overwhelmingly many exceptions—ostriches, penguins, Peking ducks, tar-coated birds, fledglings, etc. etc.—a seemingly open-ended list. Nevertheless, if told only about a particular bird, say Tweety, without being told anything else about it, we would be justified in assuming that Tweety can fly, *without knowing that it is not one of the exceptional birds*. In other words, we treat Tweety as a *typical* or *normal* bird.

We can represent the sentence "Birds fly" by instances of our patterns of plausible reasoning:

"Normal, birds fly."
"Typically, birds fly."
"It x is a bird, then assume by default that x flies."

We can now see why these are *plausible* reasoning patterns. We wish to use them to conclude that Tweety can fly, but should we subsequently learn information to the contrary—say, that Tweety is a penguin—we would retract our earlier conclusion and conclude instead that Tweety cannot fly. Thus initially we *jumped to the conclusion* or made the *default assumption* that Tweety can fly. This, of course, is what makes our rule patterns plausible rather than deductive; they sanction assumptions rather than infallible conclusions.

Notice also that there is another possible paraphrase of our reasoning pattern. In the case of Tweety the bird we were prepared to assume that Tweety can fly provided we knew of no information to the contrary, namely that Tweety is a penguin or an ostrich or the Maltese Falcon or.... So one possible reading of our pattern of plausible reasoning is:

In the absence of information to the contrary, assume A.

What is problematic here (as it is for notions like "typically" and "normally") is defining precisely what one means by "absence of information to the contrary." A natural reading is something like "nothing is known that is inconsistent with the desired assumption A." As we shall see later, this consistency-based version of the pattern motivates several formal theories of nonmonotonic reasoning. We shall also see that other intuitions are possible, leading to formalisms that apparently have little to do with consistency.

2.2 *Databases*

In the theory of databases there is an explicit convention about the representation of negative information that appeals to a particular kind of

default assumption. To see why negative information poses a problem, consider the simple example of a database for an airline flight schedule representing flight numbers and the city pairs they connect. We certainly would not want to include in this database all flights and the city pairs they do *not* connect, which clearly would be an overwhelming amount of information. For example, Air Canada flight 103 does not connect London with Paris, or Toronto with Montreal, or Moose Jaw with Athens, or.... *There is far too much negative information to represent explicity*, and this will be true for any realistic database.

Instead of explicitly representing such negative information, databases *implicitly* do so by appealing to the so-called *closed world assumption* (Reiter 1978b), which states that all relevant positive information has been explicitly represented. If a positive fact is not explicitly present in the database, its negation is assumed to hold. For simple databases consisting of atomic facts only, e.g. relational databases, this approach to negative information is straightforward. In the case of deductive databases, however, the closed world assumption (CWA) is not so easy to formulate. It is no longer sufficient that a fact not be explicitly present in order to conjecture its negation; the fact may be *derivable*. So in general we need a closed world rule that, for the flight schedule example, looks something like:

> If f is a flight and c_1, c_2 are cities, then in the absence of information to the contrary, assume $\neg$CONNECT(f, c_1, c_2).

Here, by "absence of information to the contrary" we mean that CONNECT(f, c_1, c_2) is not derivable using the database as premises. As we shall see below, there are formal difficulties with this version of the CWA; but on an intuitive level the CWA conforms to the pattern of plausible reasoning we are considering in this section. When we consider various proposed formalizations for nonmonotonic reasoning, below, we shall return to the question of the CWA since it plays a dominant role in many approaches.

2.3 *Diagnosis from First Principles*

There are two basic approaches in the AI literature to diagnostic reasoning.

Under the first approach, which might be called the experiential approach, heuristic information plays a dominant role. The corresponding systems attempt to codify the rules of thumb, statistical intuitions, and past experience of human diagnosticians considered experts in some particular task domain. In particular, the structure or design of the object being diagnosed is only weakly represented, if at all. Successful diagnoses stem primarily from the codified experience of the human expert being

modeled rather than from detailed information about the object being diagnosed. This is the basis of so-called rule-based approaches to diagnosis, of which the MYCIN system (Buchanan & Shortliffe 1984) is a notable example.

Under the second approach, often called *diagnosis from first principles*, or *diagnosis from structure and behavior*, the only information at hand is a description of some system, say a physical device or setting of interest, together with an observation of that system's behavior. If this observation conflicts with intended system behavior, then the diagnostic problem is to determine which components could by malfunctioning account for the discrepancy between observed and correct system behavior. Since components can fail in various and often unpredictable ways, their normal or default behaviors should be described. These descriptions fit the pattern of plausible reasoning we have been considering. For example, an AND-gate in a digital circuit would have the description:

> Normally, an AND-gate's output is the Boolean *and* function of its inputs.

In a medical diagnostic setting, we might want the description:

> Normally, an adult human's heart rate is between 70 and 90 beats per minute.

In diagnosis, such component descriptions are used in the following way: We first assume that all of the system components are behaving normally. Suppose, however, the system behavior *predicted* by this assumption conflicts with (i.e. is inconsistent with) the *observed* system behavior. Thus some of the components we assume to be behaving normally must really be malfunctioning. By retracting enough of the original assumptions about correctly behaving components, we can remove the inconsistency between the predicted and observed behavior. The retracted components yield a diagnosis. This approach to diagnosis from first principles forms the basis for several diagnostic reasoning systems (de Kleer & Williams 1986; Genesereth 1985; Reiter 1987). Poole (1986) took a somewhat different but closely related approach.

2.4 *Prototypes, Natural Kinds, and Frames*

Nonmonotonic reasoning is intimately connected to the notion of prototypes in psychology (Rosch 1978) and natural kinds in philosophy (Putnam 1970). To see the connection, observe that both these notions concern concepts that cannot be defined via necessary and sufficient conditions. We cannot, for example, define the natural kind "bird" by writing something like

$$(\forall x)\ \text{BIRD}(x) \equiv \text{BIPED}(x)\ \&\ \text{FEATHERED}(x)\ \&\ \cdots$$

because we can always imagine a bird that lacks one or more of the defining properties, say a one-legged bird. The best we seem capable of doing is to describe one or more "typical" members of the concept, and to define the concept as the set of individuals that do not deviate too far from the typical member(s). This notion of a "typical" member of such a concept provides the link with nonmonotonic reasoning. The rest of this section elaborates on this link.

The concepts that concern us are those lacking necessary and sufficient defining conditions. Recall that N is said to be a *necessary condition* for a predicate P if the following formula holds:

$$(\forall x)P(x) \supset N(x).$$

S is said to be a *sufficient condition* for P if the following holds:

$$(\forall x)S(x) \supset P(x).$$

Finally, P possesses a *classical definition* if there are formulas $D_1, \ldots, D_n$ that are both necessary and sufficient for P—i.e. if the following holds:

$$(\forall x)P(x) \equiv D_1(x) \,\&\, \cdots \,\&\, D_n(x).$$

As we have seen, commonsense concepts like "bird," "chair," "game," etc are not like mathematical concepts; they lack classical definitions based on necessary and sufficient conditions. Nevertheless, by appealing to conventional logic together with our pattern of plausible reasoning, we can define notions that correspond to normal necessary and sufficient conditions. For example, we have the following "necessary conditions" for the concept "bird":

If BIRD(x) then VERTEBRATE(x).
If BIRD(x) then normally FLY(x).
If BIRD(x) then assume by default BIPED(x).
If BIRD(x) then typically FEATHERED(x).
If BIRD(x) then typically HAS-AS-PART(x,beak(x)).
etc. 1.

It is natural to define a *prototypical bird* as one that enjoys all of the consequences, including the default assumptions, of the above "necessary conditions": It is a beaked, bipedal, feathered vertebrate that flies, etc.

The bird concept also possesses "sufficient conditions," some of which are logical implications while others fit our pattern for default reasoning:

If SPARROW(x) then BIRD(x).
If FLY(x) & CHIRP(x) then assume by default that BIRD(x).
If FLY(x) & FEATHERED(x) then assume by default that BIRD(x).
etc. 2.

It is natural, then, to take the concept of a bird to be defined by the above "necessary and sufficient conditions."

Now the obvious problem for AI knowledge representation is this: How do we characterize, represent, and compute with prototypes, or concepts like natural kinds, where default assumptions play such a prominent role? In his very influential "frames paper," Minsky (1975) proposed the notion of a frame, a complex data structure meant to represent certain stereotyped information. While Minsky's description of a frame is informal and often impressionistic, central to his notion are the issues we have just considered: prototypes, default assumptions, and the unsuitability of classical definitions for commonsense concepts like natural kinds. A few quotations from (Minsky 1975, p. 212) serve to illustrate this point.

> Here is the essence of the theory: When one encounters a new situation (or makes a substantial change in one's view of the present problem) one selects from memory a substantial structure called a frame. This is a remembered framework to be adapted to fit reality by changing details as necessary....
>
> A *frame* is a data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party....
>
> We can think of a frame as a network of nodes and relations. The "top levels" of a frame are fixed, and represent things that are always true about the supposed situation. The lower levels have many *terminals*—"slots" that must be filled by specific instances or data....
>
> Much of the phenomenological power of the theory hinges on the inclusion of expectations and other kinds of presumptions. A frame's terminals are normally already filled with "default" assignments. Thus, a frame may contain a great many details whose supposition is not specifically warranted by the situation.

Frames, therefore, are representations of stereotyped information. As Hayes (1979) points out, formally a frame has a logical status consisting of a collection of "necessary and sufficient" conditions on the concept defined by the frame. (Here, the quotation marks remind us that these conditions may be default assumptions.) Thus, a frame for the concept of a bird might contain bundle 1 above, of "necessary conditions" and bundle 2, of "sufficient conditions." What Minsky called the "top levels" of a frame, which represent things always true of the frame, are logical implications like the first formula of the bundle 1 or 2. The lower-level terminals or slots are predicates representing the default assumptions normally made of an instance of the frame. Thus FLY(.) and HAS-AS-PART(.,.) are slots of our BIRD(.) frame. The arguments of these slot predicates are the "fillers" in Minsky's description, so that if Tweety is an instance of the bird frame, i.e. BIRD(Tweety) holds, then the frame instance's terminals FLY(.), HAS-AS-PART(.,.), etc will be filled by Tweety, so that the

default assignments FLY(Tweety), and HAS-AS-PART(Tweety,beak (Tweety)) will be assumed.

We can now see that the "necessary and sufficient" conditions defining a frame play different roles.

"Necessary conditions" are used for frame instantiation. Given an instance, say BIRD(Tweety), of the BIRD(.) frame, we can infer some of Tweety's other properties, many of them default values. These are the expectations or presumptions referred to by Minsky, the "details whose supposition is not specifically warranted by the situation." Because some of these default assumptions may be specifically contradicted in certain cases, e.g. in the case of a bird that doesn't fly, not all the frame's terminals will be assumed. This corresponds to Minsky's assertion that "the default assumptions are attached loosely to their terminals, so that they can be easily displaced by new items that better fit the current situation."

"Sufficient conditions" are used for frame *selection* or *recognition*. Here recognition means determination of what kind of thing one might have in hand based upon knowledge of some of its properties. Of what frame might this thing be an instance? For example, the BIRD frame has as one of its sufficient conditions:

If CHIRP(x) and FLY(x) then assume by default BIRD(x).

If we have in hand something that we know chirps and flies, then we might select and instantiate the bird frame. This frame-selection or concept-recognition process is determined by some of the concept's sufficient conditions. These are normally taken to be *criterial*; chirping and flying are taken here to be criterial properties for BIRDness. The understanding that such properties do not guarantee the concept—it might be a flying cricket for example—is reflected in the default character of the sufficient condition.

