The Knowledge Engineering Review, Vol. 4: 2, 1989, 119–139

Logics for reasoning about knowledge and belief

HAN REICHGELT*

Artificial Intelligence Group, Department of Psychology, University of Nottingham, Nottingham NG7 2RD, England

Abstract

AI researchers are becoming increasingly aware of the importance of reasoning about knowledge and belief. This paper reviews epistemic logic, a logic designed specifically for this type of reasoning. I introduce epistemic logic, and discuss some of the philosophical problems associated with it. I then compare two different styles of implementing theorem provers for epistemic logic. I also briefly discuss autoepistemic logic, a form of epistemic logic intended to model an agent's introspective reasoning, i.e. an agent's reasoning about its own beliefs. Finally, I discuss some of the proposals in the AI literature that are aimed at avoiding some of the philosophical problems that dog both epistemic and autoepistemic logic. This paper is not a full introduction to the field. Rather, it is intended to give the reader some flavour of the problems that research in this area faces, as well as some of the proposals for solving these problems.

1 Introduction

There is a growing awareness in Artificial Intelligence of the importance of reasoning about knowledge and belief, or *epistemic* reasoning.[1] There have been two AI conferences on this topic (Halpern, 1986; Vardi, 1988), and each year at least a dozen papers dealing with various aspects of the problem of epistemic reasoning appear in other AI conferences and journals.

One can give many examples of areas in which agents need to be able to reason about their own and other agents' beliefs. One example is multi-agent problem solving systems where agents need to coordinate their actions, and hence need to be able to reason about the likely actions of other agents. But which actions an agent may perform depends on its beliefs about the world. Therefore, if agent *a* needs to reason about the actions agent *b* can take, *a* needs to be able to reason about what *b* believes.

Another example of the need to reason about other agents' beliefs is natural language processing. The way in which I refer to a particular entity often depends on what knowledge I ascribe to you. Whether I refer to my wife as *my wife*, *the woman sitting beside you*, or *Nikki* partly depends on what I think you know.

A final area that I want to mention is common sense reasoning. We often draw conclusions about the world by introspecting on our own knowledge base, and noticing the absence of certain pieces of information. A standard example concerns the reasons that I have for believing that I have no older brother. I have no direct evidence to this effect. Rather I draw this inference on the basis of my belief that if I had any older brothers, then I would know them, and the fact that I do not know of any. It follows that I do not have any older brothers. Moore (1985b, 1988) calls the process of reasoning from one's own beliefs *autoepistemic reasoning*. He developed a logic to model this type of reasoning, called *autoepistemic* logic. In section 4 I briefly discuss Moore's logic.

* I'd like to thank Mike Burton, Peter Jackson and an anonymous reviewer for their comments on an earlier draft of this paper.

[1] In addition to the term "*epistemic* reasoning", authors also use the term "*doxastic* reasoning". The term "doxastic reasoning" is normally used to refer to reasoning about belief, whereas the term "epistemic reasoning" is used both to refer to reasoning about knowledge and reasoning about belief. In this paper I will use the term "epistemic reasoning".

Most of the work in AI on reasoning about knowledge and belief has used some form of logic as its main knowledge representation language.[2] Arguments in favour of logic include the fact that it has a clear model-theoretic semantics (Hayes, 1977), which allows you to determine the meanings of the representations independently of the inference machinery that is part of the knowledge representation language. Another argument is the greater expressive power that logic affords (Moore, 1984; Levesque and Brachman, 1985). For example, logic allows you to express the information that something has some property ϕ without forcing you to name this entity. This is of course of special importance in epistemic reasoning. For example, I may believe that you believe that some person proved Fermat's last theorem without me being able to name that person.

A further reason for concentrating on the use of logic in epistemic reasoning is the fact that there is a custom-built logic for this type of reasoning. This logic, called *epistemic logic*, was developed by Hintikka (1962, 1969). Because epistemic logic has proved to be an important starting point in research on reasoning about knowledge and belief, I discuss it in some detail in section 2.1. However, epistemic logic also leads to a number of problems. These are discussed in section 2.2.

Although many authors have used epistemic logic in their research on reasoning about knowledge and belief, there is no clear consensus about the best way to implement a theorem prover for it. In section 3, I will discuss two different approaches to this problem, and compare them. Not surprisingly perhaps, the conclusion will be that one approach is more expressive but computationally less efficient than the other, thus providing yet another example of the tradeoff between expressive power and computational tractability discussed in Levesque and Brachman (1985).

Many of the approaches discussed in section 3 simply ignore the problems that dog any logical analysis of epistemic reasoning. In section 5 I discuss some formal systems that are aimed specifically at these problems.

This paper is intended as an introduction to the different logic-based aproaches to reasoning about knowledge and belief. It does not pretend to be an exhaustive overview of the field. Nevertheless, I hope that this introduction will give the reader some feel for the various approaches that have been taken, and their advantages. McArthur (1988) provides an alternative review of much of the material surveyed here.

2 Epistemic logic

Epistemic logic was developed by Hintikka in an attempt to provide a logic that could be used to analyze reasoning about knowledge and belief. Epistemic logic is a form of modal logic. Modal logics are obtained from classical propositional or predicate calculus by adding a necessity operator, $\Box$ or L[3]. $\Box\phi$ is interpreted as "ϕ is necessary". In epistemic logic, however, this intuitive interpretation changes, and $\Box\phi$ is interpreted as "ϕ is believed (or known)". The operator $\Box$ is therefore normally replaced by a belief operator B or a knowledge operator K. In this paper, I use B. Moreover, the single necessity operator of standard modal logic is often replaced by a number of belief operators or knowledge operators, one for each agent about whose beliefs one wants to reason. Thus, one has separate B-operators, B_a, B_b, etc, for the different agents a, b, etc. In this paper however, I restrict myself to the single agent logics.

This section is divided into two subsections. In 2.1, I discuss epistemic logic in some technical detail. I have tried to avoid excessive technicalities but certain of the approaches discussed in section 3 can only be understood if one has at least an intuitive understanding of epistemic logic. I first discuss the intuitions underlying epistemic logic. I then briefly describe epistemic propositional logic in 2.1.1, and I generalize this to epistemic predicate logic in section 2.1.2. In 2.2, I discuss some of the

[2] Rapaport (1986) is an exception. He investigates the representation of reasoning about belief in the semantic network formalism, SNePS. Moore and Hendrix (1979) also look at the use of semantic networks for the representation of belief.

[3] For an introduction to first-order predicate calculus, the reader is referred to any introductory book on logic, such as Mates (1972), Boolos and Jeffrey (1982) or Mendelson (1987). For an introduction to modal logic, the reader is referred to Hughes and Cresswell (1968) or Chellas (1980).

philosophical objections that have been raised against epistemic logic. Some of the formal systems discussed in Section 5 are directly aimed at avoiding these problems.

2.1 A short introduction to epistemic logic

As any modal logic, epistemic logic has a so-called *possible world* semantics. Possible world semantics was originally developed by Kripke (1963). Intuitively, a possible world is simply a possible state of affairs. Clearly, the current state of affairs, or the *actual* world, is also a possible world. With each proposition ϕ in the language, one can then associate the set of possible worlds in which ϕ is true. Thus, the current world would be associated with the proposition *The prime minister of England is female.* However, the possible world which is exactly like the present world, except that Margaret Thatcher did not win the last election, but rather a male politician, would not be associated with this sentence. Kripke also postulated a so-called *accessibility* relation between possible worlds. Intuitively, if a possible world w' is accessible from w, then w' would be an "imaginable" state of affairs if w were the actual world. Thus, a world in which pigs fly might not be accessible from the actual world. However, such a world might be imaginable, and hence accessible, from a world in which, by some quirk of nature, dogs fly.

As I have said, Hintikka uses Kripke's possible world semantics in his epistemic logic. The basic idea is to associate with an agent a the set of possible worlds in which all a's beliefs are true. The accessibility relation between possible worlds is then also slightly reinterpreted. Intuitively, w' is accessible from w means that there is nothing in the beliefs of the agent that allows him or her to distinguished between w and w'. Thus, as far as the agent's beliefs are concerned, if w were the real world, then w' could have been the real world as well. If w' is accessible from w in epistemic logic, we also say that w' is an *epistemic alternative* to w.

2.1.1 Epistemic propositional logic

In order to make these intuitive ideas more precise, I first discuss epistemic propositional logic. The language of this logic can be obtained from the language of classical propositional logic by adding the belief operator B. B is a one-place operator: if ϕ is a well-formed sentence, then so is $B\phi$.

In order to define the meanings of the sentences of epistemic propositional logic, we define the notion of a *propositional epistemic frame* E. E is a triple consisting of a set of possible worlds W, a distinguished element aw of W, and a binary relation R on W. They are the set of possible worlds, the actual world, and the accessibility relation between worlds respectively.

A lot of discussion has taken place about the additional constraints that should be put on the accessibility relation R. As we will see below, the interpretation of sentences containing the B-operator crucially depends on the accessibility relation. As a result, different constraints on the accessibility relation lead to different interpretations of the belief operator. Over the last few years, a consensus seems to have formed that that accessibility relation should be euclidean, transitive, and serial. The accessibility relation is euclidean if wRw' and wRw'' imply $w'Rw''$; the accessibility relation is transitive if wRw' and $w'Rw''$ imply wRw''; and the accessibility relation is serial if for every w, there is a w' such that wRw'. The resulting modal system is known as KD45 (Chellas, 1980). I return to the reasons behind these restrictions on the accessibility relation below.

