An Abstract General Model and Logic of Resource-Bounded Believers

Michael Wooldridge
Department of Computing
Manchester Metropolitan University
Chester Street, Manchester M1 5GD
United Kingdom
M.Wooldridge@doc.mmu.ac.uk

Abstract

This paper presents an abstract general model for representing the belief systems of resource-bounded reasoning agents. The intuition which underlies this new model is that it is possible to capture the key properties of many different types of belief system in a structure called a belief extension relation. The paper shows how such a relation may be derived for any system that satisfies some basic properties. The resulting formalism is simple, and yet sufficiently rich that it generalises many other frameworks for representing belief. A logic is defined, using the new model to give a semantics to belief modalities. The properties of the model and logic are discussed in detail. The paper closes with a discussion and remarks on future work.

Introduction

For many applications in Artificial Intelligence (AI), it is necessary to build systems that include symbolic representations of other systems. For example, Distributed AI (DAI) is concerned with understanding and constructing computer systems that contain multiple interacting *agents*, each of which is an AI system in its own right; it is widely accepted that to (co-)operate effectively in a multi-agent environment, agents need to be able to manipulate representations of the state and behaviour of other agents (Bond and Gasser, 1988, pp25–29). An obvious research problem is to devise knowledge representation formalisms that are suitable for this purpose. This paper contributes to the theoretical foundations of such formalisms; it considers the representation of belief in multi-agent AI systems. In this paper, the term *belief* is used to mean an agent's symbolic representation of the world it occupies, which may include other agents. (It is worth emphasising that *human* belief is *not* the object of study in this paper and in particular, no claims are made about the validity or usefulness of the model for representing human believers.)

Specifically, the paper presents a new model for representing the belief systems of resource-bounded reasoning agents. This new model is both general and abstract. It is *general* in that it can be used to represent the properties of a wide range of agents, and can be seen to generalise a number of other formalisms for representing belief. It is *abstract* in that the problem of modelling the complex and intricate reasoning processes actually used by agents is side-stepped. The idea underlying the new model is to capture the key properties of an agent's belief system in a simple, uniform framework called the *belief extension relation*; after introducing these structures, we show how it is possible to derive such a relation for any system that satisfies some simple properties. In order to conveniently represent the properties of belief systems, a logic containing belief modalities is defined; the semantics of this logic are given in terms of the new model. The new model is then compared to two other formalisms, (the deduction model (Konolige, 1986a) and normal modal logics (Halpern and Moses, 1992)), and is shown to generalise them. Some remarks on implementing the new model are then presented. The paper begins, in the following section, by reviewing previous attempts to formally model belief.

Notational Conventions: If $\mathcal{L}$ is a logical language, then we write $Form(\mathcal{L})$ for the set of (well-formed) formulae of $\mathcal{L}$. We use the lowercase Greek letters φ and ψ as meta-variables ranging over formulae of the logical languages we consider, and the uppercase Greek letter Δ as a meta-variable ranging over sets of formulae. Is S is a set, then $\wp(S)$ is the powerset of S. We use $\emptyset$ for the empty set. *Note that all proofs have been omitted due to space restrictions; they may be found in the associated technical report* (Wooldridge, 1994).

Background

The commonest technique for modelling belief is to use a modal logic with possible worlds semantics (Chellas, 1980; Halpern and Moses, 1992); this approach was first developed by Hintikka (Hintikka, 1962). Normal modal logics have properties that make them simple and interesting tools to work with, and have proved to be valuable in the formal study of belief. However, there are a number of problems associated with normal modal logics of belief. Chief among these is the so-called *logical omniscience* problem: any normal modal logic of belief predicts that (i) agents believe all valid formulae (including all propositional tautologies); and (ii) agents believe all the logical consequences of their beliefs. While logical omniscience may be acceptable in the study of theoretically perfect believers, any model of belief with this property will be unacceptable for representing *resource bounded* believers — and any realistic believer is resource bounded. Given that normal modal logics are unacceptable for representing real believers, one seems to be faced by at least two options: (i) weaken possible worlds semantics in some way, to get rid of logical omniscience; or (ii) seek an alternative semantics.