3. THE NEED FOR A FORMAL THEORY

Having isolated a common pattern of reasoning, namely "Typically A holds," or "Assume A by default," we are still left with the problem of defining what this means. In addition, we shall need a theory of so-called *truth maintenance*. While an exploration of truth-maintenance systems is beyond the scope of this paper, it is important to note their intimate connection with the kinds of plausible reasoning considered thus far. Because our reasoning pattern sanctions default assumptions, some of these assumptions may have to be retracted in the light of new information. But these retracted assumptions might themselves have supported other

conclusions, which therefore also ought to be retracted, and so on. It is the job of truth-maintenance systems, in the style of Doyle's (1979), to manage this retraction process. One reason that truth-maintenance systems are as complex as they are is that default conclusions are normally based on two things: (*a*) *the presence*, either explicit or inferred, of certain information (e.g. the presence of the fact that Tweety is a bird), and (*b*) the *absence* of certain information, either explicit or inferred (e.g. the absence of ¬FLY(Tweety)). A truth-maintenance system must maintain a dependency record with each inferred fact indicating its justification in terms of both the presence and absence of information. This will obviously complicate both the system's bookkeeping and its process of belief revision whenever the knowledge base is modified.

One reason a formal account is required for default-reasoning is that the inferences they sanction can be complicated (Reiter & Criscuolo 1983). For example, two default assumptions can conflict, as the following example shows:

The typical Quaker is a pacifist.
The typical Republican is not a pacifist.

Suppose Dick is both a Quaker and a Republican. Then he inherits contradictory default assumptions, so that intuitively neither should be ascribed to him.

A second example illustrates that typicality is not necessarily transitive, in the sense tht "Typical *A*s are *C*s" need not follow from both "Typical *A*s are *B*s" and "Typical *B*s are *C*s." For if typicality were transitive, then from

"Typical high-school dropouts are adults"

and

"Typical adults are employed"

we could conclude the intuitively incorrect

"Typical high-school dropouts are employed."

As a final example of the complexities of reasoning about typicality, consider inheritance hierarchies, which form the backbone of almost all semantic networks and knowledge-representation languages. The classes in any such hierarchy are organized into a taxonomy via ISA links. These classes normally also have attributes. Now, suppose one wants to find out whether an individual in class *C* has attribute *A*. To do this, simply search from the node *C* up the hierarchy via ISA links to find if there is a higher node with attribute *A*. If so, then the individual inherits this attribute.

Unfortunately, this simple graphical processing fails when exceptions to attributes are allowed in the hierarchy. In a nice example of this, provided by Fahlman et al (1981), we have an exception to an exception to an exception:

A mollusc typically is a shell-bearer.
A cephalopod ISA mollusc except it typically is not a shell-bearer.
A nautilus ISA cephalopod except it typically is a shell-bearer.
A naked nautilus ISA nautilus except it typically is not a shell-bearer.

Here, the class mollusc has a default attribute shell-bearer. The class cephalopod has a default attribute non-shell-bearer, and so on. Now, suppose all we know of Fred is that he is a nautilus. Fred gets the default attribute shell-bearer by virtue of being a nautilus. But Fred is also a cephalopod via an ISA link, so at the same time he gets to be a non-shell-bearer by default. To deal with this kind of problem, most implementations adopt a shortest-path heuristic. A concept inherits the attribute nearest it in the hierarchy. Unfortunately, this can be shown to fail (Reiter & Criscuolo 1983), so other criteria are necessary. Any formal theory of default reasoning must allow us to sort out inheritance problems like this.

4. CLASSICAL LOGIC IS INADEQUATE

There are two arguments against classical logic for formalizing the reasoning patterns we have been considering. The first simply notes that even if we could enumerate all exceptions to flight with an axiom of the form

$$(\forall x)\ \mathrm{BIRD}(x) \,\&\, \neg\mathrm{EMU}(x) \,\&\, \neg\mathrm{DEAD}(x) \,\&\, \ldots \supset \mathrm{FLY}(x)$$

we still could not derive FLY(Tweety) from BIRD(Tweety) alone. This is so since we are not given that Tweety is not an emu, or dead, etc. The antecedent of the implication cannot be derived, in which case there is no way of deriving the consequent of the implication.

The second argument against classical logic is the so-called *monotonicity argument*. Classical logics share a common property of being monotonic. This means that whenever T is a set of sentences in such a logic and w is a sentence, then $T \models w$ implies $T \cup N \models w$ *for any set* N of sentences. In other words, new information N preserves old conclusions w.

Now suppose default reasoning could be represented in some classical logic, and T are axioms entailing that Tweety flies—i.e. $T \models \mathrm{FLY}(\mathrm{Tweety})$. If later we learn that Tweety is an ostrich, we want the enlarged axiom set not to entail that Tweety flies, i.e. we want

$$T \cup \{\mathrm{OSTRICH}(\mathrm{Tweety})\} \not\models \mathrm{FLY}(\mathrm{Tweety}).$$

But this is impossible in a classical logic. So whatever the logical mechanism that formalizes default reasoning, it must be *nonmonotonic*; its conclusions must be retractable or *defeasible*.

5. PROCEDURAL NONMONOTONICITY IN AI

AI researchers have routinely been implementing nonmonotonic reasoning systems for some time, usually without consciously focussing on the underlying reasoning patterns on which their programs rely. Typically these patterns are implemented using the so-called negation-as-failure mechanism, which occurs as an explicit operator in AI programming languages like PROLOG, or in rule-based systems. In PROLOG, for example, the goal *not G* succeeds iff *G* finitely fails. Since failing on *G* amounts to failing to find a proof of *G* using the PROLOG program as axioms, the *not* operator implements finite nonprovability. From this observation we can see that PROLOG's negation is a nonmonotonic operator. If *G* is nonprovable from some axioms, it needn't remain nonprovable from an enlarged axiom set.

Procedural negation is almost always identified with real—i.e. *logical*—negation. The way procedural negation is actually used in AI programs amounts to invoking the rule of inference "From failure of *G*, infer $\neg G$." This is really the closed world assumption, which we encountered earlier in the context of representing negative information in databases. Partly because it is a nonmonotonic operator, procedural negation can often be used to implement other forms of default reasoning. The following example, a PROLOG program for reasoning about flying birds, illustrates this.

fly (*X*) ← *bird* (*X*) & *not ab* (*X*).
bird (*X*) ← *emu* (*X*).
bird (*X*) ← *canary* (*X*).
ab (*X*) ← *emu* (*X*).
emu (*fred*).
canary (*tweety*).

Goal: *not fly* (*fred*) succeeds.
Goal: *fly* (*tweety*) succeeds.

Notice that the first rule uses a predicate *ab*, standing for abnormal. So this rule says that *X* flies if *X* is not an abnormal bird, in other words if *X* is a normal bird. The fourth rule describes a circumstance under which something is abnormal, namely when it is an emu. This device of the *ab* predicate for representing defaults is due to McCarthy, who introduced

it in conjunction with his circumscription formalism for nonmonotonic reasoning. We shall see it again in Section 6.3.1, where circumscription is described. Continuing with the current example, we see that by identifying procedural negation with real negation we can derive that the emu fred doesn't fly, while the bird tweety does.

For a nontrivial, formally precise application of procedural negation for reasoning about time and events see Kowalski & Sergot (1986).

6. SOME FORMALIZATIONS OF NONMONOTONIC REASONING

The need for nonmonotonic reasoning in AI had been recognized long before formal theories were proposed. In support of his argument against logic in AI, Minsky invoked the nonmonotonic nature of commonsense reasoning in one version of his 1975 "frames" paper (reprinted in Haugland 1981). Partial formalizations for such reasoning were proposed by McCarthy & Hayes (1969), Sandewall (1972), and Hayes (1973). Several knowledge-representation languages, most notably KRL (Bobrow & Winograd 1977), specifically provided for default reasoning. Hayes (1979) emphasized the central role of defaults in Minsky's notion of a frame and in KRL in particular. Reiter (1978a) described various settings in AI where default reasoning is prominent.

The rest of this section is devoted to a critical examination of several formalizations of nonmonotonic inferences.

6.1 *The Closed World Assumption*

As we remarked earlier, the closed world assumption (CWA) arises most prominently in the theory of databases, where it is assumed that all of the relevant positive information has been specified. Any positive fact not so specified is assumed false. In the case of deductive databases it is natural to understand that a positive fact has been specified if it is entailed by the database, and that any fact not so entailed is taken to be false. This is the intuition behind Reiter's (1978b) formalization of the CWA. Let DB be a first-order database (i.e. any first-order theory). Reiter defines the closure of DB by

$$\text{CLOSURE(DB)} = \text{DB} \cup \{\neg P(\mathbf{t}) | \text{DB} \nvdash P(\mathbf{t}) \text{ where } P \text{ is an } n\text{-ary predicate symbol of DB and } \mathbf{t} \text{ is an } n\text{-tuple of ground terms formed using the function symbols of DB}\}.^2$$

[2] In this paper (Reiter 1978b) the database is taken to be function-free, so that $\mathbf{t}$ is an n-tuple of constant symbols; but this restriction is unnecessary in general.

In other words, the implicit negative information of a database sanctioned by the CWA are those negative ground literals whose (positive) ground atoms are not entailed by the database. Under the CWA, queries are evaluated with respect to CLOSURE(DB), rather than DB itself.

There are several problems with this view of the CWA. The most obvious is that the database closure might be inconsistent, as would be the case for DB = $\{P \vee Q\}$. [In the case of Horn databases, Reiter (1978b) shows that closure preserves the consistency of DB.] Even for nondeductive relational databases consisting only of ground atoms, Reiter's notion yields incorrect results in the presence of so-called null values. A null value is a constant symbol meant to denote an existing individual whose identity is unknown. For example, if SUPPLIES(s, p) denotes that supplier s supplies part p, then the following is a simple database DB, where ω is meant to denote a null value:

SUPPLIES(Acme, p_1) SUPPLIES(Sears, p_2) SUPPLIES(ω, p_1)

So we know that some supplier, possibly the same as Acme or Sears, possibly not, supplies p_1. Since DB $\nvdash$ SUPPLIES(ω, p_2), Reiter's CWA sanctions $\neg$SUPPLIES(ω, p_2) which, coupled with SUPPLIES(Sears, p_2) entails $\omega \neq$ Sears. But this violates the intended interpretation of the null value ω as a totally unknown supplier; we have inferred *something* about ω, namely that it is not Sears.

A different formalization of the CWA was proposed by Clark (1978) in connection with his attempt to provide a formal semantics for negation in PROLOG. Clark begins with the observation that PROLOG clauses, being of the form $\alpha \supset P(\mathbf{t})$, provide sufficient but not necessary conditions on the predicate P. Such clauses are said to be *about* P. Clark's intuition is that the CWA is the assumption that these sufficient conditions are also necessary. In other words, the implicit information in a PROLOG database sanctioned by the CWA consists of the necessary conditions on all of the predicates of the database. Clark provides a simple effective procedure for transforming a set of clauses defining sufficient conditions on a predicate P into a single formula representing its necessary conditions. We illustrate this procedure with the following example:

$P(a, b)$ 3.

$P(a, c)$ 4.

$(\forall u, v, w) \neg Q(u, v) \;\&\; R(v, w) \;\&\; P(u, w) \supset P(g(u), w)$ 5.

$(\forall u)Q(u, f(u))$ 6.