In order to assign truth values to propositions, we add an *interpretation function*, I, to a propositional epistemic frame $E = \langle W, aw, R \rangle$. A pair consisting of an epistemic frame E and an interpretation function I is called an *epistemic interpretation.* The interpretation function I assigns to each atomic sentence in the language a truth value in each possible world in W. Given an assignment of truth values to atomic propositions in each world, one can recursively determine the truth values of other propositions. Thus, $\neg\chi$ is true in some world w under some assignment I if and only if x is false in w under I. Similarly, $(\phi \,\&\, \chi)$ is true in w under I if and only if both ϕ and χ are true in w under I. There are similar conditions for the other connectives. The meaning of the belief operator B is given by the rule: $B\phi$ is true in a world w under an assignment I if I assigns the truth value *true*

to ϕ in all worlds w' such that wRw'. In other words, we say that $B\phi$ is true in w if ϕ is true in all epistemic alternatives to w.

Having defined the truth of a proposition in a world w, we can now define the notion of truth in an epistemic interpretation $\langle\langle W, aw, R\rangle, I\rangle$. A proposition ϕ is true in $\langle\langle W, aw, R\rangle, I\rangle$ if and only if I assigns the truth value *true* to ϕ in aw. In other words, a proposition ϕ is true in an epistemic interpretation if it is true in the actual world in this interpretation. The notions of validity and valid consequence are then defined in the normal way: a proposition is *valid* if and only if it is true in all epistemic interpretations; a proposition ϕ is a *valid consequence* of a set of sentences S if and only if ϕ is true in all interpretations in which all the sentences in S are true.

The rule for determining the truth value of a proposition $B\phi$ depends crucially on the notion of accessibility between worlds. A consensus seems to have developed that the best modal logic for reasoning about belief is one in which the accessibility relation is euclidean, transitive and serial. Such a logic is known as (propositional) KD45. In it, the following propositions can be shown to be valid:

a all axioms of classical propositional calculus
b $(B\phi \,\&\, B(\phi \rightarrow \chi)) \rightarrow B\chi$
c $\neg B\phi \rightarrow B\neg B\phi$ (negative introspection)
d $B\phi \rightarrow BB\phi$ (positive introspection)
e $B\phi \rightarrow \neg B\neg\phi$

Propositions (b) can be paraphrased as follows: if an agent believes that ϕ, and it believes that ϕ implies χ, then it believes χ as well. In other words, an agent believes all the logical consequences of its beliefs; its beliefs are closed under logical deduction. In 2.2, we will see that this aspect of epistemic logic has attracted a lot of criticism. Propositions (c) and (d) mean that an agent is assumed to have perfect introspective ability: they know exactly what they know and what they do not know. Proposition (c) corresponds to the fact that the accessibility relation is euclidean, while (d) corresponds to its transitivity, (e) finally simply means that an agent can never believe a proposition and its negation at the same time. In other words, an agent's knowledge base always has to be consistent. It corresponds to the seriality requirement on the accessibility relation.

KD45 is intended to be a logic for reasoning about belief. However, it is relatively straightforward to design a logic for reasoning about knowledge, at least as long as we analyze knowledge as true belief.[4] All we have to do is replace the requirement that the accessibility relation be serial by the stronger requirement that it be reflexive. The accessibility relation is reflexive if for every w, wRw. The resulting system is known as (propositional) S5. This means that the following proposition is valid as well:

$$B\phi \rightarrow \phi.$$

Thus, if an agent knows that ϕ, then ϕ must be true. Reflexivity implies seriality. Proposition (e) is therefore still valid. Thus, in this logic of knowledge, an agent's beliefs are still assumed to be closed under logical deduction; agents are still assumed to have perfect introspective capabilities; their beliefs are still consistent. But in addition, their beliefs are assumed to be true in the actual world.

2.1.2 Epistemic predicate logic

Although the analysis in the previous section is a first step towards providing a logic for epistemic reasoning, it is not complete. The main reason is that it is restricted to propositional epistemic logic, and one cannot express things like *John believes that every logician is boring.* The most straightforward way of generalizing to epistemic predicate logic is analogous to the way in which one generalizes standard propositional logic to standard predicate logic: we add a set of individuals

[4] Many philosophers would argue that the definition of knowledge as true belief is too weak. What is missing, they would argue, is some notion of justification. Knowledge should be analyzed not merely as true belief, but rather as true *justified* belief. However, logics for knowledge generally rely on the weaker analysis.

to the epistemic frame and complicate the interpretation function I. We assume that I assigns to each individual constant α and each world w an individual, and to each n-place predicate P and each world w a set if n-tuples of individuals. We can then say that an atomic proposition, such as $P(\alpha)$, is true in some world w if $I(w, \alpha) \in I(w, P)$. Similarly, we say that a proposition such as $(\forall x)P(x)$ is true in w if all individuals have the property P in w.

However, there are a number of problems with this generalization. For example, how would we express the proposition that an agent believes of the individual that is prime minister in the actual world, Margaret Thatcher, that she is the queen of England? We cannot represent this as

$$B(queen(prime\text{-}minister)).$$

This proposition would be true in some world w, if the proposition *queen*(*prime-minister*) is true in all worlds w' accessible to w. But the individual assigned to the constant *prime-minister* need not be the same in w' as it is in w. For example, our agent may well believe that the individual who, unbeknown to the agent, happens to be prime minister in the actual world, is queen of England, without also believing that the person that *it* thinks is prime minister is queen. For example, our agent may be under the misconception that Denis Thatcher is prime minister, and our agent certainly does not believe of him that he is queen. In other words, the individual assigned to some constant in the actual world need not be the same as the individual assigned to it in the agent's belief worlds.

This problem is often solved by drawing a distinction between *rigid* and *fluent* expressions. Rigid expressions are expressions whose meaning does not change between worlds, where fluent expressions may change their meaning. Thus, if ε is a rigid expression, then $I(w, \varepsilon) = I(w', \varepsilon)$ for all worlds w and w'. The above problem can be solved by using a rigid individual constant, also called a *rigid designator*. We would express the fact that an agent believes that the individual who happens to be prime minister in the real world is queen of England, by using a rigid designator, such as *Margaret Thatcher*, to refer to this individual. Thus, the constant *Margaret Thatcher* would denote the same individual, no matter in which world we evaluate it. We could then express our agent's belief that the person who happens to be prime minister in the real world is queen as

$$B(queen(Margaret\ Thatcher)).$$

Following Kripke (1972), individual constants are often treated as rigid designators, whereas predicates are usually regarded as fluents.

However, there is a second problem. Assume that $(\forall x)B\phi x$ is true in world w. Then for any individual a, $B\phi a$ is true in w, and hence ϕa must be true in all worlds w' accessible from w. But since a is an arbitrary individual, this means that $(\forall x)\phi x$ must be true in all worlds w', and hence that $B(\forall x)\phi x$ is true in w. By a similar argument, we can deduce $(\forall x)B\phi x$ from $B(\forall x)\phi x$. These formulas are known as the Barcan formula (BF) and its converse (FB). Thus, the following two formulas would be valid:

$$\text{BF: } (\forall x)B\phi x \rightarrow B(\forall x)\phi x$$

$$\text{FB: } B(\forall x)\phi x \rightarrow (\forall x)B\phi x.$$

The problem is that for most applications neither BF nor FB should be valid. BF is not valid because, even though an agent may believe of everything that actually exists that it has the property ϕ, thus making $(\forall x)B\phi x$ true, it might believe that something which happens to be a figment of its imagination does not have the property ϕ, thus falsifying $B(\forall x)\phi x$. On the other hand, FB is not valid because the fact that an agent believes of all the entities that it knows about that they have the property ϕ, does not warrant any inferences about the set of individuals in the actual world. There might well be real world individuals that do not exist in the agent's beliefs.

The standard way in epistemic logic of avoiding this problem is by complicating the notion of an interpretation. From a model-theoretic point of view, both BF and FB indicate that the set of individuals is the same between the different possible worlds. So, if we can find a way of avoiding this

assumption, then we have a way of avoiding both BF and FB. The way in which this is achieved formally is by complicating the notion of a propositional epistemic frame in a slightly different way than before. One adds a set of possible individuals and a function that assigns to each possible world a subset of this set of possible individuals. The set of possible individuals assigned to some world w is to be regarded as the set of individuals actually existing at w. Quantification is now restricted to individuals existing in some world. Thus, whereas before $(\forall x)\phi x$ was true in world w if all individuals had the property ϕ in w, we now say that $(\forall x)\phi x$ is true in world w if all individuals existing in w have the property ϕ. Since the set of individuals existing in one world is not necessarily identical to the set of individuals existing in some other world, neither BF nor FB is true in this complicated model theory. Readers are encouraged to check this.[5]

2.2 Philosophical problems with epistemic logic

Although epistemic logic is possibly acceptable as a formal system. it is less clear how satisfactory epistemic logics are as models of reasoning about knowledge and belief. The main problem is the so-called problem of *logical omniscience*. We can distinguish between two aspects of this problem.

The first aspect of the problem is that an agent is assumed to know all tautologies. A tautology is a proposition that is true in all possible worlds in all interpretations. $B\phi$ is true in a world w if and only if ϕ is true in all words accessible from w. Now, if a proposition χ is true in all possible worlds, then it is certainly true in all possible worlds accessible from w, and therefore $B\chi$ is true in w. Clearly, the above argument applies not just to w, but to all possible worlds, and it therefore follows that $B\chi$ is valid for any valid proposition χ.

A second aspect of the problem of logical omniscience is the fact that agents are assumed to know all the logical consequences of their beliefs. If ϕ is a logical consequence of a set of sentences S, then ϕ is true in all worlds in which the sentences of S are true. If S is the set of beliefs of the agent, then all sentences in S will be true in all accessible worlds. Obviously, ϕ will be true in those worlds as well, and hence the agent must believe ϕ too. This problem, which is also known as the problem of *consequential closure*, ignores the fact that humans are resource-limited reasoners, and do not have the resources to infer all the logical consequences of their beliefs.