A number of attempts have been made to weaken possible worlds semantics. One of the best-known is due to Levesque (Levesque, 1984). To weaken the notion of a world, Levesque borrowed some ideas from situation semantics; situations (worlds) in his logic may assign either **true**, **false**,

or both to a proposition, and thus they do not act like propositional valuations. He then defined an 'explicit belief' operator, with a semantics given in terms of situations. The logic of explicit belief is much weaker than that of normal modal belief. Levesque's original proposal has been extended into the first-order realm by Lakermeyer (Lakemeyer, 1991). Another attempt to weaken possible worlds semantics is due to Fagin-Halpern, who developed a logic containing an operator to represent those formulae an agent is 'aware of'; the semantics of this operator are given by simply associating a set of formulae with an agent (Fagin and Halpern, 1985). The logic also contains an implicit belief operator, with a normal modal semantics. 'Explicit' belief is then defined as implicit belief plus awareness; the resulting logic of explicit belief is weaker than implicit belief. Although both of these formalisms do get rid of logical omniscience, they have been criticised for being essentially *ad hoc* and unmotivated (Konolige, 1986b). Moreover, under some circumstances, they still predict that agents have unreasonable deductive capabilities.

The second alternative is to reject possible worlds altogether, and seek an alternative semantic base. The best-known example of such work is the *deduction model of belief*, due to Konolige (Konolige, 1986a). In essence, the deduction model is an attempt to directly model the belief systems of agents in AI; the concerns of the deduction model are thus very much the concerns of this paper. Konolige models an agent's reasoning process by associating it with a set of deduction rules; if this set of rules is logically incomplete, then the agent is not logically omniscient (for details, see below). The simplicity and directness of the deduction model has led to much interest in the (D)AI community. However, the approach is not without its problems, the most significant of which is that modelling an agent's reasoning via incomplete proof is, in general, still too strong to represent resource-bounded reasoners.

Belief Models

The new structures we develop to represent belief systems are called *belief models*. A belief model representing an agent i's belief system is a pair. The first component of this pair is a set of observations that have been made about i's beliefs. These observations are expressed in some *internal language*; throughout this paper, we shall call this internal language $\mathcal{L}$. In general, the internal language may be one of rules, frames, semantic nets, or some other kind of KR formalism, but for simplicity, we shall assume that $\mathcal{L}$ is a *logical* language. Thus the first component of i's belief model is a set of $\mathcal{L}$-formulae representing observations that have been made about i's beliefs. The second component of i's belief model is a relation, which holds between sets of $\mathcal{L}$-formulae and $\mathcal{L}$-formulae. This relation is called a *belief extension* relation, and it is intended to model i's reasoning ability. We generally abbreviate 'belief extension relation' to 'b.e. relation'. Let BE_i be the b.e. relation for agent i. Then the way we interpret this relation is:

if
$\quad i$ **believes** Δ **and** $(\Delta, \varphi) \in BE_i$ (1)
then
$\quad i$ **also believes** φ.

It is through an agent's b.e. relation that we are able to make deductions about what other beliefs it has. Note that we do not require the b.e. relation to be based on logical inference. It need not be a proof relation (as in (Konolige, 1986a)); we can model agents whose 'reasoning' is not based on logical inference, as well as those that are. However, we *are* obliged to explain where an agent's b.e. relation comes from; later, we show how a b.e. relation that correctly describes the behaviour of an agent's belief system may be derived, in a principled way, for any system that satisfies some basic properties. We now formally define belief models.

Definition 1 *A belief model, b, is a pair $b = (\Delta, BE)$ where*

- $\Delta \subseteq Form(\mathcal{L})$; *and*
- $BE \subseteq (\wp(Form(\mathcal{L})) \times Form(\mathcal{L}))$ *is a countable, non-empty binary relation between sets of $\mathcal{L}$-formulae and $\mathcal{L}$-formulae, which must satisfy the following requirements:*
  1. *Reflexivity: if* $(\Delta, \varphi) \in BE$, *then* $\forall \psi \in \Delta, (\Delta, \psi) \in BE$;
  2. *Monotonicity: if* $(\Delta, \varphi) \in BE$, $(\Delta', \psi) \in BE$, *and* $\Delta \subseteq \Delta'$, *then* $(\Delta', \varphi) \in BE$;
  3. *Transitivity: if* $(\Delta, \varphi) \in BE$ *and* $(\{\varphi\}, \psi) \in BE$, *then* $(\Delta, \psi) \in BE$.

If $b = (\Delta, BE)$ is a belief model, then Δ is said to be its *base set*, and BE its b.e. relation. We now define a function *bel* which takes as its one argument a belief model, and returns the set of $\mathcal{L}$-formulae representing the *belief set* of the model.