Clauses 3–5 are the only ones in the database about P. These are logically equivalent, respectively, to

$$(\forall x, y)x = a \,\&\, y = b \supset P(x, y)$$

$$(\forall x, y)x = a \,\&\, y = c \supset P(x, y)$$

$$(\forall x, y)((\exists u, v, w)x = g(u) \,\&\, y = w \,\&\, \neg Q(u, v) \,\&\, R(v, w) \,\&\, P(u, w)) \supset P(x, y),$$

and these three clauses are in turn logically equivalent to

$$(\forall x, y)[(x = a \,\&\, y = b) \vee (x = a \,\&\, y = c) \vee ((\exists u, v, w)x = g(u) \,\&\, y = w \,\&\, \neg Q(u, v) \,\&\, R(v, w) \,\&\, P(u, w))] \supset P(x, y). \quad 7.$$

This is a single formula representing all the sufficient conditions on P given by the original database. Similarly, clause 6 is logically equivalent to

$$(\forall x, y)((\exists u)x = u \,\&\, y = f(u)) \supset Q(x, y), \quad 8.$$

and this represents Q's sufficient conditions. Finally, we must determine R's sufficient conditions. No clause of the database is about R, so we take R's sufficient conditions to be

$$(\forall x, y) \text{ false} \supset R(x, y). \quad 9.$$

Formulas 7, 8 and 9 are logically equivalent to the original database and represent that database's sufficient conditions on, respectively, the predicates P, Q, and R. To determine the implicit information about the predicates P, Q, and R sanctioned by Clark's CWA, assume that these sufficient conditions are also necessary—i.e. simply reverse the implications of formulas 7, 8, and 9. The resulting *completed* database represents the closure of the original database according to Clark. For the example at hand, the completed database is:

$$(\forall x, y)P(x, y) \equiv [(x = a \,\&\, y = b) \vee (x = a \,\&\, y = c) \vee ((\exists u, v, w)x = g(u) \,\&\, y = w \,\&\, \neg Q(u, v) \,\&\, R(v, w) \,\&\, P(u, w))]$$

$$(\forall x, y)Q(x, y) \equiv (\exists u)x = u \,\&\, y = f(u)$$

$$(\forall x, y)R(x, y) \equiv \text{false}.$$

On Clark's view of the CWA, queries are evaluated with respect to the completed database, rather than the original database.

As intuitively appealing as Clark's notion is, it suffers from a number of problems. To begin, it lacks generality. It is defined only for PROLOG-like databases and hence is restricted to universally quantified formulas.

Moreover, each clause must be about some predicate, so for example $\neg P$, which cannot be construed as being about P, cannot be accommodated. The approach is also sensitive to the syntactic form of the database clauses. Thus $\neg P \supset Q$ is about Q, while its logically equivalent form $\neg Q \supset P$ is about P. In particular, as Shepherdson (1984) observes, the completed database corresponding to $\neg P \supset P$ is $P \equiv \neg P$, which is inconsistent.

6.2 *Consistency-Based Approaches*

Some of the early attempts at formalizing nonmonotonic reasoning ground this notion in logical consistency. They interpret the pattern "In the absence of information to the contrary, assume A" as something like "If A can be consistently assumed, then assume it."

6.2.1 NONMONOTONIC LOGIC McDermott & Doyle's *nonmonotonic logic* (1980) appeals to a modal operator M in conjunction with the language of first-order logic. MA is intended to mean "A is consistent," so the flying birds example translates in their logic to

$$(\forall x)\ \text{BIRD}(x)\ \&\ M\ \text{FLY}(x) \supset \text{FLY}(x).$$

The technical problem is to make precise this notion of consistency, since we want consistency with respect to the entire knowledge base. But this means that a formula involving the M operator is in part referring to itself since as a formula it is part of the very knowledge base with respect to which it is claiming consistency. McDermott & Doyle capture this self-referential property by a fixed-point construction, and they define the theorems of a nonmonotonic knowledge base to be the intersection of all its fixed points. Specifically, if A is a nonmonotonic theory, then T is a *fixed point* of A if

$$T = Th(A \cup \{Mw \mid \neg w \notin T\}).^3$$

The intuition behind this definition is to capture the notion that if $\neg w$ is not derivable, then Mw (whose intended meaning is "w is consistent") is.

As a simple example, consider the nonmonotonic theory $A = \{E\ \&\ MC \supset \neg D,\ F\ \&\ MD \supset \neg C,\ E,\ E \supset F\}$. The first formula says that if E is the case and if C is consistent then conclude $\neg D$, so we do conclude $\neg D$. Now $\neg D$ prevents D being consistent in the second formula, so this blocks concluding $\neg C$ using the second formula. Thus one fixed point is obtained by adding $\neg D$ to A. Similarly, adding $\neg C$ to A gives a second fixed point. Thus, A has two fixed points:

$$Th(A \cup \{\neg D\})$$

[3] Here Th denotes closure under first-order logical consequence.

$Th(A \cup \{\neg C\})$.

The theorems of A are therefore the intersection of these two fixed points.

This formalism turns out to have several problems. Because of the consistency requirement, neither the fixed points nor the theorems are recursively enumerable. A proof theory is known only for the propositional case. There are also serious difficulties with the semantics; the M operator fails to adequately capture the intuitive concept of consistency. For example, the nonmonotonic theory $\{MC, \neg C\}$ is consistent.

In response to this latter difficulty, McDermott (1982a) attempted to develop several stronger versions of the logic based on the entailment relation of various standard modal logics (T, $S4$, and $S5$) instead of, as in the 1980 version, first-order logic. Unfortunately, these attempts turned out either to be too weak to adequately characterize the M operator (in the case of T and $S4$), or to "collapse" the logic to (monotonic) $S5$ when $S5$'s entailment relation was used.

6.2.2 DEFAULT LOGIC The other most prominent consistency-based approach to nonmonotonic reasoning is Reiter's (1980) default logic. It differs from the nonmonotonic logic of McDermott & Doyle in that default statements are formally treated as rules of inference, not as formulas in a theory. The flying birds default is represented by the rule of inference (actually a rule schema because of the variable x)

$$\frac{\text{BIRD}(x) : \text{FLY}(x)}{\text{FLY}(x)}.$$

This may be read as

> "If x is a bird and it can be consistently assumed to fly, then you can infer that x flies."

More generally, rule schemas of the following form are permitted:

$$\frac{\alpha(\mathbf{x}) : \beta(\mathbf{x})}{\gamma(\mathbf{x})}.$$

This can be read as

> "If $\alpha(\mathbf{x})$ holds and $\beta(\mathbf{x})$ can be consistently assumed, then you can infer $\gamma(\mathbf{x})$."

The approach is to begin with a set of first-order sentences. These are things known to be true of the world. This knowledge is normally incomplete; we are not omniscient, so there are gaps in our world knowledge. Default rules act as mappings from this incomplete theory to a more

complete *extension* of the theory. They partly fill in the gaps with plausible conclusions. So if such an incomplete first-order theory contains BIRD (Tweety), and if FLY(Tweety) is consistent with the theory, then by the above default schema for flying birds we can extend this theory by adding FLY(Tweety) to it.

As in McDermott & Doyle's approach, the extensions are defined by a fixed-point construction. For simplicity, we consider only *closed* default rules, namely rules of the form $\alpha : \beta/\gamma$ for first-order sentences α, β, and γ. A *default theory* is a pair (D, W) where D is a set of closed default rules and W a set of first-order sentences. For any set of first-order sentences S, define $\Gamma(S)$ to be the smallest set satisfying the following three properties:

1. $W \subset \Gamma(S)$.
2. $\Gamma(S)$ is closed under first-order logical consequence.
3. If $\alpha : \beta/\gamma$ is a default rule of D and $\alpha \in \Gamma(S)$ and $\neg\beta \notin S$, then $\gamma \in \Gamma(S)$.

Then E is defined to be an *extension* of the default theory (D, W) iff $\Gamma(E) = E$, i.e. iff E is a fixed point of the operator Γ.

The following example corresponds closely to that used to illustrate McDermott & Doyle's logic.

$$W = \{E, E \supset F\}$$

$$\text{Defaults: } \frac{E : C}{\neg D} \qquad \frac{F : D}{\neg C}.$$

Here E and $E \supset F$ are the two things we know about a world W. The first default can be invoked since C is consistent with W, so we infer $\neg D$. $\neg D$ prevents the second default from applying, so no further inferences are possible. This yields an extension $Th(W \cup \{\neg D\})$. A second (and only other) extension $Th(W \cup \{\neg C\})$ is obtained similarly.

As we have just seen, multiple extensions are possible. The perspective adopted on these (Reiter 1980) is that any such extension is a possible belief set for an agent, although one could, as do McDermott & Doyle, insist that an agent's beliefs are defined by the intersection of all extensions.

One advantage of default logic is that there is a "proof theory" in the case that all default rules are *normal*, namely, of the form

$$\frac{\alpha(\mathbf{x}) : \beta(\mathbf{x})}{\beta(\mathbf{x})}$$

for arbitrary first-order formulas α and β with free variables $\mathbf{x}$. This turns out to be an extremely common default pattern; all of the examples of Section 2 conform to it. The sense in which normal defaults have a "proof theory" is the following: Given a set of first-order sentences W, a set of

normal defaults D, and a first-order sentence β, then β is in some extension of W wrt the defaults D iff the "proof theory" sanctions this. The quotation marks indicate that in general the consistency condition prevents the default rules from being effectively computable. So one problem with default logic is that its extensions are not recursively enumerable. Another is that as yet there is no consensus on its semantics (see Etherington 1987; Sandewall 1985; Shoham 1986). Moreover, because the defaults are represented as inference rules rather than object language formulas as in McDermott & Doyle (1980), defaults cannot be reasoned about within the logic. For example, from "Normally canaries are yellow" and "Yellow things are never green" we cannot conclude "Normally canaries are never green." Notice that whether McDermott & Doyle's nonmonotonic logic can support such reasoning is debatable. From

$$(\forall x)\ \mathrm{CANARY}(x)\ \&\ M\ \mathrm{YELLOW}(x) \supset \mathrm{YELLOW}(x)$$

$$(\forall x)\ \mathrm{YELLOW}(x) \supset \neg \mathrm{GREEN}(x)$$

we can indeed infer

$$(\forall x)\ \mathrm{CANARY}(x)\ \&\ M\ \mathrm{YELLOW}(x) \supset \neg \mathrm{GREEN}(x).$$

However, it is unclear whether this last formula can legitimately be interpreted to mean "Normally canaries are not green."

Despite these shortcomings of default logic, analyses using the logic have been applied to several settings in AI: Inheritance hierarchies with exceptions, as described in Section 3 (Etherington & Reiter 1983), presuppositions in natural language (Mercer & Reiter 1982), Diagnostic reasoning (Poole 1986; Reiter 1987), and the theory of speech acts (Perrault 1987).

Etherington (1986) provides a number of properties of default logic, together with various results on its relationship to other nonmonotonic formalisms. Lukaszewicz (1984) proposes a modification of default logic with several desirable properties.

6.3 *Approaches Based Upon Minimal Models*

A promising way of achieving nonmonotonicity is to treat as theorems those sentences true in all suitably distinguished models of a logical theory. Provided that enlarging the theory can lead to new distinguished models, then what was once a theorem may no longer remain so; it may be falsified by one of the new models. Approaches that adopt this perspective on nonmonotonicity require that these preferred models respect some minimality property.