The problem of consequential closure implies that agents with inconsistent beliefs believe everything. After all, any proposition is entailed by the contradiction. This problem may not seem too serious because not many of us would willingly admit to believing both ϕ and $\neg\phi$ for some (specific) proposition ϕ. To use terminology due to Konolige, our beliefs are in general *non-contradictory*: faced with an explicit contradiction, we would not admit to believing it. However, it seems likely that our entire set of beliefs is logically inconsistent: they logically imply a contradiction, even though, due to our limited resources, we have not explicitly derived any contradictions ourselves. Since epistemic logic identifies our beliefs with the set of possible worlds in which our beliefs are true, it is impossible to make a formal distinction between the notions of non-contradictoriness and logical inconsistency.

Another consequence of deductive closure is the fact that logically equivalent beliefs are deemed identical. The propositions *it is raining outside* and *arithmetic is incomplete and it is raining outside* are equivalent. For all possible worlds, if one is true, then the other is. Therefore, if an agent believes in one, it is also assumed to believe in the other. However, intuitively at least, these two propositions seem to express different beliefs, and therefore it should be possible for an agent to believe the first proposition without believing the second.

[5] There is in fact a further complication that one needs to take into account when one defines a proof theory for modal epistemic logic. If we allow formulas with free variables as theorems, such as $\phi x \vee \neg\phi x$, then the axioms of classical predicate calculus entail the converse of the Barcan formula. However, if we refuse to recognize strings like ϕx as well-formed formulas, and instead insist that only something like $(\forall x)\phi x$ is well-formed, and we treat constants as rigid designators (or we restrict the applications of certain inference rules to rigid designators), then the converse of the Barcan formula does not follow. As Hughes and Cresswell (1968, pp 178–183) point out, this shows that one has to be careful in choosing which form of classical predicate logic one uses as one's basis from which to construct epistemic predicate logic. However, these logical subtleties need not concern us here.

Clearly, from an intuitive point of view, logical omniscience is a problem. However, it should be said that many logicians have rejected the problem of logical omniscience as irrelevant to their work. They argue that epistemic logic, if it models anything, models the epistemic reasoning of *ideal* rational agents which are not encumbered by such things as resource limitations, or imperfect logical knowledge. Thus, the problem of logical omniscience arises only if one wants to model and reason about more natural agents, but this, so these logicians would argue, is not the aim of epistemic logic. However, in AI one is interested in building and reasoning about artificial systems that by necessity suffer from certain limitations, and this reply might therefore not be acceptable from an AI point of view.

Another way of making this point is to say that a logic that does not acknowledge the resource limitations of agents will lead to an intuitively wrong definition of logical consequence when one is reasoning about the type of agents that AI applications reason about. So, because we know that agents are resource limited, we will not in general conclude that some agent believes $\neg\phi$ even if we know that it believes $\neg\chi$ and $\phi \rightarrow \chi$. Therefore, an epistemic logic which leads to this conclusion cannot be said to capture adequately the notion of belief that underlies both reasoning about the beliefs of resource limited reasoners and the design of artificial agents. I return to this issue in section 5 where I discuss some of the systems that have developed specifically to deal with this problem.

3 Theorem provers for epistemic logic

In the previous section I briefly discussed epistemic logic, and some of the philosophical problems associated with it. However, apart from these philosophical problems, there is a further problem that has to be solved by any AI researcher who proposes to use epistemic logic for reasoning about knowledge and belief, namely how to construct a theorem prover for epistemic logic. One can distinguish between two different approaches to this problem, the *modal* and the *reified* approach. The modal approach amounts to trying to build theorem provers directly for modal epistemic logic. The refied approach on the other hand is based on the observation that the model theory for epistemic logic can itself be formulated in a classical first-order predicate logic. If one therefore translates a sentence ϕ of modal epistemic logic into the sentence of reified epistemic logic which exactly described the meaning of ϕ, then one can use a standard theorem prover to reason about the formalizations of the meanings of propositions in epistemic logic. So, rather than reasoning *with* propositions of epistemic logic, as one does in the modal approach, the reified approach reasons *about* these propositions. I briefly discuss the two approaches, and compare them.

3.1 Modal epistemic theorem provers

There are a number of modal epistemic theorem provers available. Examples are the systems discussed in Halpern and Moses (1985), a number of systems developed by Konolige (Konolige, 1986a, 1986b; Geissler and Konolige, 1986), and Jiang (1987). Moreover, a number of authors have proposed general theorem provers for modal logic, e.g. Abadi and Manna (1986), Wallen (1987), Jackson and Reichgelt (1989a,b). Since epistemic logic is a form of modal logic, these various modal theorem provers can of course be changed into epistemic theorem provers, simply by turning them into theorem provers for KD45, possibly without the Barcan formula or its converse. Rather than discuss the different modal epistemic theorem provers in great detail, I will briefly discuss some of the techniques that have been developed.

The main problem for any modal theorem prover can perhaps best be illustrated by considering the complications that would have to be made to a resolution-based theorem prover for classical predicate calculus. In such a theorem prover, the only inference rule that is used is resolution: from $(\phi \vee \phi')$ and $(\neg\chi \vee \chi')$ conclude $(\phi'\sigma \vee \chi'\sigma)$ provided that ϕ and χ unify with a substitution σ. $\phi'\sigma$ and $\chi'\sigma$ are the result of applying σ to ϕ' and χ' respectively. A substitution σ is a function which assigns to each variable in a formula a value, i.e. either another variable, an individual constant, or a

functional expression. Two formulas ϕ and χ unify if there is a substitution σ such that $\phi\sigma$ and $\chi\sigma$ are identical. Thus, the formulas $\phi(x, a)$ and $\phi(b, y)$ unify because if we apply the substitution $\{b\backslash x, a\backslash y\}$ to them then in both cases the result will be $\phi(b, a)$. We can therfore apply the resolution rule to the formulas $(\phi(x, a) \vee \chi(x, y))$ and $(\neg\phi(b, y) \vee \psi(z))$ to conclude $(\chi(b, a) \vee \psi(z))$. The intuitive motivation behind resolution can be illustrated in the above example. The formulas $(\phi(x, a) \vee \chi(x, y))$ and $(\neg\phi(b, y) \vee \psi(z))$ are assumed to be true at the same time. However, the substitution $\{b\backslash x, a\backslash y\}$ gives an example of a case in which $\phi(x\ a)$ and $\neg\phi(b, y)$ cannot be true at the same time. Now assume that $\phi(b, a)$ is not true. Then obviously $(\phi(x, a) \vee \chi(x, y))$ can only be true if $\chi(b, a)$ is true. Alternatively, assume that $\neg\phi(b, a)$ is not true, then $(\neg\phi(b, y) \vee \psi(z))$ can be true only if $\psi(z)$ is true. One can therefore conclude $(\chi(b, a) \vee \psi(z))$. The reader is referred to Nilsson (1982), Robinson (1979), and Stickel (1987) for detailed introductions to unification and resolution.

For modal logic in general, and modal epistemic logic in particular, the story needs to be complicated. Propositions are no longer true or false per se; rather they are true or false with respect to some world *w*. Therefore, in order to resolve two propositions $(\phi \vee \phi')$ and $(\neg\chi \vee \chi')$ we have to make sure not only that ϕ and χ unify, but also that they unify *in the same world.* In addition, because individual terms may denote different individuals across different worlds, we need to ensure that the different terms in the formulas do in fact denote the same individual.

There are various ways in which a theorem prover for classical first-order predicate calculus can be complicated to deal with this problem. Wallen (1987) and Jackson and Reichgelt (1989a,b) for example associate with each formula a world index indicating in which world the formula is true. Similarly, to deal with the problem of changing denotations for individual terms, they associate with each individual term in a formula a term index which indicates the world in which the individual originated. The unification algorithm is then complicated to take these indices into account. Two indexed formulas unify if the following conditions hold: (i) the formulas themselves unify under the standard definition for classical predicate calculus; (ii) the world indices asociated with the formulas can be shown to denote the same world; (iii) the term indices associated with corresponding terms in the formulas can be shown to denote the same world, so that we can be sure that the corresponding terms can be made to denote the same individual; and (iv) the various unifications obtained in (i), (ii) and (iii) are compatible.

Another possibility is explored by Konolige (1986a, 1986b; Geissler and Konolige, 1986). Konolige's theorem prover, like most other theorem provers, is based on refutation. Thus, to prove ϕ from a set of sentences S, one adds $\neg\phi$ to S, and tries to prove that the resulting set of propositions is unsatisfiable by deducing the empty formula, which is always false. Konolige introduces the notion of *modal depth.* The modal depth of a proposition ϕ is the number of embedded belief-operators in ϕ. Thus, $\phi(a)$ has a modal depth of 0; $B(\phi(a))$ and $B(\phi(a)) \vee \chi(b)$ both have a modal depth of 1; $B(B(\phi(a)))$ has a modal depth of 2 etc. The basic idea behind Konolige's theorem prover is to reduce the unsatisfiability of a set of sentences of modal depth n to those of a set of sentences of modal depth $n - 1$. This will eventually reduce to the unsatisfiability of a set of sentences of modal depth 0, i.e. the unsatisfiability of a set of sentences in classical predicate calculus. Clearly, in order to determine whether this is the case, one can simply use a standard theorem prover. In addition, in order to deal with the problem of varying denotations for individual terms, Konolige introduces a separate bullet operator •, which is put in front of a term. The denotation of a term such as $\bullet\alpha$ then is the individual denoted by α in the real world. Thus, *B(queen(•prime-minister))* means that the agent believes of the individual which is in reality the prime minister of England that he or she is queen, whereas *B(queen(prime-minister))* means that the agent believes that the individual that it believes is prime-minister, is also queen. Of course, the first sentence may well be true while the second is false. Again, Konolige then needs to complicate the unification algorithm to take this complication into account.