Definition 2

$$bel((\Delta, BE)) \stackrel{\text{def}}{=} \{\varphi \mid (\Delta, \varphi) \in BE\}$$

The idea is that if b is a belief model representing some agent's belief system, then $bel(b)$ contains all those formulae which we can assume the agent believes. Before moving on, we state a lemma which captures some obvious properties of belief sets.

Lemma 1 *If $b = (\Delta, BE)$ is a belief model, then*

1. *If* $\varphi \in \Delta$, *then* $\varphi \in bel(b)$;
2. *If* $\Delta' \subseteq \Delta$, *and* $(\Delta', \varphi) \in BE$, *then* $\varphi \in bel(b)$;
3. *If* $\Delta' \subseteq bel(b)$, *and* $(\Delta', \varphi) \in BE$, *then* $\varphi \in bel(b)$.

Suppose b_i is a belief model which represents agent i's belief system. Then the interpretation of 'belief' in this paper is as follows.

$\varphi \in bel(b_i)$	—	i believes φ
$\neg\varphi \in bel(b_i)$	—	i believes $\neg\varphi$
$\varphi \notin bel(b_i)$	—	i doesn't believe φ
$\neg\varphi \notin bel(b_i)$	—	i doesn't believe $\neg\varphi$

Thus the new model is a *sentential* model of belief; see (Konolige, 1986a, pp110–118) for a discussion of such models. Note that it is possible for a belief model to represent agents that have 'no opinion' on some formula. It is also possible for a model to represent an agent that believes both a formula and its negation.

A Belief Logic

Belief models are the basic mechanism for representing belief systems. However, manipulating models directly is somewhat awkward. We therefore introduce a logic that will allow

us to represent the properties of belief models in a more convenient way. In the interests of simplicity, we shall restrict our attention to propositional languages. Adding quantification is relatively simple in terms of syntax and semantics, but poses obvious problems when developing proof methods. We assume an underlying classical propositional language, which we shall call $\mathcal{L}_0$. This language is defined over a set Φ of primitive propositions, and is closed under the unary connective '$\neg$' (not), and the binary connective '$\vee$' (or). The remaining connectives of classical logic ('$\wedge$' (and), '$\Rightarrow$' (implies), and '$\Leftrightarrow$' (iff)) are assumed to be introduced as abbreviations, in the standard way. In addition, the language contains the usual punctuation symbols ')' and '('. A model for $\mathcal{L}_0$ is simply a valuation function $\pi : \Phi \rightarrow \{T, F\}$ which assigns T (true) or F (false) to every primitive proposition.

Syntax of $\mathcal{L}_B$: The belief models introduced in the preceding section are parameterised by the internal language $\mathcal{L}$, used by agents to represent their beliefs. We desire a language $\mathcal{L}_B$, which can be used by us to represent agent's beliefs: it follows that $\mathcal{L}$ must appear in $\mathcal{L}_B$ somewhere. For simplicity, we assume that beliefs may be arbitrary formulae of $\mathcal{L}_B$, and thus that agents are capable of having beliefs about beliefs, and beliefs about beliefs about beliefs, and so on. Syntactically, $\mathcal{L}_0$ is easily extended to $\mathcal{L}_B$. The language is enriched by an indexed set of unary modal operators $[i]$, one for each $i \in Ag$, where $Ag = \{1, \ldots, n\}$ is a set of agents. A formula such as $[i]\varphi$ is to be read 'agent i believes φ'.

Definition 3 *The language $\mathcal{L}_B$ contains the following symbols:*

1. *All symbols of $\mathcal{L}_0$;*
2. *The set $Ag = \{1, \ldots, n\}$ of agents;*
3. *The symbols '$]$' and '$[$'.*

Definition 4 *The set $Form(\mathcal{L}_B)$ of (well-formed) formulae of $\mathcal{L}_B$ is defined by the following rules:*

1. *If $\varphi \in Form(\mathcal{L}_0)$ then $\varphi \in Form(\mathcal{L}_B)$;*
2. *If $\varphi \in Form(\mathcal{L}_B)$ and $i \in Ag$ then $[i]\varphi \in Form(\mathcal{L}_B)$;*
3. *If $\varphi \in Form(\mathcal{L}_B)$ then $\neg\varphi \in Form(\mathcal{L}_B)$ and $(\varphi) \in Form(\mathcal{L}_B)$;*
4. *If $\varphi, \psi \in Form(\mathcal{L}_B)$ then $\varphi \vee \psi \in Form(\mathcal{L}_B)$.*

Semantics of $\mathcal{L}_B$: A model for $\mathcal{L}_B$ is obtained by taking a model for $\mathcal{L}_0$ and adding a set of belief models, one for each agent.