6.3.1 CIRCUMSCRIPTION McCarthy (1980, 1986) has proposed basing

nonmonotonic reasoning on the notion of truth in all minimal models of a first-order theory.[4] Since his 1986 approach generalizes that of his 1980 paper, we shall focus on his more recent theory. The notion of minimality to which McCarthy appeals is as follows (Lifschitz 1985b):

Assume L is a first-order language. Suppose $\mathbf{P}$ and $\mathbf{Z}$ are tuples of distinct predicate symbols of L. For any two structures Σ_1 and Σ_2 for L, define $\Sigma_1 \leq^{\mathbf{P};\mathbf{Z}} \Sigma_2$ if

i. $\text{domain}(\Sigma_1) = \text{domain}(\Sigma_2)$;
ii. Σ_1 and Σ_2 interpret all function symbols and predicate symbols other than those of $\mathbf{P}$ and $\mathbf{Z}$ identically; and
iii. for each predicate symbol P of $\mathbf{P}$, P's extension in Σ_1 is a subset (not necessarily proper) of its extension in Σ_2.

Notice that the relation $\leq^{\mathbf{P};\mathbf{Z}}$ places no restrictions on how Σ_1 and Σ_2 interpret the predicates of $\mathbf{Z}$.

Suppose now that $A(\mathbf{P};\mathbf{Z})$ is a sentence of L that mentions the predicate symbols of $\mathbf{P}$ and $\mathbf{Z}$. $A(\mathbf{P};\mathbf{Z})$ may mention predicate symbols other than those of $\mathbf{P}$ and $\mathbf{Z}$. In McCarthy's circumscription theory, the distinguished models of interest are those models of $A(\mathbf{P};\mathbf{Z})$ that are minimal wrt $\leq^{\mathbf{P};\mathbf{Z}}$. The sentences true in all such minimal models are taken to be the nonmonotonic entailments of $A(\mathbf{P};\mathbf{Z})$ of interest.

The above focus on minimal models and their entailments is not the approach emphasized by McCarthy (1986). McCarthy actually focussed on a syntactic approach, as follows:[5]

The circumscription of $\mathbf{P}$ *in* $A(\mathbf{P};\mathbf{Z})$ *with variable* $\mathbf{Z}$ is defined to be the (second-order) sentence

$$A(\mathbf{P};\mathbf{Z}) \,\&\, [\forall \mathbf{P}', \mathbf{Z}')\, \neg[A(\mathbf{P}';\mathbf{Z}') \,\&\, \mathbf{P}' < \mathbf{P}]. \qquad 10.$$

Here, for predicates Q and R of the same arity, $Q < R$ is defined to be

$$(\forall \mathbf{x})(Q(\mathbf{x}) \supset R(\mathbf{x})) \,\&\, \neg(\forall \mathbf{x})(R(\mathbf{x}) \supset Q(\mathbf{x})).$$

If we define $Q \leq R$ to be the formula $(\forall \mathbf{x})Q(\mathbf{x}) \supset R(\mathbf{x})$, then $Q < R$ is logically equivalent to the formula $Q \leq R \,\&\, \neg(R \leq Q)$. When $(Q_1, \ldots, Q_n)$ and $(R_1, \ldots, R_n)$ are tuples of predicate symbols with correspondingly equal arities, $(Q_1, \ldots, Q_n) < (R_1, \ldots, R_n)$ is defined to be the formula

$$Q_1 \leq R_1 \,\&\, \cdots \,\&\, Q_n \leq R_n \,\&\, \neg[R_1 \leq Q_1 \,\&\, \cdots \,\&\, R_n \leq Q_n].$$

[4] McCarthy (1986) actually treats second-order theories. For simplicity of exposition, we shall restrict ourselves to first-order theories. The more general case is elaborated by Lifschitz (1985b, 1986a).

[5] We adopt here the equivalent formulation of Lifschitz (1985b).

The second conjunct in sentence 10 is called the *circumscription axiom of* $A(\mathbf{P}; \mathbf{Z})$. It says that the extensions in $A(\mathbf{P}; \mathbf{Z})$ of the predicates $\mathbf{P}$ cannot be made smaller, even when the $\mathbf{Z}$ predicates are allowed to vary; or more succinctly, $\mathbf{P}$ is minimal in A with $\mathbf{Z}$ varying. Sentence 10 thus expresses the original sentence A further constrained by the requirement that $\mathbf{P}$ be minimized with $\mathbf{Z}$ variable.

In McCarthy's formulation, the nonmonotonic consequences of $A(\mathbf{P}; \mathbf{Z})$ of interest are those sentences entailed by 10. Because of what the circumscription axiom actually says, it is not surprising that the semantic and syntactic accounts of circumscription coincide. In other words, as proved independently by Lifschitz (1985b) and Etherington (1986), the sentences true in all models of $A(\mathbf{P}; \mathbf{Z})$ minimal wrt $\leq^{\mathbf{P};\mathbf{Z}}$ are precisely the sentences entailed by 10.

The circumscription axiom has the character of a second-order induction axiom in mathematics. In fact, McCarthy (1980) shows that, when sentence A defines a fragment of number theory, the circumscription axiom reduces to conventional Peano induction on the natural numbers. In deriving entailments of sentence 10, the circumscription axiom is used precisely the way induction axioms are used to prove theorems in mathematics. Since the predicate variables $\mathbf{P}'$ and $\mathbf{Z}'$ are universally quantified, we can substitute for them arbitrary formulas (provided they have suitable numbers of free individual variables). The entailments of any such instantiated version of sentence 10 will be some of the consequences of 10 itself.

Because of the extreme generality of sentence 10 (for example, which predicates $\mathbf{P}$, $\mathbf{Z}$ of A do we focus on?), McCarthy (1986) proposes a uniform principle for representing knowledge by sentences A in order to capture the pattern "Normally, such and such is the case." His approach appeals to a distinguished unary predicate AB (or often several such predicates $AB_1, \ldots, AB_n$) standing for "abnormal." In circumscribing the sentence A, it is these unary predicates that are minimized. The following example illustrates this use of the AB predicates, together with how the circumscription axiom is used as an induction axiom for deriving consequences of sentence 10.

$$(\forall x)\ \text{THING}(x) \ \&\ \neg AB_1(x) \supset \neg \text{FLY}(x) \qquad 11.$$

$$(\forall x)\ \text{BIRD}(x) \supset \text{THING}(x) \ \&\ AB_1(x) \qquad 12.$$

$$(\forall x)\ \text{BIRD}(x) \ \&\ \neg AB_2(x) \supset \text{FLY}(x) \qquad 13.$$

$$(\forall x)\ \text{EMU}(x) \supset \text{BIRD}(x) \ \&\ \neg \text{FLY}(x) \qquad 14.$$

Formula 11 is intended to express that normal (i.e. not AB_1 normal) things don't fly. Thus $\neg AB_1$ restricts THINGs to being normal wrt not flying.

Formula 12 states that birds are abnormal things wrt not flying and 13 has the intent of describing birds that are normal wrt being able to fly. Finally, axiom 14 distinguishes a subclass of nonflying birds.

Denote the conjunction of sentences 11–14 by $A(AB_1, AB_2; \text{FLY}]$ so that we shall minimize AB_1 and AB_2 with FLY variable using the circumscription axiom for $A(AB_1, AB_2; \text{FLY})$. The point of minimizing AB_1 and AB_2 is to allow as few abnormal individuals as possible, namely those forced by the theory A to be abnormal. The circumscription axiom is:

$$(\forall AB_1', AB_2', \text{FLY}')\, \neg[A(AB_1', AB_2'; \text{FLY}') \,\&\, AB_1' \leq AB_1 \,\&\, AB_2' \leq AB_2 \,\&\, \neg(AB_1 \leq AB_1' \,\&\, AB_2 \leq AB_2')] \quad 15.$$

In this axiom, we have three universally quantified predicate variables AB_1', AB_2', and FLY′, so we can choose these to be any fixed predicates we like. Suppose we cunningly choose

$$AB_1'(x) \equiv \text{BIRD}(x)$$

$$AB_2'(x) \equiv \text{EMU}(x)$$

$$\text{FLY}'(x) \equiv \text{BIRD}(x) \,\&\, \neg\text{EMU}(x).$$

If we make this substitution for the universally quantified predicate variables of the circumscription axiom 15, then from this instance of 15 together with $A(AB_1, AB_2; \text{FLY})$ we can derive, in first-order logic alone, the following:[6]

$$(\forall x)AB_1(x) \equiv \text{BIRD}(x)$$

$$(\forall x)AB_2(x) \equiv \text{EMU}(x)$$

i.e., the only abnormal things wrt flightlessness are birds, and the only abnormal birds wrt flight are emus. From this it follows easily that

$$(\forall x)\ \text{THING}(x) \,\&\, \neg\text{BIRD}(x) \supset \neg\text{FLY}(x)$$

$$(\forall x)\ \text{BIRD}(x) \,\&\, \neg\text{EMU}(x) \supset \text{FLY}(x)$$

neither of which is entailed by the original (uncircumscribed) theory.

As one can see from the example, it is not obvious in general how to instantiate the circumscribed theory. Lifschitz (1985b) provides some results about computing circumscription for various interesting special cases. Another formal problem is that, because circumscribed theories are second order, their valid formulas are not in general recursively enumerable. Note that this is also the case for nonmonotonic and default logic.

[6] The derivation itself is straightforward but tedious so we omit the details.

In addition, it can happen that a satisfiable theory has an unsatisfiable circumscription, although this cannot be in the case of theories all of whose sentences are universal in their prenex normal form (Etherington et al 1985). Lifschitz (1986a) generalizes this result on when circumscription preserves satisfiability.

Of all the formalisms proposed for nonmonotonic reasoning, circumscription appears to be the richest. It is certainly the most amenable to mathematical analysis. As a result, its formal properties have been extensively studied. Some completeness results are known (Perlis & Minker 1986). Its relationship to Reiter's notion of the closed world assumption of Section 6.1 has been analyzed by Lifschitz (1985a) and Gelfond et al (1986). Reiter (1982) shows that for a certain class of first-order theories, Clark's notion of theory completion (Section 6.1) is a consequence of circumscribing the theory. Lifschitz (1985b) provides the same result for a different class of first-order theories. A modification of McCarthy's circumscription, called pointwise circumscription (Lifschitz 1986b), together with priority orderings on the predicates to be minimized (McCarthy 1986), has been used to provide a semantics for negation for a large class of PROLOG programs (Lifschitz 1986c). All of this suggests that circumscription is a rich formalism whose full potential is far from being realized.

Independently of McCarthy, Bossu & Siegel (1985) have provided a semantic account of nonmonotomic reasoning for a special class of minimal models of a first-order theory. In the notation introduced above, their notion of minimality turns out to be based on the ordering $\leq^{\mathbf{P};\{\}}$, where $\mathbf{P}$ is the set of all predicate symbols mentioned by the theory. In other words, they minimize all predicates, with no variable predicates. Their analysis is strictly semantic, which is to say they provide nothing corresponding to McCarthy's circumscription axiom. Most significantly, Bossu & Siegel provide a decision procedure for first-order theories and queries of a certain kind. More specifically, suppose

1. the only function symbols are constants (the normal state of affairs in database theory),
2. the prenex form of each formula of the theory is universally quantified and satisfies a further natural syntactic constraint (which turns out to be a reasonable assumption for a database), and
3. the prenex form of the query is universally quantified (a reasonable assumption for some but far from all database queries) and satisfies a further simple syntactic constraint.

Under these conditions it is decidable whether the query is true in all minimal models of the theory (and hence is circumscriptively entailed by

the theory). The decision procedure is based upon a particular resolution theorem-proving strategy.

Minker (1982) provides a closely related minimal model analysis of the closed world assumption for database theory.