3.2 Reified epistemic theorem provers

An alternative way of implementing a theorem prover for epistemic logic is due to Moore (1985a),

partly inspired by McCarthy (1979). Moore observed that the semantics of modal epistemic logic could itself be formalized in classical predicate calculus. One can therefore give a number of axioms, formulated in the language of classical predicate calculus, which describe the conditions under which sentences in modal epistemic logic are true. It follows that there is no need to complicate the standard theorem provers for classical predicate logic. All one needs to do is to translate a modal epistemic sentence ϕ into a sentence in classical predicate calculus that describes the meaning of ϕ. So, rather than reason with sentences in modal epistemic logic, one treats these sentences themselves as objects (one "reifies" them), and one reasons about them. The resulting logic can be called "reified epistemic logic".

To illustrate the reified approach, some more notation is needed: if ϕ is an atomic sentence in modal epistemic logic, then $[\phi]$ is the name of that sentence in reified epistemic logic. If ϕ is of the form (χ connective ψ), then $[\phi]$ is of the form [connective]([χ], [ψ]). To improve readability, I will normally use [χ connective ψ]. With this in mind, we can give a number of axioms in reified epistemic logic. In modal epistemic logic, a sentence of the form $B\phi$ is true in a world w if ϕ is true in all worlds w' accessible from w. In order to formulate this in the language of reified epistemic logic, I use the predicates *true* and R. R stands for the accessibility relation between worlds. The following axiom, which I will call the *B-axiom*, corresponds to the clause in the semantics for modal epistemic logic in which the meaning of the belief-operator B is defined.

$$(\forall w)(true(w, [B\phi]) \equiv (\forall w')(R(w, w') \rightarrow true(w', [\phi])))$$

Because we are interested in the modal system KD45, we also need a number of axioms corresponding to the restrictions on the accessibility relation. The reader will recall that the accessibility relation was required to be serial, transitive and euclidean. So, we add the following axioms:

$$(\forall w)(\exists w')(R(w, w')),$$

$$(\forall w)(\forall v)(\forall u)((R(w, v) \,\&\, R(v, u)) \rightarrow R(w, u)),$$

$$(\forall w)(\forall v)(\forall u)((R(w, v) \,\&\, R(w, u)) \rightarrow R(v, u)).$$

Moore also gives a number of axioms for dealing with connectives in the reified propositions:

$$(\forall w)(true(w, [\phi \,\&\, \chi]) \equiv (true(w, [\phi]) \,\&\, true(w, [\chi]))),$$

$$(\forall w)(true(w, [\phi \vee \chi]) \equiv (true(w, [\phi]) \vee true(w, [\chi]))),$$

$$(\forall w)(true(w, [\phi \rightarrow \chi]) \equiv (true(w, [\phi]) \rightarrow true(w, [\chi]))),$$

$$(\forall w)(true(w, [\neg \phi]) \equiv \neg true(w, [\phi])).$$

Moore adds another axiom, which I will call the $\forall$*haxiom*:

$$(\forall w)(true(w, [(\forall y)(\phi(y))])) \equiv (\forall x)(true(w, [\phi(@(x)/y)])).$$

In this formula, @ associates with an object in the domain a rigid designator. Thus, if o is an object in some possible world, then $@(o)$ is a name for that object. $\phi(x/y)$ is obtained from $\phi(y)$ by replacing all occurrences of x by y.

The $\forall$ axiom is problematic as it is partly responsible for the fact that in Moore's reified logic both the Barcan formula and its converse are derivable. Below, I show how the Barcan formula itself can be derived. The proof that its converse is derivable as well goes along similar lines.

1	$true(w, [(\forall x)B\phi(x)])$	
2	$(\forall x)(true(w, [B\phi(@(x)/y)]))$	$\forall$-axiom
3	$(\forall x)(\forall w')(R(w, w') \rightarrow true(w', [\phi(@(x)/y)]))$	*B*-axiom
4	$(\forall w')(\forall x)(R(w, w') \rightarrow true(w', [\phi(@(x)/y)]))$	predicate calculus
5	$(\forall w')(R(w, w') \rightarrow (\forall x)(true(w', [\phi(@(x)/y)])))$	predicate calculus

6 $(\forall w')(R(w, w') \rightarrow true((w', [(\forall x)\phi(x)]))$	$\forall$-axiom
7 $true(w, [B(\forall x)\phi(x)])$	B-axiom

In the above arguments all the steps are licensed by the axioms that Moore himself gives, with the exception of the steps from 3 to 4 and 4 to 5 which involve quantificational reasoning.

It is possible to complicate reified logic in such a way that the Barcan formula is no longer valid. In section 2 we saw that the Barcan formula and its converse meant that the set of individuals existing at some world w was always identical to the set of individuals existing at some world w'. Bearing this in mind, one can avoid the Barcan formula and its converse by complicating the $\forall$-axiom. We first introduce a further 2-place predicate *exists* which is true of a (possible) individual i and a world w if i exists at w. We then complicate the $\forall$-axiom to:

$$(\forall w)(true(w, [(\forall y)\phi(y)]) \equiv (\forall x)(exists(x, w) \rightarrow (true(w, [f(@(x, w)/y)])))),$$

where $@(x, w)$ now associates with an individual x a name in world w. With this revised $\forall$-axiom, the Barcan formula nor its converse are any longer derivable. The following attempt shows that the last step in the original proof is no longer allowed.

1 $true(w, [(\forall x)B\phi(x)])$	
2 $(\forall x)(exists(x, w) \rightarrow true(w, [B\phi(@(x)/y)]))$	revised $\forall$-axiom
3 $(\forall x)(exists(x, w) \rightarrow (\forall w')(R(w, w') \rightarrow true(w', [\phi(@(x)/y)])))$	B-axiom
4 $(\forall w')(\forall x)((exists(x, w) \,\&\, R(w, w')) \rightarrow true(w', [\phi(@(x)/y)]))$	predicate calculus
5 $(\forall w')(R(w, w') \rightarrow (\forall x)(exists(x, w) \rightarrow true(w', [\phi(@(x)/y)])))$	predicate calculus

The revised $\forall$-axiom will no longer allow us to conclude

$$(\forall w')(R(w, w') \rightarrow true(w', [(\forall x)\phi(x)]))$$

from 5, as the individual exists at w and not at w'. We thus prevent the derivation of the Barcan formula and its converse. The cost is of course a more complicated notation.

3.3 A comparison of the reified and the modal approach

Given the distinction between modal and reified epistemic logic, the question naturally arises whether one approach is to be preferred over another. Konolige (1986, ch. 6) discusses a similar question.

There are a number of requirements that one would like a knowledge representation formalism to meet. The first is ease and naturalness of expression. On this criterion, modal epistemic logic is preferable to reified epistemic logic. As section 3.2 illustrated, reified propositions quickly become very complicated, especially when one allows for varying domains of individuals. As a consequence, writing formulas to represent some piece of epistemic information quickly becomes very difficult.

A second requirement is computational efficiency. At first sight, it might seem that the reified approach is superior. After all, it uses the language for classical predicate calculus, and one can use a theorem prover for this logic. Moreover, applying an inference rule is always faster in such a theorem prover than in a modal theorem prover: in order to determine whether two propositions $\phi \vee \phi'$ and $\neg\chi \vee \chi'$ resolve we merely need to establish that ϕ and χ unify with some substitution σ. However, in a modal theorem prover, in addition to ensuring that ϕ and χ unify, we would also have to ensure that they do so *in the same world.* Clearly, this entails a further test, and hence the basic resolution step is more expensive. The relative inefficiency of modal resolution becomes even clearer when we associate term-indices with the terms in each proposition.

But there are two arguments that plead against reified theorem provers. First, reified sentences soon become very complicated. The complicated character of reified epistemic logic, and the resulting increase in the search space, quickly offsets the efficiency gain due to the faster resolution step. This point is reinforced by the second argument against the alleged computational superiority of reified epistemic logic. A modal epistemic theorem prover will contain hard-wired routines for

determining whether the world-indices associated with some formulas unify. In a pure reified theorem prover, this is not the case. For example, all the information that is expressed in world-indices will have to be explicitly expressed in terms of the R predicate. Given that hard-wired special purpose routines are bound to be faster than general routines, this suggests that reified theorem provers will in fact be slower.

A final criterion for comparing different knowledge representation formalisms is expressive power. Here reified epistemic logics are superior. First, you can quantify over (names of) propositions, and hence you can express things that cannot be expressed in modal epistemic logic. An example is *Nigel believes everything that Sam believes.* While this might seem a somewhat unnatural sentence, one can think of examples of systems in which one may exactly want to say this. Consider for example a system that represents the knowlege of different domain experts. Systems of this kind may need to be told that for certain problems it should use the beliefs of one expert in preference to those of another. Hence, for the purposes of solving this particular kind of problem, it should believe everything that the preferred domain expert believes.

Another advantage of reified approach, as we shall see in section 5, is that it provides a more natural way to cope with the problem of logical omniscience. Reified epistemic logic can avoid this problem because the features of a possible world semantics that are responsible for the problem of logical omniscience are explicitly formulated in some axioms. Thus, in section 2.2, we saw that logical omniscience arise partly because of the following axiom:

$$(B\phi \ \& \ B(\phi \rightarrow \chi)) \rightarrow B\chi.$$

However, because of the nature of possible worlds, this axiom is forced upon you in modal epistemic logic. However, in reified epistemic logic, you can simply drop the reified counterpart of this axiom and formulate explicitly as axioms the inference rules which you believe the agents to use. I return to this point in section 5.