Definition 5 *A model, M, for $\mathcal{L}_B$ is a pair $M = \langle \pi, \{b_i\} \rangle$, where $\pi : \Phi \rightarrow \{T, F\}$ is an interpretation for $\mathcal{L}_0$, and $\{b_i\}$ is an indexed set of belief models, one for each agent $i \in Ag$. If b_i is the belief model of agent i, then Δ_i is its base set, and BE_i is its belief extension relation.*

The semantics of $\mathcal{L}_B$ are defined via the satisfaction relation '$\models$', which holds between models and $\mathcal{L}_B$-formulae. The rules defining this relation are given below.

$$
\begin{array}{lcll}
M & \models & p & \text{iff} \quad \pi(p) = T \text{ (where } p \in \Phi) \\
M & \models & \neg\varphi & \text{iff} \quad M \not\models \varphi \\
M & \models & \varphi \vee \psi & \text{iff} \quad M \models \varphi \text{ or } M \models \psi \\
M & \models & [i]\varphi & \text{iff} \quad \varphi \in bel(b_i)
\end{array}
$$

Validity for $\mathcal{L}_B$ is defined in the usual way: if $\varphi \in Form(\mathcal{L}_B)$ and there is some model M such that $M \models \varphi$, then φ is said to be *satisfiable*, otherwise it is *unsatisfiable*. If $\neg\varphi$ is unsatisfiable, then φ is said to be *valid*, i.e., satisfied by every model. If φ is valid, we indicate this by writing $\models \varphi$.

Properties of $\mathcal{L}_B$: Since the propositional connectives of $\mathcal{L}_B$ have standard semantics, all propositional tautologies will be valid; additionally, the inference rule *modus ponens* will preserve validity. In short, we can use all propositional modes of reasoning in $\mathcal{L}_B$. However, $\mathcal{L}_B$ has some properties that $\mathcal{L}_0$ does not. To illustrate this, we first establish an analogue of Konolige's attachment lemma (Konolige, 1986a, pp34–35). (Note that if $\Delta = \{\varphi_1, \ldots, \varphi_n\}$ then $[i]\Delta$ abbreviates $[i]\varphi_1, \ldots, [i]\varphi_n$.)

Lemma 2 *The set $\{[i]\Delta, \neg[i]\Delta'\}$ is unsatisfiable iff $\exists\varphi \in \Delta'$ such that $(\Delta, \varphi) \in BE_i$.*

This lemma allows us to derive a number of useful results, for example:

Theorem 1 $\models \quad [i]\varphi_1 \wedge \cdots \wedge [i]\varphi_n \quad \Rightarrow \quad [i]\varphi$ *where* $(\{\varphi_1, \ldots, \varphi_n\}, \varphi) \in BE_i$.

This result represents the basic mechanism for reasoning about belief systems: if it is known that i believes $\{\varphi_1, \ldots, \varphi_n\}$, and that $(\{\varphi_1, \ldots, \varphi_n\}, \varphi) \in BE_i$, then this implies that i also believes φ.

Note that the K axiom and the necessitation rule of normal modal logic (Chellas, 1980) do *not* in general hold for belief modalities in $\mathcal{L}_B$, and thus $\mathcal{L}_B$ does not fall prey to logical omniscience. However, this does not mean that $\mathcal{L}_B$ is incapable of representing perfect reasoners: below, we show that $\mathcal{L}_B$ can be used to represent any agent that can be represented using normal modal logics of belief.

The Belief Extension Relation

We have now developed belief models, the basic mathematical tool for representing belief systems, and a language called $\mathcal{L}_B$, which can be used to express properties of belief systems. However, we have said little about the *meaning* of belief models, or where they come from. How can a belief model be associated with an agent? Under what circumstances can we say a belief model truly represents an agent's belief system? It is these questions that we address in this section.

In practice, it is possible to associate a b.e. relation with any system that satisfies the following two properties:

- it must be possible to characterise the system's 'belief state' as a set of formulae in some logical language $\mathcal{L}$;
- it must be possible to identify the 'legal belief states' of the system, in a way to be described below.

It is argued that the first requirement is quite weak. The beliefs of almost any conceivable agent can be described via a set of formulae of some language. In particular, the beliefs of AI systems are generally *directly represented* as a set of formulae. (It was this observation, of course, that gave the impetus to Konolige's deduction model (Konolige, 1986a, pp12–13).)