6.3.2 MINIMALITY AND THE FRAME PROBLEM The frame problem (McCarthy & Hayes 1969) concerns the representation of those aspects of a dynamically changing world that remain invariant under state changes. For example, walking to your front door or starting your automobile will not change the colors of any objects in the world. In a first-order representation of such worlds, it is necessary to explicitly represent all of these invariants under all state changes by so-called *frame axioms*. Thus, to represent the fact that turning on a light switch does not alter the colors of objects requires, in the situational calculus of McCarthy & Hayes (1969), a frame axiom of the form

$$(\forall x, c, s, l)\ \mathrm{COLOR}(x, c, s) \supset \mathrm{COLOR}(x, c, \mathrm{result}(\text{turn-on}, l, s))$$

where s is a state variable, x an object, c a color, and l a light switch.

The problem is that in general a vast number of such axioms will be required; object colors also remain invariant when lights are switched off, when someone speaks, etc, so there is a major difficulty even articulating a complete set of frame axioms for a given world, not to mention the computational problems associated with deduction in the presence of so many axioms.

A solution to the frame problem is a representation of the world that provides correct conclusions to be drawn about the dynamics of that world without explicitly representing, or reasoning with, the frame axioms. One of the principal motivations for the study of nonmonotonic reasoning was the belief that it would provide a solution to the frame problem (McCarthy 1977; Reiter 1978a); we required some way of saying that in the absence of information to the contrary, a state-changing event preserves the truth of an assertion.

Hanks & McDermott (1986) have investigated various nonmonotonic proposals for solving the frame problem and conclude that the apparently natural approaches fail. Specifically, they consider the simple setting where, in initial state s_0, a person is alive, then a gun is loaded, some time passes, and then the gun is fired at the person. They ask whether the person's resulting death can be deduced nonmonotonically, i.e. without explicit use of frame axioms. The axiomatization used appeals to McCarthy's AB predicate. It also appeals to a binary predicate T (for true) where $T(f, s)$ denotes that fact f is true in world state s. Syntactically, facts are first-

order sentences and so are treated as terms. Their axioms for the shooting scenario are simple and seemingly natural:

$T(\text{alive}, s_0)$

$(\forall s)\, T(\text{loaded}, \text{result}(\text{load}, s))$

$(\forall s)\, T(\text{loaded}, s) \supset AB(\text{alive}, \text{shoot}, s) \,\&\, T(\text{dead}, \text{result}(\text{shoot}, s))$

$(\forall f, e, s)\, T(f, s) \,\&\, \neg AB(f, e, s) \supset T(f, \text{result}(e, s))$.

Here $AB(f, e, s)$ means that fact f is abnormal when event e occurs in world state s. The last axiom, intended to circumvent the need for frame axioms, says that normally a fact f, true in state s, will remain true in the state that results from event e occurring in state s.

Hanks & McDermott consider circumscribing the above axioms, minimizing AB with T varying and ask us to consider the following situations:

s_0, $s_1 = \text{result}(\text{load}, s_0)$, $s_2 = \text{result}(\text{wait}, s_1)$, $s_3 = \text{result}(\text{shoot}, s_2)$.

Intuitively, we want $T(\text{dead}, s_3)$ to be circumscriptively derivable. Somewhat surprisingly, it is not. The reason is that the circumscribed theory has two models minimal in AB. In one, $AB(\text{alive}, \text{shoot}, s_2)$ is the only true AB atom, and it is easy to see that $T(\text{dead}, s_3)$ is true in this model, as required. But there is another model minimal in AB, namely that in which $AB(\text{loaded}, \text{wait}, s_1)$ is the only true AB atom, and in this model, corresponding to the gun mysteriously being unloaded during the wait event, $T(\text{alive}, s_3)$ is true. It follows that $T(\text{dead}, s_3)$ is not circumscriptively derivable from the above theory. Hanks & McDermott also show that default logic leads to an analogous result, in the sense that the above axioms, together with the default rule schema

$$\frac{:\neg AB(f, e, s)}{\neg AB(f, e, s)}$$

has two extensions, one containing $T(\text{dead}, s_3)$, the other containing $T(\text{alive}, s_3)$.

One might argue that this failure to solve the frame problem stems from an inappropriate set of axioms. Indeed, Lifschitz (1986d) has proposed an axiomatization that circumscriptively does yield the correct conclusions. Others, e.g. Kowalski & Sergot (1986), have argued that time plays a distinguished role in the frame problem, and that any nonmonotonic approach must respect this special status of time. It is towards this perspective that we now turn.

By explicitly providing for time, we obtain a finer-grained representation of dynamically changing worlds than with the situational calculus. We

can, for example, represent overlapping events, event durations, etc (Allen 1984; Kowalski & Sergot 1986; McDermott 1982b). In such temporal representations the frame problem becomes the persistence problem—determining that a fact known to be true at time t remains true over a future time interval provided no event is known to occur during that time interval to change the fact's truth value. In the case of the shooting scenario, assuming discrete time, we have that at $t = 0$ the person is alive and the gun is loaded and at $t = 2$ the gun is fired.[7] The problem is to infer that at $t = 2$ the person is still alive and the gun still loaded, i.e. that the truth of the facts "alive" and "loaded" persists from $t = 0$ to $t = 2$. Intuitively, since we were not informed of an unloading event occurring at $t = 1$, we want to infer that at $t = 2$ the gun is still loaded. This, of course, must be a defeasible inference since it could have been the case that the gun was unloaded at $t = 1$.[8]

Kautz (1986) proposes a minimal model solution to the persistence problem, and shows that there is a second-order circumscription-like axiom corresponding to this semantics. Shoham (1986) adopts an *S5* modal logic for representing an agent's knowledge, and proposes a minimal knowledge semantics for the persistence problem. Kowalski & Sergot (1986) propose a PROLOG-based temporal calculus of events that addresses the nonmonotonic character of the persistence problem using PROLOG's negation-as-failure mechanism. This is currently perhaps the most sophisticated approach to the persistence problem and the representation of events. It suffers primarily from its reliance on negation-as-failure, whose semantics is far from clear, so that it is somewhat closer to an implementation than a specification.

Shoham (1986) speculates on foundations for nonmonotonic reasoning for general settings, not just the temporal domain. He argues two perspectives.

1. There should be a shift in emphasis away from syntactic characterizations [as in default and nonmonotonic logic, or autoepistemic logic (Section 6.4.1, below)] in favor of semantic ones. This means that, having first fixed upon a logical language (not necessarily first order) one next provides a semantics for this language appropriate to the intended entailment relation for the application in mind.[9]
2. This entailment relation will be defined in terms of truth in all those models of a given axiomatization minimal with respect to some

[7] Recall that in the scenario we wait some time before firing the gun.

[8] Recall that in Hanks & McDermott's situational calculus version, the undesired model was one in which the gun was mysteriously unloaded during the wait event.

[9] Such an approach to knowledge representation was earlier provided by Levesque (1984).

application dependent criterion. The ability to characterize such minimality criteria axiomatically (as is the case for example with a circumscription axiom in McCarthy's theory), while perhaps desirable, is not essential. In effect, on Shoham's view, an axiomatization of an application domain coupled with a characterization of its preferred minimal models is a sufficient specification of the required entailments.

In support of his conclusion that nonmonotonicity necessarily involves minimality of one kind or another, Shoham offers his own theory of temporal minimization, as well as McCarthy's minimal semantics of circumscription. In addition, he proposes a minimal model semantics for a modification of Reiter's default logic.

Shoham's thesis—that nonmonotonic reasoning can be identified with truth in minimal models of one kind or another—is attractive. It provides a unifying perspective. Moreover, it suggests a methodology with which one can approach novel applications by considering which notion of minimality is to be preferred. The considerable successes of different forms of circumscription is strong evidence in its favor. Nevertheless, the fact that so few applications have been thoroughly explored, coupled with the unexpected difficulty of the frame problem, should caution us against overly hasty generalizations when it comes to nonmonotonic reasoning.

6.4 *Epistemic Approaches*

A number of approaches to nonmonotonic reasoning appeal to logics of belief or knowledge. The intuitive idea behind these is that a possible paraphrase of our favorite "Typically, birds fly" is something like "If x is a bird and if you don't believe (know) that x cannot fly, then x can fly." Since the standard epistemic logics (*S4*, *S5* etc) are all monotonic, direct appeals to these cannot work. However, nonmonotonicity can be achieved by a logic that sanctions $\neg B\alpha$[10] whenever α is absent from an agent's belief set, a property possessed by none of the standard epistemic logics. Under these circumstances, if an agent's belief set contains BIRD(Tweety) together with the default sentence

$$(\forall x)\ \mathrm{BIRD}(x) \,\&\, \neg B \neg \mathrm{FLY}(x) \supset \mathrm{FLY}(x)$$

but not $\neg$FLY(Tweety), then the belief set will contain $\neg B\neg$FLY(Tweety) whence, by modus ponens, the belief set will contain FLY(Tweety).

This, then, is the basic intuition behind epistemic approaches to non-

[10] We use $B\alpha$ to denote that an agent believes α.

monotonicity. Notice that nonmonotonicity is achieved by virtue of endowing an agent with the ability to reflect on its own beliefs in order to infer sentences expressing what it doesn't believe. The sentences contained in such a belief set depend on the entire belief set and hence are *indexical.*

We now consider several proposals for nonmonotonic epistemic logics.

6.4.1 AUTOEPISTEMIC LOGIC In response to the semantic deficiencies of McDermott & Doyle's nonmonotonic logic, Moore (1984, 1985) provides a reconstruction of their logic based upon belief rather than consistency, which he calls autoepistemic logic. Recall that the former logic appeals to a modal operator M with consistency as its intended meaning. Autoepistemic logic invokes a dual operator B[11] corresponding (roughly) to $\neg M \neg$. Moore's is a propositional logic only with the usual formulas formed from a propositional language augmented with the modal operator B. Given some set of premises A, a set T of formulas is a *stable expansion of A* just in case

$$T = Th(A \cup \{Bw \mid w \in T\} \cup \{\neg Bw \mid w \notin T\}).^{12}$$

Notice that this is a fixed-point definition much like that of McDermott & Doyle. In fact, under the dual correspondence of B with $\neg M \neg$ Moore's definition of a stable expansion differs from the fixed points of McDermott & Doyle (Section 6.2.1) only by the inclusion of $\{Bw \mid w \in T\}$ in his fixed-point construction. This set provides for an agent's perfect positive introspection; if w is in its belief set, then it believes w so that Bw is also in its belief set. The second set in the definition provides for perfect negative introspection; if w is nót in an agent's belief set, the agent does not believe w.

Levesque (1987) generalizes Moore's notion of a stable expansion to the full first-order case (which includes quantification into modal contexts). He also provides a semantic account of stable expansions in terms of a second modal operator O, where Ow is read as "w is all that is believed." Levesque then goes on to characterize stable expansions as follows: Ow is true exactly when all the formulas that are believed form a stable expansion of $\{w\}$.

As observed by Konolige (1987), stable expansions have some undesirable properties. Konolige notes that there are two stable expansions of $\{Bp \supset p\}$, one containing $\neg Bp$ but not p, the other containing both Bp and p. The first expansion is intuitively appropriate; an agent whose only initial belief is $Bp \supset p$ has no grounds for entering p into her belief set and

[11] We use B here for belief. In his papers, Moore uses the symbol L.

[12] Here Th denotes closure under the entailment relation of propositional logic.

should therefore enter $\neg Bp$. The second expansion, containing both Bp and p, is intuitively unacceptable. It corresponds to an agent arbitrarily entering p, hence also Bp, into her belief set.

