# Common and Mutual Belief for Agent Modeling

Ullrich Hustadt*

Max-Planck-Institut für Informatik, Im Stadtwald, D-66123 Saarbrücken
Telefon 0681/302-5431, Telefax 0681/302-5401, E-mail hustadt@mpi-sb.mpg.de

## 1 Introduction

We want to consider a dialog situation between a system and a heterogeneous group of dialog partners. In the following, we use the term '*agent*' for all participants of a dialog. Our problem is to find adequate representational means for describing the beliefs, goals, and plans of each agent. We assume that we can provide a sufficiently detailed description of the knowledge base of the system, but we don't have complete descriptions of the knowledge bases of all other participating agents. However, we assume that there is a minimal amount of knowledge common to all knowledge bases. Knowledge required as a basis for producing a sensible dialog belongs to this common part of knowledge. For example, it can be assumed that all the agents have the knowledge that saying "Hello" is a greeting and the starting point of a dialog. Another form of common knowledge we want to model is that of *stereotypes*. A stereotype is a collection of sentences assigned to members of a specific group of individuals. As soon as we assume that an agent belongs to such a specific group of individuals, we can ascribe all the sentences attached to this group to the agent. I present an example in Section 2.

Representational schemes developed in the field of knowledge representation are often employed in natural-language dialog systems for representing knowledge about the world. Such schemes are, for instance, formulas of first-order predicate logic, semantic networks, and frames. But these are inadequate for representing the required detail of information we need.

Hendrix [3] proposes to extend the semantic network formalism by *partitions* to solve this problem. The basic idea is to maintain a number of separate partitions to store the system's beliefs about the world, the system's goals, the system's assumptions about the dialog partner's beliefs about the domain, the system's assumptions about the dialog partner's assumption about the system's beliefs about the domain, etc. Within each partition the standard semantic network formalism can be used. In the BGP-MS system [5] this representation scheme has been enhanced by *partition inheritance* which, for instance, allows the system's and dialog partner's mutual beliefs to be stored in a separate partition whose contents are inherited by the partitions containing the system's

* **Acknowledgments:** This work is supported by the German Ministry for Research and Technology (BMFT) under grant ITS 9102 (Project Logo). Responsibility for the contents lies with the author.

assumptions about the world and the system's assumptions about the user's beliefs about the world.

Knowledge representation frameworks based on the partition approach have been used in a number of applications, and their utility has been demonstrated. They are still widely used (cf. [5, 2]). However, they have restrictions.

- Partitioned semantic networks have no formal semantics.
- All reasoning is confined to one partition.
- The inheritance between partitions is implemented by an ad hoc mechanism which cannot be controlled by the knowledge engineer.

The approach I propose here is in line with the *modal logic approach* of Allgayer, Ohlbach, and Reddig [1]. The basic idea is to enhance a decidable fragment of first-order logic with modal operators for modeling the notions of belief, knowledge, and desires. To provide the initial knowledge base for agents, we support mutual and group beliefs, knowledge, and desires.

## 2 Examples Using Belief Modalities

The language we use to describe individual as well as stereotypical information is called Mod-$\mathcal{ALC}$. It is based on the terminological logic $\mathcal{ALC}$ [7] and extends the language of Hustadt and Nonnengart [4] with the modalities $\Box^{\mathtt{m}}_{(m,C)}$ and $\Box^{\mathtt{c}}_{(m,C)}$. These are used for describing information about groups of agents and generalize those operators available in our previous papers.

Before I define the syntax and semantics in Section 3, I present some motivating examples. Suppose our signature contains a modal operator symbol 'believe' and agent symbols 'Tom' and 'Tim'. The terminological sentence

$$\Box_{(\text{believe,Tim})}\,\text{Tom} \in \text{speeder} \tag{1}$$

describes that Tim believes that Tom is a person tending to drive too fast, i.e. in our possible worlds semantics, in any world in the belief space of Tim, Tom is a speeder. The terminological sentence

$$\Box_{(\text{believe,Tim})}\Box^{\mathtt{c}}_{(\text{believe,speeder})}\;\;(\text{2cv} \sqsubseteq \text{slow\_car}) \tag{2}$$

defines that Tim believes that anybody Tim regards as a speeder believes that a 2cv is a slow car. In this example, speeder is a concept representing a group of individuals. Such a concept is called *stereotype concept*. One has to be careful about the interpretation of the concepts speeder, 2cv, and slow_car. Whereas speeder is interpreted from the viewpoint of Tim, the concepts 2cv and slow_car are interpreted from the viewpoint of a speeder. Furthermore, the sentence specifies only what Tim believes that every speeder believes on his own. This form of belief is called *common belief* (indicated by the superscript $\mathtt{c}$ in $\Box^{\mathtt{c}}_{(\text{believe,speeder})}$). If we want to specify that any speeder believes in addition that any other speeder also believes that 2cv's are slow cars, then we use

$$\Box_{(\text{believe,Tim})}\Box^{\mathtt{m}}_{(\text{believe,speeder})}\;\;(\text{2cv} \sqsubseteq \text{slow\_car}) \tag{3}$$

which describes a *mutual belief* among speeders (indicated by the superscript $\mathtt{m}$ in $\Box^{\mathtt{m}}_{(\text{believe,speeder})}$).