The conclusion from the comparison between reified and modal epistemic logics is that reified logics have the edge in terms of expressive power, whereas modal epistemic logics are preferable on computational grounds. Thus we see another example of the trade-off of expressive power and computational efficiency described in Levesque and Brachman (1985).

4 Autoepistemic logic

Epistemic logic is aimed primarily at reasoning about the beliefs of another agent. However, a lot of reasoning about knowledge and belief involves an agent introspecting on his or her own knowledge base. Moore (1985b, 1988) calls this type of reasoning *autoepistemic reasoning*, and he has developed a logic designed specifically for this type of introspective reasoning. The example of autoepistemic reasoning given in section 1 concerned the reasons that I have for believing that I have no older brothers: if I had any older brothers, then I would know about them; introspection of my knowledge base shows me that I do not know of any older brothers; hence, I can conclude that I do not have any.

As Moore points out, one of the interesting aspects of autoepistemic reasoning is the fact that it is nonmonotonic. An autoepistemic reasoner may draw certain conclusions about the world because of the absence of certain pieces of information from its knowledge base. When these pieces of information are added at a later stage, the original inference is no longer warranted. Thus, adding further premises to the set of one's beliefs may mean that inferences that were drawn earlier are no longer correct. In the above example of autoepistemic reasoning, if I receive new information which suggests that I do have an older brother, then the original autoepistemic inference based on the absence of this piece of information from the knowledge base is no longer warranted.

Moore points out that the reason for the nonmonotonicity of autoepistemic reasoning is different from the reason behind the nonmonotonicity of default reasoning. Moore defines default reasoning as the drawing of plausible inferences from less-than-conclusive evidence in the absence of information to the contrary. A default, such as birds typically fly, does not provide a cast-iron

guarantee that Tweety flies. Rather, it provides less-than-conclusive evidence, and the inference that Tweety flies can only be regarded as a plausible inference, rather than as a logically valid inference. Default reasoning is *defeasible*, and the original plausible inference may simply turn out to be wrong. However, autoepistemic reasoning is not defeasible: if you really believe that if you had any older brothers, then you would know them, then you cannot believe that you do not have any older brothers, and at the same time accept that you have an older brother. The reason for the nonmonotonic character of autoepistemic reasoning lies in its *context-sensitivity*. An agent draws autoepistemic inferences on the basis of the explicit beliefs that it has about the world. As these beliefs change, the autoepistemic inferences that are allowed change. A number of authors, e.g. Konolige (1988b) and Marek and Truszczyński (1989), further explore the relationships between logics for autoepistemic and default reasoning.

Having introduced autoepistemic reasoning, Moore then proposes an autoepistemic logic. Moore's system is restricted to propositional logic. Since autoepistemic logic relies on a set of propositions that are believed by the reflecting agent, we need an autoepistemic theory T. Given an autoepistemic theory T formulated in some language $\mathscr{L}$, Moore then defines an autoepistemic interpretation I as an assignment of truth values to the sentences of $\mathscr{L}$, which conforms to the standard truth recursion for propositional logic (for example, I should assign *true* to $\phi \,\&\, \chi$ only if it assigns *true* to both ϕ and χ), and such that $B(\phi)$ is true in I if and only if $\phi \in T$. The latter clause of course also ensures that if $\phi \notin T$, then $B\phi$ is false, and $\neg B\phi$ is true. We call an autoepistemic interpretation I an autoepistemic model of T if I also assigns *true* to all the sentences in T. Thus, in an autoepistemic model, all the sentences in the belief set T of the agent turn out to be true in the real world as well.

A particularly interesting class of autoepistemic theories are the so-called *stable* autoepistemic theories. A stable autoepistemic theory T meets the following three conditions:

- if $\phi, \ldots, \phi' \in T$ and $\phi, \ldots, \phi' \vdash \chi$, then $\chi \in T$,
- if $\phi \in T$, then $B\phi \in T$,
- if $\phi \notin T$, then $\neg B\phi \in T$.

Autoepistemic theories that meet these requirements are called "stable" because no amount of reasoning, logical or autoepistemic, will add anything to them. The notion of stability can then be defined to define the notion of a *stable expansion*: an autoepistemic theory T is a stable expansion of a set of premises A if and only if T equals the set of ordinary logical consequences of the set $A \cup \{B\phi \mid \phi \in T\} \cup \{\neg B\phi \notin T\}$. Moore considers stable expansions interesting because they represent the beliefs of an ideally rational agent.[6] Also, stable expansions have attractive technical properties, e.g. if T is a stable expansion, then if $B\phi \in T$, then $\phi \in T$, and if $\neg B\phi \in T$, then $\phi \notin T$. Stable expansions also turn out to be important in some meta-logical results, such as soundness and completeness, which need not concern us here.

Autoepistemic logic suffers from a number of shortcomings. First, at the moment only propositional autoepistemic logic has been defined. Of course these results in severe expressive limitations. One really needs autoepistemic predicate logic. However, it is not clear how to generalize this definition to autoepistemic predicate logic.

Second, the reader will have noticed that logical omniscience becomes a problem in stable expansions: if $B\phi$ and $B(\phi \rightarrow \chi)$ are both true, then both $\phi \in T$ and $(\phi \rightarrow \chi) \in T$. Because T is a stable expansion, it will be closed under logical deduction, and therefore $\chi \in T$. It follows from that of course that $B\chi$ is true. However, Moore would dismiss this problem, as stable expansions are meant to describe the reasoning of ideal rational agents, and ideal rational agents do not suffer from resource limitations.

A final problem with autoepistemic logic is that, as far as I know, there are no theorem provers. Moore (1988) suggests a possible theorem proving strategy, which involves enumerating all the

[6] Readers familiar with for example Reiter's default logic will recognize a similarity between Moore's stable autoepistemic theories and the extensions in default logic.

members of T. Clearly, this strategy is at best computationally infeasible, and, when the underlying non-autoepistemic logic is undecidable, impossible. Jackson and Reichgelt (1988, 1989b) improve on this by adding a further inference rule to their modal theorem prover, which allow the construction of (parts of) stable expansions as the need arises. Konolige (1988a) also addresses this problem.

5 Logical omniscience

Both epistemic logic and autoepistemic logic raise a number of philosophical problems. The most important problem is the problem of *logical omniscience*. In section 2.2, I distinguished between several subproblems. For example, one aspect of the problem of logical omniscience is the problem of *consequential closure*: an agent is assumed to believe all the logical consequences of its beliefs. Other problems are the fact that an agent is assumed to believe all valid sentences; that if it believes a contradiction, then it is assumed to believe all other sentences in the language as well; and that if an agent believes some sentence ϕ, then it believes all sentences that are logically equivalent to ϕ as well.

There are a number of proposals in the AI literature for avoiding the problem(s) of logical omniscience, at least in epistemic logic. Before I discuss these, I want to look briefly at the various reasons why logical omniscience arises as a problem. The various attempts at solving logical omniscience can be seen as dealing with the different reasons for it. Fagin and Halpern (1985) distinguish between four sources.

- *Lack of awareness.* The agent may simply be unaware of some concept and therefore cannot have any beliefs about it. The standard example is the Bantu tribesman who has no beliefs about whether the price of personal computers is rising because he is not aware of the concept of personal computers.
- *Resource-boundedness.* An agent does not have the computational resources to derive all the logical consequences of its beliefs.
- *Lack of inference rules.* An agent may simply not be aware of certain inference rules. For example, someone may not know that $\phi \rightarrow \chi$ and $\neg\chi$ imply $\neg\phi$.
- *Limited focus of attention.* Agents may simply fail to put sentences that were derived in different contexts together. In particular, although the beliefs that the agent will assent to in one context may always be consistent, there may be inconsistencies in the total set of beliefs. For example, an agent may deny racial prejudices when they are discussed in the context of academic debate, while behaving according to them when it comes to everyday dealings with members of other races.

Having discussed the different reasons why logical omniscience is a problem, let us now turn to some of the proposals that have been made to avoid the problem. One can distinguish between two approaches, a *syntactic* and a *semantic* one.

5.1 Syntactic approaches

A first line of attack on the problem of logical omniscience is what Levesque calls the *syntactic* approach. In this approach, the structures that are used to model belief include syntactic entities such as sentences in some language, and possibly inference rules. An early example of the syntactic approach in AI is provided by Moore and Hendrix (1979). A more sophisticated example is Konolige's (1986a) *deduction model of belief.*

The basic idea underlying Konolige's system is to restrict the inference rules that can be used to derive new beliefs from old beliefs. He argues that the beliefs of some agent can be analyzed in terms of the notion of a *deduction structure.* A deduction consists of a set of sentences $\mathscr{B}$ in a non-modal first-order language, called the *base set*, and a set of inference rules $\mathscr{R}$, such as modus ponens. The base set contains the basic beliefs of some agent, whereas the inference rules are the rules that it can use to derive new sentences. Given a set of inference rules $\mathscr{R}$, we can define a relation $\vdash_{\mathscr{R}}$ which holds between a set of sentences S and a sentence ϕ just in case ϕ can be derived from S by using only the inference rules in $\mathscr{R}$. We can use this relation to define the belief set, bel($\langle\mathscr{B}, \mathscr{R}\rangle$), the set of believed

sentences generated in a deduction $\langle \mathscr{B}, \mathscr{R} \rangle$, as $\{\phi \mid \mathscr{B} \vdash_{\mathscr{R}} \phi\}$. Konolige then defines the semantics of the belief operator as follows:

$B\phi$ is true in a deduction structure $\langle \mathscr{B}, \mathscr{R} \rangle$ iff $\phi \in \text{bel}(\langle \mathscr{B}, \mathscr{R} \rangle)$.