To eliminate this undesirable property of Moore's autoepistemic logic, Konolige proposes the notion of a strongly grounded expansion of a set of premises A.[13] For any set Σ of formulas of our modal propositional language, denote by Σ_0 those formulas of Σ with no occurrence of the modal operator B, i.e. Σ_0 is the purely propositional part of Σ. Call a stable expansion T of A *minimal* iff there is no stable expansion S of A such that S_0 is a proper subset of T_0. Finally, call a set of formulas a *strongly grounded expansion* of A iff it is a minimal stable expansion of A. Konolige (1987) proposes strongly grounded expansions "as candidates for ideal introspective belief sets, because they limit the assumptions an agent makes about the world." Notice that the premise set $\{Bp \supset p\}$, which was problematic under Moore's account, has just one strongly grounded expansion, namely, the intuitively appropriate expansion containing $\neg Bp$ but not p.

Konolige provides several characterizations of strongly grounded expansions of A, all appealing to fixed-point constructions. Perhaps the most interesting characterization is in terms of the modal logic *KU*45, which is axiomatic *S5*, with *S5*'s axiom schema $B\phi \supset \phi$ replaced by the weaker $B(B\phi \supset \phi)$. Denote *KU45*'s provability relation by $\vdash_{KU45}$. Konolige shows that T is a strongly grounded expansion of A iff T satisfies the fixed point equation

$$T = \{w \mid A \cup \{B\alpha \mid \alpha \in A\} \cup \{\neg B\alpha \mid \alpha \notin T_0\} \vdash_{KU45} w\}.$$

Suppose (D, W) is a default theory (Section 6.2.2). Define its *autoepistemic transform* to be

$$W \cup \left\{ B\alpha \,\&\, \neg B\neg\beta \supset \gamma \;\middle|\; \frac{\alpha : \beta}{\gamma} \in D \right\}.$$

Thus, the transform translates default rules to sentences of autoepistemic logic. Konolige proves that autoepistemic logic is at least as expressive as default logic in the following sense:

Let A be the autoepistemic transform of a default theory. Then E is an

[13] Konolige (1987) calls these "strongly grounded autoepistemic extensions of A." He also deals with a first-order modal language, generalizing Moore's (1984, 1985) propositional language, but without quantifying into modal contexts. Here I continue to use a propositional modal language since the differences are inessential when quantification into modal contexts is forbidden.

extension of this default theory iff $E = S_0$[14] for some strongly grounded expansion S of A.

The question remains whether autoepistemic logic is strictly more expressive than default logic. Is there a set A of sentences with a strongly grounded expansion S for which S_0 is not an extension of any default theory? Surprisingly, the answer is no; Konolige shows:

For any set A of sentences there is an effectively constructable default theory such that E is an extension of this theory iff $E = S_0$ for some strongly grounded expansion S of A.

The above two results yield the unexpected conclusion that there is an exact correspondence between the extensions of default logic and strongly grounded expansions of autoepistemic logic.

6.4.2 SELF-KNOWLEDGE AND IGNORANCE Levesque (1982, 1984) is concerned with the following question: What is an appropriate notion of knowledge that would endow with self-knowledge a database *KB* of information about a world? Levesque's concept of self-knowledge includes knowledge about lack of knowledge; not only should *KB* know the information (and the entailments thereof) it contains, it should also know that it doesn't know a fact when indeed that fact is unknown to it.

To simplify the discussion, we shall consider a knowledge language called KFOPCE by Levesque (1982) which, though elementary, is sufficient to convey how nonmonotonicity and default reasoning can be achieved. In a subsequent paper Levesque (1984) treats a much richer such language.

KFOPCE is a first-order modal language with equality and with a single modal operator K (for "know"), constructed in the usual way from a set of predicate and variable symbols and a countably infinite set of symbols called *parameters*. Predicate symbols take variables and parameters as their arguments. Parameters can be thought of as constants. Their distinguishing feature is that they are pairwise distinct and they define the domain over which quantifiers range, i.e. the parameters represent a single universal domain of discourse.

A database *KB* of information about a world is a first-order sentence, i.e. a sentence of KFOPCE with no occurrence of the K operator. We consider how Levesque defines the result of querying *KB* with a sentence of KFOPCE. This requires first specifying a semantics for KFOPCE. A *primitive sentence* (of KFOPCE) is any atom of the form $P(p_1, \ldots, p_n)$, where P is an n-ary predicate symbol and $p_1, \ldots, p_n$ are parameters. A *world structure* is any set of primitive sentences that includes $p = p$ for

[14] Recall that S_0 is the purely propositional part of S.

each parameter p, and that does not include $p_1 = p_2$ for different parameters p_1 and p_2. The effect of this requirement on the equality predicate is that semantically the parameters are all pairwise distinct. A world structure is understood to be a set of true atomic facts. A *structure* is any set of world structures. The truth value of a sentence of KFOPCE with respect to a world structure W and a structure Σ is defined as follows:

1. If p is a primitive sentence, p is true wrt W and Σ iff $p \in W$.
2. $\neg w$ is true wrt W and Σ iff w is false wrt W and Σ.
3. $w_1 \vee w_2$ is true wrt W and Σ iff w_1 or w_2 is true wrt W and Σ.
4. $(\forall x)w(x)$ is true wrt W and Σ iff for every parameter p, $w(p)$ is true wrt W and Σ.
5. Kw is true wrt W and Σ iff for every $S \in \Sigma$, w is true wrt S and Σ.

Notice that condition 4 implies that, insofar as KFOPCE is concerned, the parameters constitute a single universal domain of discourse. The parameters are used to identify the known individuals. Notice also that when f is a first-order sentence (so that condition 5 need never be invoked in the truth recursion for f) then the truth value of f wrt W and Σ is independent of Σ, and we can speak of the truth value of f wrt W alone.

Given this semantics, Levesque defines the result of querying KB with an arbitrary sentence of KFOPCE as follows:

> Let $M(KB)$ be the set of the world structures W for which KB is true wrt W. $M(KB)$ is thus the set of models of KB. The result of querying KB with a sentence k of KFOPCE is defined to be
> ASK(KB, k) = *yes* if for all $W \in M(KB)$ k is true wrt W and $M(KB)$.
> = *no* if for all $W \in M(KB)$ k is false wrt W and $M(KB)$.
> = *unknown* otherwise.

Notice that this is an *S5* semantics with $M(KB)$ the equivalence class of mutually accessible possible worlds. It is this semantics that justifies interpreting the modal operator K of KFOPCE as a knowledge operator.

As an example, suppose KB is the conjunction of the following formulas:

> ENROLLED(Bill, cs100)
> TEACH(Mary, cs100) $\vee$ TEACH(Susan, cs100)
> $(\exists x)$ TEACH(x, math100)

Here, Bill, Mary, cs100, ..., are among the parameters. The following are some sample queries, together with the answers sanctioned by the above definition:

1. Is anyone known to be enrolled in cs100?
   $\neg(\exists x)K$ ENROLLED(x, cs100): yes

2. Does anyone teach cs100?
   $(\exists x)$ TEACH(x, cs100): yes
3. Is anyone known to teach cs100?
   $(\exists x)K$ TEACH(x, cs100): no
4. Is anyone known to teach math100?
   $(\exists x)K$ TEACH(x, math100): no
5. Is there a course in which Bill is enrolled and in which he is not known to be enrolled?
   $(\exists x)$ ENROLLED(Bill, x) & $\neg K$ ENROLLED(Bill, x): unknown.

Notice that ASK is nonmonotonic. For example, updating *KB* with TEACH(Sam, math100) would change the answer to question 4 from no to yes.

Levesque provides a noneffective way, requiring only an oracle for first-order theoremhood, of determining the result of ASKing *KB* an arbitrary sentence of KFOPCE.

In order to represent defaults like flying birds Levesque proposes

$$(\forall x)\ \text{BIRD}(x)\ \&\ \neg K \neg \text{FLY}(x) \supset \text{FLY}(x). \qquad 16.$$

This creates a technical problem; we must be able to update *KB* with non-first-order formulas like this, which requires first specifying the semantics of such updates. Levesque provides such a semantics, whose details we omit here. He then shows how to (noneffectively) determine a first-order formula $|\alpha|_{KB}$ such that the result of updating *KB* with α is *KB* & $|\alpha_{KB}|$. Thus, updating *KB* with a default like statement 16 has the effect of conjoining with *KB* a certain first-order formula.

Levesque's approach to (nonmonotonically) querying a first-order database has several advantages. It is semantically precise and well motivated. It allows one to ASK a database about its states of knowledge (witness the above simple example of an educational database), thus providing a far more expressive query language than conventional approaches using first-order logic (Green 1969). Moreover, the ASK operator can be realized in terms of first-order theoremhood, albeit by appealing to an oracle.

On the other hand, Levesque's treatment of default reasoning is problematic. Because defaults like statement 16 are assimilated into *KB* as suitable first-order formulas, they lose their character as defaults and hence cannot be reasoned about within the logic. In this respect they are akin to the default rules of default logic (Section 6.2.2). Moreover, inconsistencies can arise when intuitively they should not. For example, using the default sentence 16 to update the following *KB* leads to an inconsistent database:

$$\text{BIRD(Tweety)} \quad \text{BIRD(Opus)} \quad \neg \text{FLY(Tweety)} \vee \neg \text{FLY(Opus)}.$$

Intuitively, this is so since *KB* does not know $\neg$FLY(Tweety), and it does

not know $\neg$FLY(Opus), so by sentence 16 it deduces both FLY(Tweety) and FLY(Opus). Most other formalisms for handling defaults—e.g. circumscription, nonmonotonic logic, and default logic—do not lead to inconsistencies like this.

Despite such problems, Levesque (1982) provides a variety of interesting ideas for representing and structuring default information, including a proposal that, in many respects, anticipates McCarthy's (1986) use of the *AB* predicate for representing typicality. In its simplest form, Levesque's proposal is to introduce the concept of a typical-*P*, written ∇P, understood to be a new predicate. Thus ∇BIRD denotes a typical bird, and we can write a first-order axiom

$$(\forall x)\ \nabla\text{BIRD}(x) \supset \text{FLY}(x).$$

Certain birds are not typical:

$$(\forall x)\ \text{OSTRICH}(x) \supset \neg\nabla\text{BIRD}(x).$$

Defaults now state conditions under which instances of typical-birds may be inferred.

$$(\forall x)\ \text{BIRD}(x) \,\&\, \neg K\neg\nabla\text{BIRD}(x) \supset \nabla\text{BIRD}(x).$$

Using such representations for typicality, Levesque (1982) shows how to structure these to deal with many problems involving interacting defaults (Reiter & Criscuolo 1983) like the Quaker-Republican and shell-bearing examples of Section 3.

There have been a few other theories of knowledge in which an agent's ability to introspect on his ignorance leads to nonmonotonicity. Halpern & Moses (1984) propose a propositional approach very like Moore's autoepistemic logic (Section 6.4.1) but based upon an agent's knowledge rather than (as in Moore's case) belief. Unfortunately, as Halpern & Moses observe, their formalism cannot accommodate default reasoning. Konolige (1982) proposes a multi-agent logic of knowledge grounded in the propositional modal logic *S4*. This achieves nonmonotonicity by means of a closed world rule of inference based upon *S4* nonprovability. Using this logic, Konolige solves the Wise Man Puzzle, which requires a wise man to reason about the states of knowledge of two other wise men. However, the logic does not allow an agent to conclude that he does not know some fact, and hence it cannot provide a theory for default reasoning.

6.5 *Conditional Logics*

A few recent attempts to formalize nonmonotonic reasoning have been based upon conditional logics, which have been studied by several philosophical logicians, e.g. Lewis (1973) and Stalnaker (1968).