An example, for the use of concept terms build using modal operators, is the following terminological sentence where speeder and nice_car are interpreted from the viewpoint of Tim and bad_car from the viewpoint of a speeder.

$$\Box_{(\mathsf{believe},\mathrm{Tim})} \ \ (\mathrm{nice\_car} \sqsubseteq \Box^{\mathsf{C}}_{(\mathsf{believe},\mathrm{speeder})} \ \mathrm{bad\_car})$$

## 3 Syntax and Semantics for Mod-$\mathcal{ALC}$

We assume four disjoint alphabets, the set $\mathsf{C}$ of *concept symbols*, the set $\mathsf{R}$ of *role symbols*, the set $\mathsf{M}$ of *modal operator symbols*, and the set $\mathsf{O}$ of *object symbols*. There is a distinguished subset $\mathsf{A}$ of the object symbols, called the set of *agent symbols*. The set $\mathsf{C}$ contains two distinguished elements *top* and *all* which denote the set of all objects and the set of all agents, respectively. The tuple $\Sigma = (\mathsf{O}, \mathsf{A}, \mathsf{M}, \mathsf{C}, \mathsf{R})$ is called the *signature*.

The set of *concepts* and *roles* is inductively defined as follows. Every concept symbol is a concept and every role symbol is a role. Now assume that $C$ and $D$ denote concepts, $R$ and $S$ denote roles, $m$ is a modal operator symbol, and $a$ is an agent symbol. Then $C \sqcap D$, $C \sqcup D$, $\neg C$, $\forall\, R.C$, $\exists\, R.C$, $\Box_{(m,a)}\, C$, $\Box^{\mathsf{C}}_{(m,C)}\, D$, $\Box^{\mathsf{m}}_{(m,C)}\, D$, and $\Diamond_{(m,a)}\, C$ are concepts.

The set of sentences of Mod-$\mathcal{ALC}$ is divided into the set of *terminological sentences* and the set of *assertional sentences*. If $C$ and $D$ are concepts, then $C \sqsubseteq D$ is a terminological sentence. If $C$ is a concept, $R$ is a role, and $x$, $y$, and $z$ are object symbols then $x \in C$ and $(y, z) \in R$ are assertional sentences. Moreover, if $\Phi$ is a terminological (respectively assertional) sentence and if $m$ is a modal operator symbol and $a$ is an agent symbol then $\Box_{(m,a)}\, \Phi$, $\Box^{\mathsf{C}}_{(m,C)}\, \Phi$, $\Box^{\mathsf{m}}_{(m,C)}\, \Phi$, and $\Diamond_{(m,a)}\, \Phi$, are terminological (respectively assertional) sentences. A *knowledge base* is a finite set of terminological and assertional sentences.

A note on notation: we use $A$ for concept symbols, $m$ for modal operator symbols, $a$ for agent symbols, $x$, $y$, and $z$ for object symbols, $C$, $D$, and $E$ for concepts, $R$ and $S$ for roles, and $\Phi$ for sentences.

This defines the syntax of Mod-$\mathcal{ALC}$. Now we provide the semantics. In essence, we are using the standard Kripke (possible worlds) semantics adjusted for our language.

**Definition 1 $\Sigma$-Structures.** As usual we define a *$\Sigma$-structure* as a pair $(\mathcal{D}, \mathcal{I})$ which consists of a domain $\mathcal{D}$ and an interpretation function $\mathcal{I}$ which maps the object symbols to elements of $\mathcal{D}$, concept symbols to subsets of $\mathcal{D}$ and the role symbols to subsets of $\mathcal{D} \times \mathcal{D}$. The interpretation of the concept symbol *top* is $\mathcal{D}$ and the interpretation of *all* is the set $\mathcal{A} = \{a \mid \mathcal{I}(x) = a \wedge x \in \mathsf{A}\}$.

**Definition 2 Frames and Interpretations.** By a frame $\mathcal{F}$ we understand any pair $(\mathcal{W}, \Re)$ where

- $\mathcal{W}$ is a non-empty set (of worlds).
- $\Re$ is the disjoint union $\biguplus_{m \in \mathsf{M}, a \in \mathsf{A}} \Re^a_m$ of binary relations $\Re^a_m$ on $\mathcal{W}$, the so-called *accessibility relations* between worlds.