As can be seen from the above description, Konolige assumes that the entire set of beliefs of some agent can be obtained by exhaustively applying the inference rules to the base set and the propositions that result from earlier rule applications. The set of beliefs generated in some deduction structure is closed under the inference rules that are part of the deduction structure. This might seem to leave us with the problem of logical omniscience but, as Konolige points out, this form of deductive closure is weaker than the consequential closure that results from a possible-world semantics for epistemic logic. The reason is that the set of inference rules in a deduction model does not have to be logically complete. In other words, there may be sentences that follow from the base set of beliefs $\mathscr{B}$ if we use the inference rules of classical predicate calculus, but which do not follow if we use the restricted set of inference rules $\mathscr{R}$. In fact, the propositions that are valid in the Kripke models of modal epistemic logic are also true in any deduction structure whose set of inference rules is complete with respect to classical predicate calculus.

Konolige's approach deals with only some of the sources of the problem of logical omniscience. Because the base set of sentences need not include for every atomic sentence ϕ either ϕ or $\neg\phi$, it can deal with lack of awareness. Similarly, because the set of inference rules in a deduction structure need not be complete, it can also cope with a lack of inference rules. Finally, Konolige does not insist that the base set is consistent. So, his system can also cope with the problem of inconsistent beliefs. The danger that the agent believes every sentence in the language, may be avoided because the incompleteness in the inference rules may block the derivation of other sentences.

Although Konolige's deduction model of belief goes some way towards solving the problem of logical omniscience, it does not provide a full answer. The problem of consequential closure is merely replaced by the weaker assumption that the agent's beliefs are closed under an incomplete set of inference rules. So, we still face the problem of deductive closure, and one could argue that if the consequential closure was intuitively unacceptable because it did not take into account the resource limitations of an agent, then similar arguments apply to weaker forms of deductive closure.

Konolige himself sketches one way of avoiding this problem. He proposes to add to each sentence ϕ some indication about the cost necessary to derive ϕ from the set of base sentences. So, if $D(n)$ represents the fact that n applications of inference rules were necessary, then we can redefine *modus ponens* as

$$(D(k) \,\&\, \phi), (D(l) \,\&\, (\phi \rightarrow \chi)) \vdash (D(k + l + 1) \,\&\, \chi).$$

As it stands, this proposal will not work, if we have conjunction elimination (from $(\phi \,\&\, \chi)$ conclude ϕ and χ) as an inference rule. After all, we can always infer χ from $D(k + l + 1) \,\&\, \chi$. So, even if $k + l + 1$ is larger than the largest allowable cost, and the inference of χ should be blocked, then we can still infer χ. The underlying reason is that D really is a two-place predicate which should take a sentence ϕ and some integer n representing the cost of deriving ϕ as arguments. But since Konolige uses a non-reified language, it is impossible to have predicates that take sentences as their arguments.

The reified approach discussed in section 3.2 does allow for such predicates, and it is relatively straightforward to reformulate Konolige's deduction model of belief, and his sketch of dealing with the problem of resource limitations in a reified style. One way of doing this would be to introduce a further predicate *bel*, which corresponds to Konolige's belief set. We then give a number of axioms for *bel*. Suppose that we have a deduction structure, whose base set is $\{\phi, \chi, \phi \rightarrow \pi\}$ and which has modus ponens as its only inference rule. This gives the following axioms, where, as before, if ϕ is a sentence, then $[\phi]$ denotes the name of that sentence:

$bel([\phi]) \,\&\, D(0, [\phi])$,

$bel([\chi] \,\&\, D(0, [\chi])$,

$$bel([\phi \rightarrow \pi]) \,\&\, D(0, [\phi \rightarrow \pi]),$$

$$(\forall x)(\forall y)((bel(x) \,\&\, D(n, x) \,\&\, bel([x \rightarrow y]) \,\&\, D(m, [x \rightarrow y])) \rightarrow (bel(y) \,\&\, D(n + m + 1, y))).$$

We can now redefine the meaning of the belief-operator to take this new predicate D into account. Suppose that we want to impose an upper limit of i on the amount of reasoning effort, then we define the following axiom:

$$true([B\phi]) \equiv (\exists n)(n < i \,\&\, bel(\phi) \,\&\, D(n, [\phi])).$$

It is left to the reader to check that if we impose an upper limit of 1, then the proposition $B\pi$ would not be true in the above deduction structure.

The reified style solution to the problem of logical omniscience is a generalization of Konolige's proposals, and therefore inherits from it the ability to deal with the lack of awareness problem, the lack of certain inference rules, and the inconsistency problem. In addition, it can also deal with the resource limitation problem.

However, there are a number of problems with this reified approach and the syntactic approach in particular. First, in this reified treatment the original intuition that we could reason about the meanings of sentences in modal epistemic logic in reified epistemic logic is lost. We can now express things in reified epistemic logic that cannot be expressed in modal epistemic logic. On the other hand, one might argue that modal epistemic logic leads to the problem of logical omniscience, and if we want to avoid it, then we will have to replace modal epistemic logic, and the original reified system that was directly based on it, by something else.

A second problem is computational. In order to deal with the various problems satisfactorily, a large number of additional axioms need to be added. However, this will considerably complicate the search for a proof of a given formula. It is possible that the resulting system will be prohibitively slow, and that in order to implement a workable system, one needs to hardwire some of the axioms that were formulated before.

In addition to these specific criticisms, there are also more general problems. Levesque (1984) for example argues that whereas the possible world semantics of modal epistemic logic is in a sense too coarse grained (in that we cannot distinguish between different but logically equivalent propositions), syntactic approaches are in a sense too fine grained in that they force you to distinguish between propositions that should not be distinguished. For example, Levesque points out that under the syntactic approach there is no reason why an agent should believe $\phi \,\&\, \chi$ if it believes $\chi \,\&\, \phi$. However, it is not clear whether this is actually a problem. Fagin and Halpern (1985) argue, and I would agree with them, that this is not a shortcoming, but is in fact a positive feature: there is indeed no reason why an agent should believe $\phi \,\&\, \chi$ if it believes $\chi \,\&\, \phi$. Alternatively, if this argument fails to convince, one may argue that Levesque simply points out that one needs to add the following axiom:

$$bel([x \,\&\, y]) \equiv bel([y \,\&\, x]).$$

Another argument against the syntactic approach is that one can ask oneself to what extent one really gives a *logic* of reasoning about knowledge and belief. The resulting system is in reality a description of resource limited reasoning in a fairly complicated first-order language. The resulting system therefore strikes me more as a formalization exercise than as the development of a truly new logic, in the sense that modal epistemic logic is a new logic. On the other hand, this is perhaps not a bad thing. The above could be seen as an attempt to formalize a theory of reasoning about knowledge and belief in classical predicate logic, similar in a sense to the attempt by Hayes (1985) to formalize naive physics in this representation language.

5.2 Semantic approaches

In contrast to the syntactic approach the semantic solutions try to deal with with the problem of logical omniscience by proposing more complicated model structures. Levesque's (1984) logic of

explicit and implicit belief is a clear example of this strategy. Levesque draws a distinction between *implicit* and *explicit* belief. Explicit beliefs are defined as those beliefs that the agent actively holds. One can say that an agent explicitly believes ϕ if it answers "yes" when asked if it believes ϕ. Explicit beliefs are therefore somewhat similar to the base set of beliefs in Konolige's deduction model of belief (see 5.1). Implicit beliefs, on the other hand, are those beliefs that logically follow from what the agent believes, whether the agent realizes it or not. An implicit belief is everything that would be true if the world was exactly as the agent believed it to be.

Levesque argues that the possible world semantics may be appropriate for the analysis of implicit belief. The problem of logical omniscience and the failure to take resource limitations into account only apply when we try to analyse what the agent explicitly believes. If we restrict ourselves to implicit belief, then resource limitations are irrelevant. Levesque therefore proposes to maintain the possible world analysis for implicit belief, and proposes a new logic for explicit belief.

Levesque's logic of explicit belief relies on the notion of a *situation*, which is derived from Barwise and Perry (1983). Roughly, a situation is a partial world. In the model theory for modal epistemic logic, a possible world supports the truth of every sentence in the language. A situation, on the other hand, supports the truth or falsity of only some sentences of the language, and may fail to support the truth or falsity of some sentences. We say that a situation s is compatible with a possible world w if w supports the truth of each sentence whose truth is supported by s, and the falsity of each sentence whose falsity is supported by s. Levesque calls situations which are not compatible with any possible world (because they for example support both the truth and falsity of some sentence ϕ), *incoherent*.

Levesque's formal definition starts with the definition of a *model structure*, which consists of a set of situations $\mathcal{S}$, a subset of $\mathcal{S}$, called $\mathcal{B}$, and two functions $\mathcal{T}$ and $\mathcal{F}$. $\mathcal{B}$ is the set of situations that, as far as the agent is concerned, could correspond to the actual world. $\mathcal{T}$ and $\mathcal{F}$ assign to each atomic sentence in the language a set of situations in $\mathcal{S}$ that support the truth and falsity of the proposition respectively. Levesque then defines two entailment relations $\models_T$ and $\models_F$. $\models_T$ ($\models_F$) is true of a situation s and a sentence ϕ if s supports the truth (falsity) of ϕ. The definitions for the non-modal propositions are relatively straightforward:

1 $s \models_T \phi$ iff $s \in \mathcal{T}(\phi)$ for atomic ϕ
$s \models_F \phi$ iff $s \in \mathcal{F}(\phi)$ for atomic ϕ.
2 $s \models_T \neg\phi$ iff $s \models_F \phi$
$s \models_F \neg\phi$ iff $s \models_T \phi$.
3 $s \models_T (\phi \,\&\, \chi)$ iff $s \models_T \phi$ $s \models_T \chi$
$s \models_F (\phi \,\&\, \chi)$ iff $s \models_F \phi$ or $s \models_F \chi$.