We shall focus here on *subjunctive conditionals*, i.e. statements of the form "If A were the case, then B would be the case," which we denote by $A \Rightarrow B$. The classic example from the philosophical literature is "If a match were to be struck, then it would light," which intuitively we all take to be true. But we also take to be true that "If a wet match were to be struck, then it would not light," and there is nothing peculiar about these two statements in the presence of a wet match. This means that the subjunctive if–then, $\Rightarrow$, is not the same as $\supset$, material implication, for otherwise the match example would have the form $A \supset C$ and $A \,\&\, B \supset \neg C$ which, in the presence of $A \,\&\, B$, a wet match, leads to a contradiction.

Now all of this certainly feels nonmonotonic. We can rephrase our bird example by subjunctive conditionals like "If x were a bird then x would fly," whereas "If x were a featherless bird then x would not fly." It is this intuition that suggests appealing to a suitable logic of conditionals to formalize nonmonotonic reasoning.

Such logics do exist (e.g. Delgrande 1986). Typically, these are based upon a possible-worlds semantics in which the truth value of a conditional $A \Rightarrow B$ in a world depends on a subset of those worlds in which A is true. Conditional logics differ primarily in how these worlds-in-which-A-is-true are distinguished. Axiomatizations of conditional logics correspond to these differing semantics—e.g. Delgrande's (1986).

As Delgrande (1986) observes, one motivation for considering conditional logics is that they allow us to reason about typicality within the logic. For example, "Typical canaries are not green" should be derivable (see Section 6.2.2). The logic should mandate the inconsistency of "All ravens are birds" with "Typical ravens are not birds," provided some raven exists. Indeed, Delgrande's logic has these properties.

Unfortunately, for our purposes, these logics have a fatal flaw; they are monotonic. Moreover, they are extremely weak. For example, modus ponens cannot be a rule of inference for conditional statements. This is so since otherwise, in our wet match example, from $A \Rightarrow C$, $A \,\&\, B \Rightarrow \neg C$, and $A \,\&\, B$ we could derive both C and $\neg C$. This failure of modus ponens means that we cannot infer default conclusions. BIRD(Tweety) and $(\forall x)$ BIRD$(x) \Rightarrow$ FLY(x) does not entail FLY(Tweety) in any conditional logic.

Despite these shortcomings, a few researchers (Delgrande 1986; Ginsberg 1986; Nute 1984) have proposed basing nonmonotonic reasoning systems on such logics. In all cases, nonmonotonicity is achieved by pragmatic considerations affecting how the logic is used. Unfortunately, this destroys the principled semantics on which these logics were originally based, so it is unclear what the advantages are of pursuing this approach to nonmonotonic reasoning.

7. SOME OBJECTIONS

Formalisms for nonmonotonic reasoning, grounded as they are in more or less conventional logics, have often been criticized. The most common objection is that probability theory is more appropriate (e.g. Cheeseman 1985). Numerically inclined nonprobabilists argue in favor of fuzzy reasoning (Zadeh 1985) or likelihood reasoning (Rich 1983), etc. In effect, all such proposals identify statements like "Typically birds fly" with "Most birds fly." In other words, they identify prototypical properties with statistical properties. Now, in certain settings a statistical reading is warranted. Regardless of my concept of a prototypical bird, if I find myself lost and hungry in a remote part of the world, my design of a bird-catching trap will depend upon my observation of the frequency with which the local birds fly. But to appeal exclusively to a statistical reading for plausible inference is to misunderstand the intended purpose of nonmonotonic reasoning.

In a wide variety of settings, nonmonotonic reasoning is necessary precisely because the information associated with such settings requires that certain *conventions* be respected. Such conventions may be explicit, as in the closed world assumption for the representation of negative information in databases. More commonly, these conventions are implicit, as in various principles of cooperative communication of information where it is understood by all participants that the informant is conveying all of the relevant information. Any relevant item of information not so conveyed is justifiably (and nonmonotonically) assumed false. For example, if someone were to tell you that John has not stopped beating the rug, you would justifiably infer that John was beating the rug despite the fact that the original statement might be true precisely because John never was beating the rug to begin with.[15] The point is that if this were the case, your informant should have told you. Since she didn't, convention dictates the appropriateness of your conclusion, despite its defeasibility.

Pictures and diagrams provide another interesting example. There is a kind of closed world convention to the effect that if an entity is not depicted in a picture or diagram, then it is not present in the world or the device the diagram represents.

It would seem that with respect to such conventions, statistical reasoning has no role to play whatsoever. It is difficult to imagine, for example, what it could mean to assign a probability to the failure of a circuit diagram to

[15] In linguistics, the original statement is said to *presuppose* the conclusion that John was beating the rug. Presupposition is well known to involve defeasible inferences (Levinson 1983, Ch. 4).

depict a device's power supply, or what advantage there could possibly be in doing so. McCarthy (1980) makes a similar point in discussing the missionaries-and-cannibals problem; he observes that the situation described by the puzzle is so wildly implausible that it would be meaningless to try to assign a conditional probability to the proposition that the boat is not leaky. In this connection, notice that puzzle solving is perhaps the clearest example of how convention sanctions nonmonotonic reasoning independently of any probabilistic interpretation. In fact, the preceding discussion suggests that much of what passes for human commonsense reasoning may at heart be puzzle solving.

The above argument from convention does not address all objections to logically based formalizations of nonmonotonic reasoning. Many nonmonotonic inferences are *abductive* in nature, which is to say they provide plausible explanations for some state of affairs. In this setting, an explanation can be taken to be a set of formulas that, together with the available background knowledge, entails the given state of affairs. The problem, of course, is that not just any explanation will do; it must, in some sense, be a "best" explanation. An explanation might be judged "best" because it is simplest, most general, or most probable, or because it is the outcome of weighing explicit evidence pro and con, etc. No such criteria are embodied in any current formalisms for nonmonotonic reasoning.

Israel (1980) criticizes nonmonotonic formalisms on similar, though more general grounds. He objects to the centrality of deductive logic in these formalisms as a mechanism for justifying an agent's beliefs. For Israel, "a heuristic treatment [of nonmonotonic reasoning], that is a treatment in terms of rational epistemic policies, is not just the best we could hope for. It is the only thing that makes sense." Abductively reasoning to a best explanation would, in Israel's view, require rational epistemic policies that necessarily lie outside the province of nonmonotonic logics. McDermott (1986) levies a similar criticism (among others) but is pessimistic about the very existence, currently, of formal theories of such rational epistemic policies for abductive reasoning. Nevertheless, as he observes:

> This state of affairs does not stop us from writing medical diagnosis programs. But it does keep us from understanding them. There is no independent theory to appeal to that can justify the inferences a program makes.... these programs embody *tacit* theories of abduction; these theories would be the first nontrivial formal theories of abduction, if only one could make them explicit.

We shall pursue McDermott's example of diagnostic reasoning because it will allow us to draw an important distinction. This, in turn, will reveal a significant role for nonmonotonic logics in situations requiring Israel's rational epistemic policies.

The proper way of viewing diagnosis is as a process of theory formation (Poole 1986): What is the best theory that accounts for the given evidence? But if there is a best theory, there must be poor ones; so diagnostic reasoning really consists of two problems: (*a*) What is the space of possible theories that account for the given evidence? (*b*) What are the best theories in this space? Most rule-based diagnostic systems conflate these two questions, attempting to converge on a best theory (usually by statistical means) without explicitly accounting for the space of possible theories through which they are searching. However, once this distinction is made, the proper role of nonmonotonic logics in diagnosis is revealed: They can characterize the space of possible theories that explain the evidence. This is seen most clearly in papers by Poole (1986) and Reiter (1987). For example, Reiter shows that the space of possible theories is precisely the set of extensions of a suitable formalization in default logic (Section 6.2.2) of the diagnostic setting. Poole's characterization, while somewhat different, is also based on default logic. Other approaches to diagnosis that emphasize characterizing the space of all theories are given by de Kleer & Williams (1986) and Reggia et al (1985).

The second problem—choosing a best theory from the space of possible theories—is currently beyond the province of nonmonotonic logics. In this respect, Israel's criticism is correct. However, given the space of possible theories as provided by nonmonotonic logics, we can at least begin a principled study of the rational epistemic policies for theory selection that Israel rightly emphasizes. This is the approach of de Kleer & Williams (1986) and Peng & Reggia (1986), who provide probabilistic grounds for diagnostic theory preference. In a different setting Poole (1985) proposes a preference ordering on theories that favors the most specific theories.

In brief, a proper analysis of diagnostic reasoning, and more generally abductive reasoning, must address two distinct problems. The first—that of characterizing the space of possible explanatory theories—is an appropriate role for nonmonotonic logics. The second—that of determining theory preference—requires rational epistemic policies that appear to have little to do with current approaches to nonmonotonic reasoning.

8. CONCLUSIONS

Nonmonotonicity appears to be the rule, rather than the exception, in much of what passes for human commonsense reasoning. The formal study of such reasoning patterns and their applications has made impressive, and rapidly accelerating progress. Nevertheless, much remains to be done.

The unexpected complexity of the frame problem suggests that many more non-toy examples need to be thoroughly explored in order for us to

gain a deeper understanding of the essential nature of nonmonotonic reasoning. In this connection, note that most potential applications have barely been touched, if at all. Apart from those discussed in this paper, examples include implicatures and presuppositions in natural language, high-level vision, qualitative physics, and learning.

With the possible exception of PROLOG's negation-as-failure mechanism, we know almost nothing about reasonable ways to compute nonmonotonic inferences. Truth maintenance systems must be integrated components of nonmonotonic reasoners, yet we have no adequate formal account of such systems. All current nonmonotonic formalisms deal with single agent reasoners. However, it is clear that agents must frequently ascribe nonmonotonic inferences to other agents, for example in cooperative planning or speech acts.[16] Such multi-agent settings require appropriate formal theories, which currently we lack.

The ultimate quest, of course, is to discover a single theory embracing all the seemingly disparate settings in AI where nonmonotonic reasoning arises. Undoubtedly, there will be surprises en route, but AI will profit from the journey, in the process becoming much more the science we all wish it to be.

ACKNOWLEDGMENTS

Many thanks to David Etherington, Russ Greiner and Hector Levesque for providing valuable suggestions on improving an earlier draft of this paper. My thanks also to Teresa Miao for carefully and patiently preparing this manuscript.

This research was done with the financial support of the National Sciences and Engineering Research Council of Canada, under operating grant A9044.

[16] See Perrault (1987). Incidentally, the requirement that an agent must be able to ascribe default rules to another agent argues for an epistemic approach to nonmonotonic reasoning (Section 6.4). See Halpern & Moses (1985) for a (monotonic) multi-agent logic of knowledge.