By a $\Sigma$-interpretation $\Im$ based on $\mathcal{F}$ we understand any tuple $(\mathcal{D}, \mathcal{F}, \Im_{\mathrm{loc}}, \epsilon)$ where

- $\mathcal{D}$ denotes the common domain of all $\Sigma$-structures in the range of $\Im_{\mathrm{loc}}$.
- $\epsilon$ denotes the actual world (the current situation).
- $\mathcal{F}$ is a frame.
- $\Im_{\mathrm{loc}}$ maps worlds to $\Sigma$-structures with common domain $\mathcal{D}$ which interpret object symbols equally.

The accessibility relation for mutual belief is defined by:

**Definition 3.** Let $\Im = (\mathcal{D}, \mathcal{F}, \Im_{\mathrm{loc}}, \epsilon)$ be a $\Sigma$-interpretation, $\Im_{\mathrm{loc}}(\epsilon) = (\mathcal{D}, \mathcal{I})$, $m$ a modal operator name, and $C$ a concept. The set $\Re^{\mathtt{m}}_m(C)$ is the smallest set $S$ satisfying

$$S = \{\chi_2 \mid \exists\, a \in \mathcal{A}\colon \Re^a_m(\chi_1, \chi_2) \wedge \mathcal{I}(a) \in \Im(C) \wedge (\chi_1 \in S \vee \chi_1 = \epsilon)\}.$$

**Definition 4 Interpretation of Terms.** Let $\Im = (\mathcal{D}, \mathcal{F}, \Im_{\mathrm{loc}}, \epsilon)$ be a $\Sigma$-interpretation and let $\Im_{\mathrm{loc}}(\epsilon) = (\mathcal{D}, \mathcal{I})$. We define the interpretation of terms inductively over their structure:

$$\begin{aligned}
\Im(A) &= \mathcal{I}(A) \text{ if } A \text{ is a concept symbol}\\
\Im(P) &= \mathcal{I}(P) \text{ if } P \text{ is a role symbol}\\
\Im(C \sqcap D) &= \Im(C) \cap \Im(D)\\
\Im(C \sqcup D) &= \Im(C) \cup \Im(D)\\
\Im(\neg C) &= \mathcal{D} \setminus \Im(C)\\
\Im(\forall\, R.C) &= \{d \in \mathcal{D} \mid \forall\, e \in \mathcal{D}\colon (d, e) \in \Im(R) \Rightarrow e \in \Im(C)\}\\
\Im(\exists\, R.C) &= \{d \in \mathcal{D} \mid \exists\, e \in \mathcal{D}\colon (d, e) \in \Im(R) \wedge e \in \Im(C)\}\\
\Im(\Box_{(m,a)}\, C) &= \{d \in \mathcal{D} \mid \forall\, \chi \in \mathcal{W}\colon \Re^a_m(\epsilon, \chi) \Rightarrow d \in \Im[\chi](C)\}\\
\Im(\Box^{\mathtt{C}}_{(m,C)}\, D) &= \{d \in \mathcal{D} \mid \forall\, a \in \mathcal{A}\colon \forall\, \chi \in \mathcal{W}\colon\\
&\qquad\qquad \mathcal{I}(a) \in \Im(C) \wedge \Re^a_m(\epsilon, \chi) \Rightarrow d \in \Im[\chi](D)\}\\
\Im(\Box^{\mathtt{m}}_{(m,C)}\, D) &= \{d \in \mathcal{D} \mid \forall\, \chi \in \mathcal{W}\colon \chi \in \Re^{\mathtt{m}}_m(C) \Rightarrow d \in \Im[\chi](D)\}\\
\Im(\Diamond_{(m,a)}\, C) &= \{d \in \mathcal{D} \mid \exists\, \chi \in \mathcal{W}\colon \Re^a_m(\epsilon, \chi) \wedge d \in \Im[\chi](C)\}
\end{aligned}$$

where $\Im[\chi] = (\mathcal{D}, \mathcal{F}, \Im_{\mathrm{loc}}, \chi)$.

Note that $\Diamond_{(m,a)}$ is dual of $\Box_{(m,a)}$, i.e. $\Diamond_{(m,a)}\Phi$ is equivalent to $\neg\Box_{(m,a)}\neg\Phi$.