The purpose of the second requirement is simply to ensure that the set of all sets of $\mathcal{L}$-formulae can be partitioned into two disjoint sets: one representing *legal* belief states,

the other representing *illegal* belief states. The idea is that the system can *never* be in one of its illegal states, whereas for each of the legal states, there is some chain of events through which the system could come to be in that state. To illustrate this requirement, consider the following simple example. Suppose we have a *non-contradictory* agent: one that never simultaneously believes both a formula and its negation. Then the set of illegal belief states for this agent will include all those in which the agent believed both φ and $\neg\varphi$, for any $\varphi \in Form(\mathcal{L})$.

We shall now introduce some notation.

Definition 6 *If i is an agent, then the set of legal belief states of i is denoted BS_i. Note that $BS_i \subseteq \wp(Form(\mathcal{L}))$.*

Given the set BS_i, it is possible to derive a b.e. relation that correctly describes the behaviour of i's belief system. Suppose it has been observed that agent i has beliefs Δ. What *safe* predictions can we make about i's other beliefs? That is, what predictions could we make about i's beliefs that were *guaranteed* to be correct?

We could only safely say that if i believes Δ then it also believes φ iff whenever i believes Δ, it *necessarily* also believes φ. What interpretation can be given to the term 'necessarily'? One might say that if i believes Δ, then it necessarily also believes φ iff in *all legal states* where it believes Δ, it also believes φ. This notion of necessity is that at the heart of normal modal logics, where φ is said to be necessary iff φ is true in all possibilities (Chellas, 1980). However, it is important to note that belief is <u>not</u> being given a normal modal interpretation here. This leads to the following derivation of an agent's b.e. relation.

Definition 7 *If i is an agent, then BE_i, the derived b.e. relation of i, is defined thus:*

$$BE_i \stackrel{\text{def}}{=} \{(\Delta, \varphi) \mid \forall \Delta' \in BS_i,\ \text{if } \Delta \subseteq \Delta' \text{ then } \varphi \in \Delta'\}.$$

Before we can move on, we need to establish that derived b.e. relations *are* actually b.e. relations, i.e., that they satisfy the properties stated in Definition 1.

Theorem 2 *If BE_i is the derived b.e. relation of agent i, $\Delta \subseteq Form(\mathcal{L})$, and $\exists \Delta' \in BS_i$, s.t. $\Delta \subseteq \Delta'$, then (Δ, BE_i) is a belief model, i.e., BE_i satisfies the reflexivity, monotonicity, and transitivity conditions of Definition 1.*

If we have an agent's derived b.e. relation, then it is obvious that this relation correctly describes the behaviour of that agent's belief system.

Relationship to Other Formalisms

An obvious question to ask of any new formalism for representing belief is: how *expressive* is it? In this section, we show that the new model is sufficiently expressive that it can be viewed as a generalisation of two other well-known formalisms for modelling belief.

The Deduction Model of Belief

The model of belief systems presented in this paper is similar in some respects to Konolige's deduction model (Konolige, 1986a). In fact, as we shall demonstrate formally, the new model actually *generalises* the deduction model, in that the behaviour of any belief system in the deduction model can be represented using the new model. Before this result is established, a review of the deduction model is given.

The deduction model uses *deduction structures* to represent belief systems. A deduction structure d is a pair $d = (\Delta, \rho)$, where Δ is a base set of beliefs, (in much the same way as in the new model), and ρ is a set of *deduction rules*. A deduction rule is a rule of inference with the following properties:

- it has a fixed, finite number of premises; and
- it is an effectively computable function of those premises.

If $\Delta \subseteq Form(\mathcal{L})$, and ρ is a set of deduction rules for the language $\mathcal{L}$, then we write $\Delta \vdash_\rho \varphi$ iff there is a proof of φ from Δ using only the rules in ρ. The *deductive closure* of a set Δ under rules ρ is the set of formulae that may be derived from Δ using ρ. Formally, the deductive closure of a deduction structure is given by the function *close*:

$$close((\Delta, \rho)) \stackrel{\text{def}}{=} \{\varphi \mid \Delta \vdash_\rho \varphi\}.$$

A model for a propositional version of Konolige's language L^B (which corresponds syntactically to $\mathcal{L}_B$) is a pair $\langle \pi, \{d_i\} \rangle$, where π is a propositional valuation, and $\{d_i\}$ is an indexed set of deduction structures, one for each agent. The semantics of the modal belief operator $[i]$ is then:

$$\langle \pi, \{d_i\} \rangle \models [i]\varphi \quad \text{iff} \quad \varphi \in close(d_i).$$

We now state the key results of this section.