Literature Cited

Allen, J. F. 1984. Towards a general theory of action and time. *Artif. Intell.* 23: 123–54

Bobrow, D. G., Winograd, T. 1977. An overview of KRL, a knowledge representation language. *Cognitive Sci.* 1: 3–46

Bossu, G., Siegel, P. 1985. Saturation, nonmonotonic reasoning and the closed-world assumption. *Artif. Intell.* 25: 13–63

Buchanan, B. G., Shortliffe, E. ed. 1984. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Reading, Mass: Addison-Wesley

Bundy, A. 1985. Incidence calculus: a mechanism for probabilistic reasoning. *J. Automat. Reason.* 1: 263–83

Cheeseman, P. 1985. In defense of prob-

ability. *Proc. Int. Joint. Conf. Artif. Intell.*, Los Angeles, pp. 1002–9
Clark, K. 1978. Negation as failure. See Gallaire & Minker 1978, pp. 293–322
de Kleer J., Williams, B. C. 1986. Reasoning about multiple faults. *Proc. Am. Assoc. Artif. Intell., Natl. Conf.*, Philadelphia, pp. 132–45
Delgrande, J. P. 1986. A first-order conditional logic for reasoning about prototypical properties. *Simon Fraser Univ., Dept. Comput. Sci. Tech. Rep., Burnaby.* 24 pp.
Doyle, J. 1979. A truth maintenance system. *Artif. Intell.* 12: 231–72
Etherington, D. W. 1986. *Reasoning with incomplete information: investigations of non-monotonic reasoning.* PhD thesis. Univ. British Columbia, Vancouver. 151 pp.
Etherington, D. W. 1987. A semantics for default logic. *Proc. Int. Joint. Conf. Artif. Intell. Milan.* In press
Etherington, D., Mercer, R., Reiter, R. 1985. On the adequacy of predicate circumscription for closed world reasoning. *Comput. Intell.* 1: 11–15
Etherington, D. W., Reiter, R. 1983. On inheritance hierarchies with exceptions. *Proc. Am. Assoc. Artif. Intell. Natl. Conf.*, Washington, pp. 104–8
Fahlman, S. E., Touretzky, D. S., van Roggen, W. 1981. Cancellation in a parallel semantic network. *Proc. Int. Joint Conf. Artif. Intell. 81*, Vancouver, pp. 257–63
Genesereth, M. R. 1984. The use of design descriptions in automated diagnosis. *Artif. Intell.* 24: 411–36
Gallaire, H., Minker, J., ed. 1978. *Logic and Data Bases.* New York/London: Plenum 458 pp.
Gelfond, M., Przymusinska, H., Przymusinska, T. 1986. The extended closed world assumption and its relation to parallel circumscription. Univ. Texas, Dept. Math. Sci. Work. Pap. El Paso
Ginsberg, M. L. 1986. Counterfactuals. *Artif. Intell.* 30: 35–79
Green, C. 1969. *The application of theorem-proving to question answering systems.* PhD thesis. Stanford Univ., Stanford
Halpern, J. Y., Moses, Y. 1984. Towards a theory of knowledge and ignorance: preliminary report. *Proc. AAAI Workshop Non-Monotonic Reason.*, New Paltz, pp. 125–43
Halpern, J. Y., Moses, Y. 1985. A guide to the modal logics of knowledge and belief: preliminary draft. *Proc. Int. Joint Conf. Artif. Intell.*, Los Angeles, 480–90
Hanks, S., McDermott, D. 1986. Default reasoning, nonmonotonic logics, and the frame problem. *Proc. Am. Assoc. Artif. Intell. Natl. Conf.*, Philadelphia, pp. 328–33
Haugland, J. 1981. *Mind Design.* Cambridge, Mass: MIT Press. 368 pp.
Hayes, P. J. 1973. The frame problem and related problems in artificial intelligence. In *Artificial and Human Thinking*, ed. A. Elithorn, D. Jones, pp. 45–59. San Francisco: Jossey-Bass
Hayes, P. J. 1979. The logic of frames. In *Frame Concepts and Text Understanding*, ed. D. Metzing, pp. 46–61, Berlin: Walter de Gruyter
Israel, D. J. 1980. What's wrong with nonmonotonic logic? *Proc. Am. Assoc. Artif. Intell. Natl. Conf.*, pp. 99–101
Kautz, H. A. 1986. The logic of persistence. *Proc. Am. Assoc. Artif. Intell. Natl. Conf.*, Philadelphia, pp. 401–5
Konolige, K. 1982. Circumscriptive ignorance. *Proc. Am. Assoc. Artif. Intell. Natl. Conf.*, Pittsburgh, pp. 202–4
Konolige, K. 1987. On the relation between default theories and autoepistemic logic. *SRI Int., Artif. Intell. Cent. Tech. Rep., Palo Alto*
Kowalski, R., Sergot, M. 1986. A logic-based calculus of events. *New Generation Comput.* 4: 67–95
Levesque, H. J. 1982. A formal treatment of incomplete knowledge bases. *Fairchild Lab. Artif. Intell. Res. Tech. Rep. 3*
Levesque, H. J. 1984. Foundations of a functional approach to knowledge representation. *Artif. Intell.* 23: 155–212
Levesque, H. J. 1986. Knowledge representation and reasoning. *Ann. Rev. Comput. Sci.* 1: 255–87
Levesque, H. J. 1987. All I know: preliminary report. *Univ. Toronto, Dept. Comput. Sci. Tech. Rep.*, Toronto
Levinson, S. C. 1983. *Pragmatics.* London: Cambridge Univ. Press. 420 pp.
Lewis, D. 1973. *Counterfactuals.* Cambridge Mass: Harvard Univ. Press. 150 pp.
Lifschitz, V. 1985a. Closed-world databases and circumscription. *Artif. Intell.* 27. 229–35
Lifschitz, V. 1985b. Computing circumscription. *Proc. Int. Joint Conf. Artif. Intell.*, Los Angeles, pp. 121–27
Lifschitz, V. 1986a. On the satisfiability of circumscription. *Artif. Intell.* 28: 17–27
Lifschitz, V. 1986b. Pointwise circumscription: a preliminary report. *Proc. Am. Assoc. Artif. Intell. Natl. Conf.*, Philadelphia, pp. 406–10
Lifschitz, V. 1986c. On the declarative semantics of logic programs with negation. *Stanford Univ., Comput. Sci. Dept. Tech. Rep.*, Stanford
Lifschitz, V. 1986d. Formal theories of

action. *Stanford Univ., Comput. Sci. Dept. Tech. Rep.*, Stanford
Lukaszewicz, W. 1984. Considerations on default logic. *Proc. AAAI Workshop Non-Monotonic Reason.*, New Paltz, pp. 165–93
McCarthy, J. 1977. Epistemological problems of artificial intelligence. *Proc. Int. Joint Conf. Artif. Intell.*, Cambridge, Mass., pp. 223–27
McCarthy, J. 1980. Circumscription—a form of non-monotonic reasoning. *Artif. Intell.* 13: 27–39
McCarthy, J. 1986. Applications of circumscription to formalizing commonsense knowledge. *Artif. Intell.* 28: 89–116
McCarthy, J., Hayes, P. J. 1969. Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence 4*, ed. B. Meltzer, D. Michie, pp. 463–502. Edinburgh: Edinburgh Univ. Press
McDermott, D. 1982a. Non-monotonic logic II: non-monotonic modal theories. *J. ACM* 29: 33–57
McDermott, D. V. 1982b. A temporal logic for reasoning about processes and plans. *Cognitive Sci.* 6: 101–55
McDermott, D. 1986. A critique of pure reason. *Yale Univ., Dept. Comput. Sci. Tech. Rep.*, New Haven
McDermott, D., Doyle, J. 1980. Non-monotonic logic I. *Artif. Intell.* 25: 41–72
Mercer, R. E., Reiter, R. 1982. The representation of presuppositions using defaults. *Proc. Can. Soc. Computat. Stud. Intell. Natl. Conf.*, Saskatoon, pp. 103–7
Michalski, R. S. 1983. A theory and methodology of inductive learning. In *Machine Learning*, ed. R. S: Michalski, J. G. Carbonell, T. M. Mitchell, pp. 83–129. Palo Alto: Tioga Publishing
Minker, J. 1982. On indefinite databases and the closed world assumption. *Proc. 6th Conf. Automat. Deduct.*, New York, pp. 292–308
Minsky, M. 1975. A framework for representing knowledge. In *The Psychology of Computer Vision*, ed. P. H. Winston, pp. 211–77. New York: McGraw-Hill. 282 pp.
Moore, R. C. 1984. Possible-world semantics for autoepistemic logic. *Proc. AAAI Workshop Non-Monotonic Reason.*, New Paltz, pp. 396–401
Moore, R. C. 1985. Semantical considerations on nonmonotonic logic. *Artif. Intell.* 25: 75–94
Nilsson, N. 1986. Probabilistic logic. *Artif. Intell.* 28: 71–87
Nute, D. 1984. Non-monotonic reasoning and conditionals. *Univ. Georgia, Adv. Comput. Meth. Cent. Res. Rep. 01-0002, Athens*
Peng, Y., Reggia, J. A. 1986. Plausibility of diagnostic hypotheses: the nature of simplicity. *Proc. Am. Assoc. Artif. Intell. Natl. Conf.*, Philadelphia, pp. 140–45
Perlis, D., Minker, J. 1986. Completeness results for circumscription. *Artif. Intell.* 28: 29–42
Perrault, C. R. 1987. An application of default logic to speech act theory. *SRI Int. Artif. Intell. Cent. Tech. Rep., Palo Alto.* 26 pp.
Poole, D. L. 1985. On the comparison of theories: preferring the most specific explanation. *Proc. Int. Joint Conf. Artif. Intell.*, Los Angeles, pp. 144–47
Poole, D. 1986. Default reasoning and diagnosis as theory formation. *Univ. Waterloo, Dept. Comput. Sci. Tech. Rep. CS-86-08, Waterloo*
Putnam, H. 1970. Is semantics possible? In *Language, Belief, and Metaphysics*, ed. H. E. Kiefer, M. K. Munitz, pp. 50–63. Albany: State Univ. New York Press
Reggia, J. A., Nau, D. S., Wang, Y., Peng, Y. 1985. A formal model of diagnostic inference. *Inform. Sci.* 37: 227–85
Reiter, R. 1978a. On reasoning by default. *Proc. TINLAP-2, Theor. Issues Nat. Lang. Process. 2*, Univ. Illinois, Urbana-Champaign, pp. 210–18
Reiter, R. 1978b. On closed world data bases. See Gallaire & Minker 1978, pp. 55–76
Reiter, R. 1980. A logic for default reasoning. *Artif. Intell.* 13: 81–132
Reiter, R. 1982. Circumscription implies predicate completion (sometimes). *Proc. Am. Assoc. Artif. Intell. Natl. Conf.*, Pittsburgh, pp. 418–20
Reiter, R. 1987. A theory of diagnosis from first principles. *Artif. Intell.* In press
Reiter, R., Criscuolo, G. 1983. Some representational issues in default reasoning. *J. Comput. Math. Appl.* 9: 1–13 (Special issue on computational linguistics)
Rich, E. 1983. Default reasoning as likelihood reasoning. *Proc. Am. Assoc. Artif. Intell. Natl. Conf.*, Washington, pp. 348–51
Rosch, E. 1978. Principles of categorization. In *Cognition and Categorization*, ed. E. Rosch, B. B. Lloyds. New York: Lawrence Erlbaum Assoc. 328 pp.
Sandewall, E. 1972. An approach to the frame problem and its implementation. In *Machine Intelligence 7*, ed. B. Meltzer, D. Michie, pp. 195–204. Edinburgh: Edinburgh Univ. Press
Sandewall, E. 1985. A functional approach to non-monotonic logic. *Comput. Intell.* 1: 80–87
Shepherdson, J. C. 1984. Negation as failure:

a comparison of Clark's completed data base and Reiter's closed world assumption. *J. Logic Program.* 1: 51–79

Shoham, Y. 1986. *Reasoning about change: time and causation from the standpoint of artificial intelligence*. PhD thesis. Yale Univ., New Haven

Stalnaker, R. 1968. A theory of conditionals. In *Studies in Logical Theory*, ed. N. Rescher, pp. 98–112. Oxford: Blackwell

Zadeh, L. 1981. PRUF—a meaning representational language for natural languages. In *Fuzzy Reasoning and Its Applications*, ed. E. Mamdani, B. Gaines. New York: Academic. 381 pp.

Zadeh, L. A. 1985. Syllogistic reasoning as a basis for combination of evidence in expert systems. *Proc. Int. Joint Conf. Artif. Intell.*, Los Angeles, pp. 417–19