**Definition 5 Satisfiability.** Let $\Im = (\mathcal{D}, \mathcal{F}, \Im_{\mathrm{loc}}, \epsilon)$ be a $\Sigma$-interpretation and $\Im_{\mathrm{loc}}(\epsilon) = (\mathcal{D}, \mathcal{I})$. We define the satisfiability relation $\models$ inductively over the structure of Mod-$\mathcal{ALC}$ sentences:

$$\begin{aligned}
&\Im \models x \in C && \text{iff } \mathcal{I}(x) \in \Im(C)\\
&\Im \models (x, y) \in R && \text{iff } (\mathcal{I}(x), \mathcal{I}(y)) \in \Im(R)\\
&\Im \models C \sqsubseteq D && \text{iff } \Im(C) \subseteq \Im(D)\\
&\Im \models \Box_{(m,a)}\, \Phi && \text{iff } \forall\, \chi \in \mathcal{W}\colon \Re^a_m(\epsilon, \chi) \Rightarrow \Im[\chi] \models \Phi\\
&\Im \models \Box^{\mathtt{C}}_{(m,C)}\, \Phi && \text{iff } \forall\, a \in \mathcal{A}\colon \forall\, \chi \in \mathcal{W}\colon \mathcal{I}(a) \in \Im(C) \wedge \Re^a_m(\epsilon, \chi) \Rightarrow \Im[\chi] \models \Phi\\
&\Im \models \Box^{\mathtt{m}}_{(m,C)}\, \Phi && \text{iff } \forall\, \chi \in \mathcal{W}\colon \chi \in \Re^{\mathtt{m}}_m(C) \Rightarrow \Im[\chi] \models \Phi\\
&\Im \models \Diamond_{(m,a)}\, \Phi && \text{iff } \exists\, \chi \in \mathcal{W}\colon \Re^a_m(\epsilon, \chi) \wedge \Im[\chi] \models \Phi
\end{aligned}$$

Let $\Phi$ be a Mod-$\mathcal{ALC}$ sentence with $\Im \models \Phi$. Then we call $\Phi$ *satisfiable* in $\Im$ and we call $\Im$ a *model* for $\Phi$. An interpretation $\Im$ is a *model* of a knowledge base $K$ if it is a model for every sentence in $K$.

So far we did not define any special properties for the modal operators. Some typical properties are

$$\Box_{(m,a)}\,\Phi \Rightarrow \Diamond_{(m,a)}\,\Phi \tag{D}$$
$$\Box_{(m,a)}\,\Phi \Rightarrow \Phi \tag{T}$$
$$\Box_{(m,a)}\,\Phi \Rightarrow \Box_{(m,a)}\,\Box_{(m,a)}\,\Phi \tag{4}$$
$$\Diamond_{(m,a)}\,\Phi \Rightarrow \Box_{(m,a)}\,\Diamond_{(m,a)}\,\Phi \tag{5}$$

Similar schemata can be given for $\Box^{\mathtt{C}}_{(m,a)}$ and $\Box^{\mathtt{m}}_{(m,a)}$. The axiom schemata correspond to well-known properties of the accessiblity relations.

**Definition 6.** Let $\mathcal{R}$ be a set of properties of the accessibility relations. An interpretation $\Im$ is called a $\mathcal{R}$-*interpretation* if the accessibility relation $\Re$ of the underlying frame $\mathcal{F}$ satisfies all properties in $\mathcal{R}$. We say a set of Mod-$\mathcal{ALC}$ sentences $T$ entails $\Phi$ in all $\mathcal{R}$-interpretations if all $\mathcal{R}$-interpretations which are models of $T$ are also models of $\Phi$.

## 4 Implementation

Providing an expressively powerful language for the purpose of agent modeling is not enough. We also need a theorem proving method that is correct and complete with respect to the semantics of the language. For Mod-$\mathcal{ALC}$, this can be done using the ideas of Ohlbach [6]. The main idea is to manipulate modal logic formulas by some set of transformation rules so that classical, i.e. first-order, proof methods can be applied.

## References


1. J. Allgayer, H. J. Ohlbach, and C. Reddig. Modelling agents with logic. In *Proceedings of the Third International Workshop on User Modeling, DFKI Document D-92-17*, August 1992.
2. Afzal Ballim. *ViewFinder: A Framework for Representing, Ascribing and Maintaining Nested Beliefs of Interacting Agents.* PhD thesis, Université de Geneève, Geneva, Swiss, 1992.
3. G. Hendrix. Extending the utility of semantic networks through partitioning. In *IJCAI'75*, pages 115–121, 1975.
4. Ullrich Hustadt and Andreas Nonnengart. Modalities in knowledge representation. In *Proceedings of the 6th Australian Joint Conference on Artificial Intelligence*, pages 249–254, Melbourne, Australia, 16–19 November 1993. World Scientific.
5. Alfred Kobsa. Modeling the user's conceptual knowledge in BGP-MS, a user modelling shell system. *Computational Intelligence*, 6:193–208, 1990.
6. Hans Jürgen Ohlbach. Semantics based translation methods for modal logics. *Journal of Logic and Computation*, 1(5):691–746, 1991.
7. M. Schmidt-Schauß and G. Smolka. Attributive concept description with complements. *AI*, 48:1–26, 1991.