Theorem 3 *If $d = (\Delta, \rho)$ is a deduction structure, and BE_ρ is the derived b.e. relation associated with ρ, then $close((\Delta, \rho)) = bel((\Delta, BE_\rho))$.*

An obvious corollary of this theorem is the following.

Theorem 4 *Belief models are at least as expressive as deduction structures. That is, for any deduction structure $d = (\Delta, \rho)$, there exists a corresponding belief model $b = (\Delta, BE)$ such that $close(d) = bel(b)$.*

Normal Modal Logics of Belief

In this section, we compare the possible worlds approach with our belief models. To do this, we first define a language $\mathcal{L}_w$, with a syntax identical to $\mathcal{L}_B$, but with a possible worlds semantics. $\mathcal{L}_w$ is essentially a normal modal logic with the single necessity operator replaced by an indexed set of necessity operators $[i]$, one for each agent. Models for $\mathcal{L}_w$ are generalisations of the models for normal modal logics (Chellas, 1980).

Definition 8 *A model, M_w, for $\mathcal{L}_w$ is a triple $M_w = \langle W, \{R_i\}, \pi_w \rangle$, where W is a non-empty set of world, $\{R_i\}$ is an indexed set of relations over W, one for each agent $i \in Ag$, and $\pi_w : W \times \Phi \rightarrow \{T, F\}$ is a valuation function that gives the truth of each primitive proposition in each world.*

The truth of a formula thus depends upon which world it is interpreted in:

$$\begin{array}{lll}
\langle M_w, w \rangle \models p & \text{iff} & \pi(w, p) = T \quad (\text{where } p \in \Phi) \\
\langle M_w, w \rangle \models \neg\varphi & \text{iff} & \langle M_w, w \rangle \not\models \varphi \\
\langle M_w, w \rangle \models \varphi \vee \psi & \text{iff} & \langle M_w, w \rangle \models \varphi \text{ or } \langle M_w, w \rangle \models \psi \\
\langle M_w, w \rangle \models [i]\varphi & \text{iff} & \forall w' \in W, \text{ if } (w, w') \in R_i \\
 & & \quad \text{then } \langle M_w, w' \rangle \models \varphi.
\end{array}$$

Name	Theorem	Condition on R_i
K	$[i](\varphi \Rightarrow \psi) \Rightarrow (([i]\varphi) \Rightarrow ([i]\psi))$	any
T	$[i]\varphi \Rightarrow \varphi$	reflexive
D	$[i]\varphi \Rightarrow \neg[i]\neg\varphi$	serial
4	$[i]\varphi \Rightarrow [i][i]\varphi$	transitive
5	$\neg[i]\varphi \Rightarrow [i]\neg[i]\varphi$	euclidean

Table 1: Theorems K, T, D, 4, and 5

The most interesting properties of $\mathcal{L}_w$ are those which relate conditions on accessibility relations in the model structure to theorems in the corresponding logic. Although there are many properties which correspond to theorems, (see, e.g., (Chellas, 1980)), only five are of real interest from the point of view of belief logics: those called K, T, D, 4, and 5. These theorems, and the conditions they correspond to, are summarised in Table 1. Theorem T is often called the knowledge theorem: it says that if i believes φ, then i is true in the world. Theorem D is the consistency theorem: it says that if i believes φ, then it does not believe $\neg\varphi$. Theorems 4 and 5 are called the positive and negative introspection theorems, respectively: together, they characterise agents that are perfectly aware about their own beliefs.

As we add conditions to accessibility relations, we get progressively more theorems in the corresponding logic. For example, if we demand that R_i is reflexive and transitive, then we have a logic with theorems K, T, and 4. We refer to this logic as the *system* KT4. As it turns out, there are just eleven distinct systems of modal logic based on the theorems K, T, D, 4, and 5 (see (Chellas, 1980, p132)). These are: K, K4, K5, KD, KT, K45, KD5, KD4, KT4, KD45, and KT5. However, axiom T is generally taken to characterise *knowledge*, not belief. For this reason, we shall consider this axiom no further here; we restrict our attention to the four remaining axioms and their eight remaining systems: K, K4, K5, KD, K45, KD5, KD4, and KD45. We shall shortly define a correspondence theorem, (cf. (Konolige, 1986a, pp104–108)), which relates $\mathcal{L}_w$ and these eight systems to $\mathcal{L}_B$, in much the same way that the previous section related $\mathcal{L}_B$ to the deduction model. First, however, some notation.