Levesque introduces two different belief operators corresponding to his notions of explicit and implicit belief. The explicit belief operator is B. Syntactically, it is slightly different from the belief operator in modal epistemic logic in that it cannot be recursively applied. So, $B\phi$ is only well-formed if ϕ does not itself contain any occurrences of B. In order to define the meaning of $B\phi$, recall that a model structure contains a subset $\mathcal{B}$ of the set of all situations $\mathcal{S}$ which represent the situations that can be actual as far as the beliefs of the agent are concerned. $\mathcal{B}$ plays an essential role in the definition of $B\phi$:

$$s \models_T B\phi \text{ iff for every } s' \in \mathcal{B},\ s' \models_T \phi$$

$$s \models_F B\phi \text{ iff not } s \models_T B\phi.$$

It is important to realize that there are no restrictions on the functions $\mathcal{T}$ and $\mathcal{F}$ in the notion of a model structure. For example, a situation s may be a member of both $\mathcal{T}(\phi)$ and $\mathcal{F}(\phi)$ for some atomic proposition ϕ. Similarly, it is possible that some situation s is not a member of either $\mathcal{F}(\phi)$ or $\mathcal{F}(\phi)$. Levesque can solve the various problems of logical omniscience.

- The problem of deductive closure is avoided. Assume that $s \models_T \phi$ and that $s \models_T (\phi \rightarrow \chi)$. The latter implies that $s \models_T (\neg\phi \vee \chi)$, and hence either $s \models_T \neg\phi$ or $s \models_T \chi$. From $s \models_T \neg\phi$, we can now conclude $s \models_F \phi$. Clearly, in modal epistemic logic, the latter would contradict $s \models_T \phi$, and we

would be able to conclude $s \models_T \chi$. However, in Levesque's logic, this inference is not warranted. $s \models_T \phi$ and $s \models_F \phi$ can be true at the same time. For example, if ϕ is atomic, then s may be a member of both $\mathcal{T}(\phi)$ and $\mathcal{F}(\phi)$.

- The problem that if the agent's beliefs are inconsistent, then it believes all propositions is also avoided. It is possible that for all situations s in $\mathcal{B}$ it is true that $s \in \mathcal{T}(\phi)$ and $s \in \mathcal{F}(\phi)$ for some atomic proposition ϕ, so that $B(\phi \,\&\, \neg\phi)$. However, it does not follow from this that every situation in $\mathcal{B}$ also supports the truth of χ. Therefore $B\chi$ need not be true.

Finally, an agent does not have to believe explicitly every valid sentence in the language. For example, consider the sentence $B(\phi \vee \neg\phi)$ for some atomic sentence ϕ. If $\mathcal{B}$ contains some situation s such that neither $s \in \mathcal{F}(\phi)$ nor $s \in \mathcal{F}(\phi)$, then clearly not $s \models_T (\phi \vee \neg\phi)$ and $B(\phi \vee \neg\phi)$ is not true.

In addition to the explicit belief operator B, Levesque also uses the implicit belief operator L. As the B operator, L cannot be recursively applied. In order to define its semantics, Levesque defines a way of extending a situation s into a set of possible worlds $\mathcal{W}(s)$. The construction is relatively straightforward. A situation s assigns truth values to some atomic sentences in the language. Every possible world which assigns the same truth values to those atomic sentences is then in $\mathcal{W}(s)$. Note incidentally that if we have incoherent situation s, i.e. a situation which assigns *true* to both ϕ and $\neg\phi$, then $\mathcal{W}(s)$ is empty.

This construction can obviously be extended to sets of situations. So, $\mathcal{W}(\mathcal{B})$ is the set of all possible worlds that can be constructed from the situations in $\mathcal{B}$. Levesque then defines the semantics of the L-operator as:

$s \models_T L\phi$ iff for every $s' \in \mathcal{W}(\mathcal{B})$, $s' \models_T \phi$

$s \models_F L\phi$ iff not $s \models_T L\phi$.

The reader can check that given this semantics $B\phi \rightarrow L\phi$ is valid. Thus, an agent which explicitly believes that ϕ also believes it implicitly.

Although Levesque goes some way towards solving the problem of logical omniscience, his treatment is restricted to propositional logic. So, sentences such as $(\exists x)B\phi x$ or $L(\exists x)\phi x$, are not allowed. Also, Levesque's system does not allow nested beliefs. Lakemeyer (1986) addresses the former problem, while Fagin and Halpern (1986) address the latter.

Another criticism that one can level against Levesque's system concerns the notion of a situation. Intuitively, a situation is described as a partial possible world. However, whereas possible worlds are always assumed to be consistent, in the sense that they never assign both the truth values *true* and *false* to the same sentence ϕ, a situation is allowed to support both the truth and falsity of ϕ. Now, the notion of an incoherent situation seems to me problematic from a philosophical point of view. After all, a situation is described as a partial possible world, and possible worlds are always consistent. Therefore, a situation would always have to be consistent as well. If Levesque wants to use the notion of an incoherent situation, then it seems to me that he owes us a better intuitive description of what a situation is.

Earlier, I mentioned Fagin and Halpern's distinction about the different sources responsible for the logical omniscience problem. Fagin and Halpern point out that Levesque's logic cannot really deal with resource limitations. In their argument they use the notion of awareness: an agent is said to be aware of a proposition ϕ, formally $A(\phi)$, if and only if $B(\phi \vee \neg\phi)$. They show that in Levesque's system any valid non-modal sentence ϕ must also be explicitly believed if the agent is aware of all the atomic propositions that ϕ is composed of. Recall that $B\phi$ is true in a model structure if $s \models_T \phi$ is true for all situations s in $\mathcal{B}$. However, as one can see from the definition of $s \models_T \phi$ and $s \models_F \phi$, the truth value of a non-modal sentence ϕ in some situation s is completely determined by the truth values in s of the component parts of ϕ. Because we know by hypothesis that $B(\chi \vee \neg\chi)$ for all component parts of ϕ, we knew that for all situations s in $\mathcal{B}$, either $s \in \mathcal{T}(\chi)$ or $s \in \mathcal{F}(\chi)$. Now, because ϕ is valid in propositional logic, it does not matter what the truth values of its component parts are: ϕ will always be true in s. Hence, $B\phi$ must be true as well. Fagin and Halpern argue that this result is not

really satisfactory when we want to deal with resource limitations. The fact that an agent is aware of every atomic sentence in some valid sentence ϕ does not mean that it also explicitly believes ϕ. It may simply take a very long time to compute the truth value of ϕ, and a resource-limited agent might not be able to do so. In addition, a similar argument can be used to show that Levesque's model of explicit belief does not allow for a lack of inferences rules. If an agent is aware of all the atomic sentences in ϕ and χ, and $\phi \vdash \chi$, then $B\phi$ logically implies $B\chi$.

Fagin and Halpern develop a whole family of logics of awareness to avoid these problems. The approach can best be described as an amalgamation of the semantic and the syntactic approach. Fagin and Halpern return to the standard possible semantics, and do not use situations, as Levesque does. Rather, they define a function $\mathscr{A}$ which associates with each possible world a set of sentences that the agent is aware of. The important point is that this set of sentences is completely arbitrary, and is not restricted to atomic propositions. Moreover, it does not have to be closed under logical deduction. They also introduce a new modal operator A, $A\phi$ is true in some world w if and only if $\phi \in \mathscr{A}(w)$. The implicit belief operator L is defined in the standard way: $L\phi$ is true in some world w if and only if ϕ is true in all worlds accessible from w. Finally, the explicit belief operator B is defined as: $B\phi$ is true in a world w if and only if $\phi \in \mathscr{A}(w)$ and ϕ is true in all worlds accessible from w. The reader will have noticed that, given this definition, we have the following equivalence

$$B\phi \equiv (L\phi \,\&\, A\phi).$$

Fagin and Halpern's treatment of explicit and implicit belief has a number of limitations. First, $B(\phi \,\&\, \neg\phi)$ can never be true, simply because $L(\phi \,\&\, \neg\phi)$ can never be true. Hence, an agent cannot have inconsistent beliefs. Second, although deductive closure is avoided, the way in which this is done is rather counter-intuitive. Assume that both $B\phi$ and $B(\phi \rightarrow \chi)$ are true. Then it follows from the definition of B that both $A\phi$ and $A(\phi \rightarrow \chi)$ are true, as well $L\phi$ and $L(\phi \rightarrow \chi)$. Given the definition of L, the latter two propositions imply that $L\chi$ is true as well. Therefore, the only way to falsify $B\chi$, which after all is equivalent to $L\chi \,\&\, A\chi$, is to say that $A\chi$ is false. But this leads to the counter-intuitive consequence that an agent can be aware of $\phi \rightarrow \chi$ without being aware of χ. Fagin and Halpern propose a further logic that gets around the first problem, but they provide no answer to the second problem.

5.3 A comparison of the syntactic and the semantic approach

In the previous two subsections, I discussed some of the ways in which AI researchers have tried to avoid the problem of logical omniscience. A number of other attempts have been made (see e.g. papers in Halpern, 1986), but it is beyond the scope of this paper to discuss all these alternative solutions in detail. In this section I want to briefly compare the two different styles of solution. The criteria that I use are similar to those used in section 3.3 to compare reified and modal theorem provers.

The first criterion concerns implementation. It is relatively straightforward to see how one might implement syntactic approaches. Consider for example Konolige's deduction model. Konolige implements a deduction structure as a natural deduction style theorem prover from which certain inference rules have been omitted. However, it is much harder to see how one could implement theorem provers based on the semantic approach. Although Levesque does give a proof theory for his logic, which could in principle be used as a specification of a theorem prover, the proof theory is formulated as a (large) set of axioms, and a small set of inference rules, and it is an open question whether one can implement an efficient theorem prover on the basis of this proof theory.