Definition 9 *If $M_w = \langle W, \{R_i\}, \pi_w \rangle$ is an $\mathcal{L}_w$-model, then φ is said to be valid in M_w, (notation $M_w \models \varphi$) iff if $w \in W$, then $\langle M_w, w \rangle \models \varphi$. In saying that Σ is a system of normal modal belief logic we mean $\Sigma \in \{K, K4, K5, KD, K45, KD5, KD4, KD45\}$. In saying that an $\mathcal{L}_w$-model M_w is a Σ-model, we mean that if φ is a Σ-theorem, then $M_w \models \varphi$. If M_w is a Σ-model, we indicate this by writing M_w^Σ.*

Definition 10 *If i is an agent, $M_w = \langle W, \ldots \rangle$ is an $\mathcal{L}_w$-model, and $w \in W$, then the belief set of i at w in M_w is given by the function bel_w:*

$$bel_w(M_w, w, i) \stackrel{\text{def}}{=} \{\varphi \mid \langle M_w, w \rangle \models [i]\varphi\}$$

The set of legal belief states associated with a system Σ is then:
$BS_\Sigma \stackrel{\text{def}}{=} \{\Delta \mid \exists M_w^\Sigma = \langle W, \ldots \rangle, \exists w \in W, \exists i \in Ag \text{ s.t. } bel_w(M_w^\Sigma, w, i) = \Delta\}$.

The derived b.e. relation for a system Σ may then be obtained in the usual way (see above). We can now state our correspondence theorem.

Theorem 5 *Let Σ be a normal modal system of belief, (Δ, BE_Σ) be a derived belief model of Σ, and φ be one of the theorems K, D, 4 or 5. Then the property expressed by φ is true of (Δ, BE_Σ) iff φ is a Σ-theorem.*

This theorem has an obvious corollary.

Theorem 6 *For every normal modal belief system Σ, there exists a corresponding class of derived $\mathcal{L}_B$ models which satisfy just the theorems of Σ.*

We can actually prove a stronger result than this, which is more in the spirit of the deduction model result, above. However, the statement of this result is a good deal more involved, and so we omit it; see (Wooldridge, 1994).

Remarks The results of this section are important for the new model, as they show that it can be used to represent the kinds of belief system that existing formalisms are capable of representing. However, there is an informal sense in which the new model is *more* expressive than those we have compared it to: because b.e. relations are an *abstract* way of representing an agent's reasoning, they can readily be used to capture properties of belief systems that would be awkward to represent using other formalisms — if they could be represented at all. This point is discussed in (Wooldridge, 1994).

Implementation Aspects

The issues surrounding the implementation of a system which makes use of $\mathcal{L}_B$ in some way are not the primary concern of this paper (see, e.g., (Stein and Barnden, 1995) for a description of the CASEMENT system for reasoning with beliefs). Nevertheless, it is worth briefly commenting on these issues. First, note that a tableau-based decision procedure for $\mathcal{L}_B$ has been developed, and is described in the associated technical report (Wooldridge, 1994). (It has not been presented here due to space restrictions.) This proof method has been used on a number of examples, including the wisest man puzzle (Konolige, 1986a, pp57–61). A PROLOG implementation of this procedure has been developed and tested on a number of problems (Gibbs, 1994).

Secondly, recall the informal interpretation given to an agent's b.e. relation, as described earlier: if i believes Δ and $(\Delta, \varphi) \in BE_i$ then i also believes φ. This interpretation corresponds to an axiom in the logic $\mathcal{L}_B$ (see Theorem 1):

$$\models [i]\varphi_1 \wedge \cdots \wedge [i]\varphi_n \Rightarrow [i]\varphi \quad \text{where } (\{\varphi_1, \ldots, \varphi_n\}, \varphi) \in BE_i.$$

This axiom readily lends itself to forward reasoning; a b.e. relation can thus be implemented as a set of rules, very much

like rules in the standard AI sense. Backward reasoning may proceed in the obvious way; to see whether i believes φ, find some $\varphi_1, \ldots, \varphi_n$ such that i believes $\varphi_1, \ldots, \varphi_n$, and $(\{\varphi_1, \ldots, \varphi_n\}, \varphi) \in BE_i$.

Finally, note that reasoning in $\mathcal{L}_B$ may utilise the technique of *semantic attachment*, described by Konolige (and attributed by him to Weyhrauch) (Konolige, 1986a, p7). The idea is that when reasoning in $\mathcal{L}_B$ we must often decide whether $(\{\varphi_1, \ldots, \varphi_n\}, \varphi) \in BE_i$, for some agent i and $\{\varphi_1, \ldots, \varphi_n, \varphi\} \subseteq Form(\mathcal{L}_B)$. Under certain circumstances, this reduces to another decision problem. For example, if we have agents whose internal language $\mathcal{L}$ is the standard propositional logic $\mathcal{L}_0$, and that are perfect $\mathcal{L}_0$ reasoners, then deciding whether $(\{\varphi_1, \ldots, \varphi_n\}, \varphi) \in BE_i$ reduces to deciding whether $\{\varphi_1, \ldots, \varphi_n\} \vdash_{\mathcal{L}_0} \varphi$. We can thus directly simulate an agent's reasoning process in order to decide whether some pair is present in its b.e. extension relation.

Concluding Remarks

Formalisms for representing the belief systems of resource bounded reasoning agents are an area of ongoing research in (D)AI. This paper has contributed to the theoretical foundations of such formalisms, by presenting a new abstract general model of resource-bounded belief. A logic called $\mathcal{L}_B$ has been developed, containing belief modalities with semantics given in terms of the new model. The properties of the model and logic have been investigated in detail, and they have been shown to generalise two other well-known formalisms for representing belief. Future work will look at integrating this model with other components of an agent's cognitive makeup (for example, the interaction between beliefs and intentions); this work has already begun, in a practical sense, in an agent-oriented DAI testbed called MyWorld (Wooldridge, 1995). Another well-known attempt to integrate models of belief with other mental attitudes is (Cohen and Levesque, 1990), where belief and goal modalities are used to define the notion of intention. Finally, note that in other work, we have considered the implications of adding *temporal* modalities to $\mathcal{L}_B$ (Wooldridge, 1994; Wooldridge and Fisher, 1994).

References

Bond, A. H. and Gasser, L., editors (1988). *Readings in Distributed Artificial Intelligence*. Morgan Kaufmann Publishers: San Mateo, CA.

Chellas, B. (1980). *Modal Logic: An Introduction*. Cambridge University Press: Cambridge, England.

Cohen, P. R. and Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence*, 42:213–261.

Fagin, R. and Halpern, J. Y. (1985). Belief, awareness, and limited reasoning. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 480–490, Los Angeles, CA.

Gibbs, J. R. (1994). Tableau-based theorem proving in a temporal belief logic. Master's thesis, Department of Computing, Manchester Metropolitan University, Chester St., Manchester M1 5GD, UK.

Halpern, J. Y. and Moses, Y. (1992). A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54:319–379.

Hintikka, J. (1962). *Knowledge and Belief*. Cornell University Press: Ithaca, NY.

Konolige, K. (1986a). *A Deduction Model of Belief*. Pitman Publishing: London and Morgan Kaufmann: San Mateo, CA.

Konolige, K. (1986b). What awareness isn't: A sentential view of implicit and explicit belief (position paper). In Halpern, J. Y., editor, *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning About Knowledge*, pages 241–250. Morgan Kaufmann Publishers: San Mateo, CA.

Lakemeyer, G. (1991). A computationally attractive first-order logic of belief. In *JELIA-90: Proceedings of the European Workshop on Logics in AI (LNAI Volume 478)*, pages 333–347. Springer-Verlag: Heidelberg, Germany.

Levesque, H. J. (1984). A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence (AAAI-84)*, pages 198–202, Austin, TX.

Stein, G. C. and Barnden, J. A. (1995). Towards more flexible and common-sensical reasoning about beliefs. In Cox, M. and Freed, M., editors, *Representing Mental States and Mechanisms — Proceedings of the 1995 Spring Symposium*. AAAI Press.

Wooldridge, M. (1994). A temporal belief logic. Technical report, Department of Computing, Manchester Metropolitan University, Chester St., Manchester M1 5GD, UK.

Wooldridge, M. (1995). This is MyWorld: The logic of an agent-oriented testbed for DAI. In Wooldridge, M. and Jennings, N. R., editors, *Intelligent Agents: Theories, Architectures, and Languages (LNAI Volume 890)*, pages 160–178. Springer-Verlag: Heidelberg, Germany.

Wooldridge, M. and Fisher, M. (1994). A decision procedure for a temporal belief logic. In Gabbay, D. M. and Ohlbach, H. J., editors, *Temporal Logic — Proceedings of the First International Conference (LNAI Volume 827)*, pages 317–331. Springer-Verlag: Heidelberg, Germany.