A second criterion is naturalness. I criticized the different semantic approaches because they relied on intuitively unclear notions. For example, Levesque uses incoherent situations to account for inconsistent beliefs. But it is not clear what an incoherent situation is. In general, in the semantic approach one tries to explain the meaning of belief in terms of some model-theoretic entity, which corresponds to, or perhaps represents, a possible state of affairs in (part of) the world. But the world can never be incoherent in the sense that both ϕ and $\neg\phi$ are true at the same time. I therefore doubt

that an intuitively satisfactory semantic solution to the problem of logical omniscience is possible.

The various syntactic solutions on the other hand also have problems. For example, in Fagin and Halpern's system, the problem of consequential closure is avoided by allowing for the possibility that an agent is aware of $\phi \rightarrow \chi$ without being aware of ϕ or χ. In Konolige's deduction model of belief, on the other hand, one still has to live with some form of closure under deduction, even if it is slightly weaker than consequential closure. However, as I have shown in 5.1, it is possible to avoid this problem in Konolige's logic. Therefore, I tentatively conclude that a syntactic approach to the problem of logical omniscience is preferable.

6 Conclusion

Reasoning about knowledge and belief is likely to become more important in AI as we move to more sophisticated problem solving architectures, as well as more "user-friendly" human computer interfaces, which allow the human user to use natural language, or the machine to tailor its response to the beliefs that it ascribes to the user. In this paper, I have discussed the use of epistemic logic to this end, and I compared two different ways of writing theorem provers for epistemic logic. I also discussed autoepistemic logic, a logic aimed specifically at introspective reasoning. I pointed to some of the intuitive problems that arise when one uses epistemic or autoepistemic logic to analyze reasoning about knowledge and belief. In particular I mentioned the problem of logical omniscience, and I discussed a number of approaches in AI to avoid these difficulties.

One can of course argue that the problem of logical omniscience should simply be ignored, and that we should try to model the reasoning of ideal rational agents. Moore for example takes this line in his discussion of autoepistemic logic. However, as I pointed out in section 2.2, it seems to me that this answer is not satisfactory in the context of AI. AI reasoners should be aware of the resource limitations of other agents, and a notation of valid consequence which ignores these limitations will therefore lead to counterintuitive and even false conclusions. I therefore believe that the search for systems that avoid the problem of logical omniscience should be continued. It must be admitted however that the proposed solutions to the problem of logical ominiscience all yield complicated and (in some cases unwieldy) formal systems. The question therefore arises whether such logical systems can be used as the foundation for the implementation of a system that is capable of reasoning about the beliefs of other agents. This of course is a legitimate question, and one that would seem to require further research. However, even if it turns out that the logical systems discussed in this paper cannot be used directly to endow a system with this ability, I nevertheless hope that the work described will prove useful as an abstract specification of the properties that the ultimate implementation must process.

References

Abadi, M and Manna, Z, 1986. "Modal theorem proving" *CADE-8*, 172–189.

Barwise, J and Perry, J, 1983. *Situations and attitudes*, Cambridge, Massachusetts: Bradford Books [In this book Barwise and Perry introduce a new approach of the study of the semantics of natural language, called *situation semantics*. Levesque uses similar ideas in his logic of implicit and explicit belief]

Boolos, G and Jeffrey, R, 1974. *Computability and logic*, Cambridge: Cambridge University Press [One of the standard introduction to logic. This book is very much written with the theory of computation in mind]

Chellas, B, 1980. *Modal logic*, Cambridge: Cambridge University Press [A good introduction to modal logic, restricted however to propositional modal logic]

Fagin, R and Halpern, R, 1985. "Belief, awareness, and limited reasoning: Preliminary report" *IJCAI-85*, 491–501

Geissler, Ch and Konolige, K, 1986. "A resolution method for quantified modal logics of knowledge and belief" In: Halpern, ed (1986)

Halpern, J, ed, 1986. *Theoretical aspects of reasoning about knowledge*, New York: Morgan Kaufmann [The proceedings of a conference on reasoning about knowledge and belief. Recommended further reading]

Halpern, J and Moses, Y, 1985. "A guide to the modal logics of knowledge and belief: Preliminary draft" *IJCAI-9*, 480–490 [A short introduction to epistemic logic, restricted to the propositional case]

Hayes, P, 1977. "In defence of logic" *IJCAI-5*, 559–565

Hayes, P, 1985. "The second naive physic manifesto" In: Hobbs and Moore, eds (1985)

Hintikka, J, 1962. *Knowledge and belief*, New York: Cornell University Press

Hintikka, J, 1969. "Semantics for propositional attitudes" In: Davis, J, et al, eds, *Philosphical Logic*, Dordrecht: Reidel

Hobbs, J and Moore, R, eds, 1985. *Formal theories of the common sense world*, Norwood, New Jersey: Ablex

Hughes, G and Cresswell, M, 1968. *An introduction to modal logics*, London: Methuen [Probably still the best introduction to modal logic]

Jackson, P and Reichgelt, H, 1988. "A modal proof method for doxastic reasoning in incomplete theories" *ECAI-88*, 480–484

Jackson, P and Reichgelt, H, 1989a. "A general proof method for arbitrary modal predicate logic" In: Jackson, P,, Reichgelt, H, and Harmelen, F, van, eds, *Logic-based knowledge representation*, Cambridge, Massachusetts: MIT Press

Jackson, P and Reichgelt, H, 1989b. "A modal proof method of doxastic reasoning" In: Jackson, P, Reichgelt, H and Harmelen, F, van, eds, *Logic-based knowledge representation*, Cambridge, Massachusetts: MIT Press

Jian, Y, 1987. "A clausal form of belief. A logic programming model" *9th International Conference of the Cognitive Science Society*, 301–314

Konolige, K, 1986a. *A deduction model of belief*, London: Pitman [Based on Konolige's thesis, this book discusses both Konolige's modal epistemic theorem prover, as well as his deduction model of belief]

Konolige, K, 1986b. "Resolution and quantified epistemic logic" *CADE-8*, 199–208

Konolige, K, 1988a, "Hierarchic autoepistemic theories for nonmonotonic reasoning: Preliminary Report" SRI Technical Note 446

Konolige, K, 1988b. "On the relation between default and autoepistemic logic" *Artificial Intelligence*, **35**, 343–382.

Kripke, S, 1963. "Semantic considerations on modal logic" *Acta Philosophica Fennica*, **16**, 83–94

Kripke, S, 1972. "Naming and necessity" In: Davidson, D and Harmon, G, eds, *Semantic of natural language, 2nd ed*, Dordrecht: Reidel

Lakemeyer, G, 1986. "Steps towards a first-order logic of explicit belief" In: Halpern, ed

Levesque, H, 1984. "A logic of implicit and explicit belief" *AAAI-84*, 198–202

Levesque, H and Brachman, R, 1985. "A fundamental tradeoff in knowledge representation and reasoning (revised version)" In: Brachman, R and Levesque, H, eds, *Readings in knowledge representation*, Los Altos, California: Morgan Kaufmann

Mates, B, 1972. *Elementary logic* (2nd ed), Oxford: Oxford University Press [A good introduction to first-order logic]

Marek, W and Truszczyński, M, 1989. "Relating autoepistemic and default logics" In: Brachman, R, Levesque, H and Reiter, R, eds, *Proceedings of the First International Conference on the Principles of Knowledge Representation and Reasoning*, San Mateo, CA: Morgan Kaufmann

McArthur, G, 1988. "Reasoning about knowledge and belief: A survey" *Computational Intelligence*, **4**, 223–243 [An alternative review of much of the material discussed in this paper]

McCarthy, J, 1979. "First order theories of individual concepts and propositions" *Machine Inteligence*, **9**, 129–147

Mendelson, E, 1987. *Introduction to mathematical logic (3rd)*, Wadsworth, California: [Another worthwhile introduction to logic]

Moore, R, 1984. "The role of logic in Artificial Intelligence" SRI Tech Report 335

Moore, R, 1985a. "A formal theory of knowledge and action" In: Hoggs and Moore, eds [Moore's reified epistemic logic]

Moore, R, 1985b. "Semantical considerations on non-monotonic logic" *Artificial Intelligence*, **25**, 75–94 [In this paper Moore introduces his autoepistemic logic]

Moore, R, 1988. "Autoepistemic logic" In: Smets, Ph, Mamdani, A, Dubois, D and Prade, H, eds, *Non-standard logics for automated reasoning*, London: Academic Press [A further introduction to Moore's autoepistemic logic. The book also contains three comments with a reply by Moore]

Moore, R and Hendrix, G, 1979. "Computational models of belief and the semantics of belief sentences" SRI Technical Note 187

Nilsson, N, 1982. *Principles of artificial intelligence*, Berlin: Springer [A very good introduction to automated theorem provers]

Rapaport, W, 1986. Logical foundations for belief representation" *Cognitive Science*, **10**, 371–442

Robinson, J, 1979. *Logic: form and function. The mechanization of deductive reasoning*, Edinburgh University press [A general introduction to logic, and the theory underlying the central concepts in automated theorem proving such as unification and resolution]

Stickel, M, 1987. "An introduction to automated deduction" In: Bibel, W and Jorrand, Ph, eds, *Fundamentals of artificial intelligence: An advanced course*, Berlin: Springer-Verlag

Vardi, M, ed, 1988. *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge.* Los Altos, California: Morgan Kaufmann [The proceedings of a conference on reasoning about knowledge and belief. Recommended further reading]

Wallen, L, 1986. "Matrix proof methods for modal logics" *IJCAI-10*, 917–923.