Using Hierarchical Autoepistemic Logic to Model Beliefs in Dialogue

J. A. Taylor

October 4, 1994

Abstract

The study of beliefs in dialogue has proceeded along two separate lines: the creation of logical formalisms which capture the inferences that it seems reasonable for agents to make, and the construction of implemented models that approximate what agents actually do. In our project to connect the two approaches we have created a belief model based on a reason maintenance system, and here we present a theory in nonmonotonic logic which describes the behaviour of this model.

We start out by examining the consequences of combining the use of modal operators (to represent nested beliefs) with autoepistemic operators (as a means of nonmonotonic inference). We then look at how some general principles of belief modelling can be represented, such as the persistence and ascription of beliefs, before applying our findings to the construction of a theory for the specific domain in which we are interested. We present a worked example of a dialogue in this domain, and outline the restrictions on the inferences sanctioned by the theory.

Although the purpose of the logic-based theory is merely to illustrate the principles behind the inferences made by the implemented model, we describe ways in which an implemented model can be created directly from the theory by the use of a theorem prover, either an off-the-shelf system or one specifically designed for this logic. Finally we look at the extentions that would need to be made to the theory in order to model further inferences available to human agents that can be made in our chosen domain.

1 Introduction: purpose of the HAEL description

One of the objectives of the BELMOD project is to bridge the gulf between logics of belief used in AI, and the properties of human beliefs as determined by psychologists. To this end we want to have a description of the system we are using in terms of formal logic.

The system which has been constructed as a model of agents with a limited set of belief states [12] uses the ATMS to keep track of which beliefs agents can consistently hold. It is a non-standard application of the ATMS in that the interpretations placed on ATMS labels are not the usual ones, and that inputs to the system can bring about updates to the state of the ATMS designed to cause a particular fact to be true of it by other means than simply asserting that fact.

There is no logic which directly describes the behaviour of the ATMS; Ginsberg's multi-valued logic [4] can be used, but the number of truth values required in a model of a fairly complex ATMS state is prohibitive. But we can capture the logical process which it is modelling using Hierarchical Autoepistemic Logic (HAEL [6]). This is because when reasoning about utterances we are only interested in the confidence agents have in their beliefs, which is indicated in the ATMS by the fact that they follow from more or less sound assumptions. HAEL allows propositions to hold in different "evidential spaces" which can be thought of as corresponding to levels of certainty, but does not associate each proposition with the actual assumptions that lead to it.

1.1 Scope of the model

The ATMS model at any single point in time is a model of a single agent at a corresponding point in time. In order to model a dialogue with contributions from multiple agents, an instance of the ATMS is created for each agent. As a simulated dialogue proceeds, the new information relating to the utterances is presented as input to each ATMS model, and causes updates in its state corresponding to the changes in the beliefs of the agent being represented. The identities of ATMS nodes thus correspond to propositions such as 'The first section is left of the palm beach', or 'Fred asked at time T3 whether I have the swamp'. We call an instantiation of the ATMS in this type of belief model a *simulated agent*. A model consisting of the mental states of a group of simulated agents plus some domain mechanics (e.g., audibility of utterances) would be an *electronically synthesized universe (ESU)*. Such a model might contain information about points in time other than the current one, such as that certain propositions are true "eventually" or "always" (Cohen & Levesque [2]) but such expressions are interpreted relative to a moving present.

Our logical model, however, has to express explicitly the relationship between an agent's beliefs at one time point and those at the next. This means adding a representation of a time point to all the agent's beliefs, including nested beliefs. Thus singly nested beliefs have two time points, e.g., "Doris believed at T1 that Fred believed at T2 that her references to "the swamp" meant the swamp in the top right hand corner of her map". A model of an individual agent with this type of beliefs we would call a *diary*. However we also choose to have a single theory of the world rather than separate theories for each agent, thus enabling direct representations of axioms such as "All agents are aware of any utterances that happen", so rather than a diary it is a *history*.

We are not attempting to replicate in the theoretical HAEL model the limit on the discrimination of individual deeply-nested beliefs that is the important feature of the ATMS model; the idea of the HAEL model is to represent the *ideal* of which the ATMS model is a restricted implementation. However, in the HAEL implementations to be described later we have made some concessions to practicality which restrict inferences on deeply-nested beliefs.

1.2 The role of HAEL's evidential spaces

As usual in HAEL, we will put propositions following from stronger evidence in the lower-numbered evidential speces, and those following from weaker evidence, such as defaults, in the higher-numbered spaces. Propositions are inherited from lower to higher spaces, and default rules in higher spaces refer to the absence of contradictory propositions in lower spaces.

Within each evidential space in HAEL, propositions can be written using the full expressiveness of weak S5, as a modal logic of belief. This allows us to represent agents who make assumptions which later turn out to be mistaken, and to model one agent reasoning about another agent's state of mind. What is notably absent from such a description is any representation of agents' beliefs about other agents' degrees of certainty. Nested belief sentences in weak S5 can capture the distinction between eg. being unsure whether another agent knows something and knowing that another agent is unsure of something, but the HAEL evidential space reflects only the strength of the actual evidence underlying the belief.

This is not entirely realistic; an agent may be dead certain that another agent holds a belief in a rather doubtful way, or doubtfully believe that another agent is dead certain about something, both cases being represented by the same sentence in the same evidential space. This, however, is exactly what happens in the ATMS implementation. The alternative would be to have agents model one another as HAEL engines, potentially very problematic.

1.3 Need for beliefs to be timed

The ATMS model has no explicit reference to time; it models an agent's beliefs at a particular moment, and that agent's beliefs with respect to other agents refer to what the latter agent believes at the same moment, and so on. However, there are reasons for including references to the times at which beliefs hold in the HAEL model. Firstly, although there is no explicit mention of time in the ATMS model, it does implicitly implement certain temporal axioms. For instance, any belief will persist if it is not explicitly altered by an addition made to the database; a nested belief that 'persists' in fact becomes a belief about a new belief (referring to a later point in time) of the agent to which it refers. The ATMS is able to implement such a model because of a feature of cooperative dialogue in general which is well illustrated by the Map Task. Worlds can change, but facts about how the world was at some time T_1 cannot, and agents' beliefs about such facts can only progress through a partially ordered series of states leading ultimately to full knowledge. These states are represented by assumptions justifying beliefs about the domain, which are either contradicted or reinforced into premises as the dialogue proceeds. It so happens in the Map Task that there are no state changes in the domain (the maps, rules, etc. stay the same). Thus the model does not need a temporal argument in predicates relating to the domain or in the belief operator.

Since the object of the HAEL exercise is to make the logical foundations of the system explicit, it is necessary to include time explicitly in the formalism and ensure that changes due to elapsing time are represented distinctly from changes due to nonmonotonicity. Thus, belief is represented as a three-place operator referring to an agent, a moment in time and an object sentence. Facts about the domain need no temporal place in their representations, though utterances do need one.

2 Fundamentals of our model

To construct a logical analogue of the operation of the ATMS-based model, we will combine four well-known logical formalisms, namely HAEL, the modal logic of belief weak S5 [5], and the situation calculus of temporal logic [7], and first-order predicate calculus. Each of these systems is superimposed on the preceeding one: the constructions of HAEL use weak S5 sentences as their objects, and the modal operators of the latter have been given an extra place to denote the situation in which the belief they refer to is applicable. Propositions (to which beliefs at times refer) are expressed in FOPC. The syntax and semantics presented here are not intended to be exhaustive, merely to express clearly what we intend when the interpretation of our formalisms would otherwise become unclear due to differences between the component systems upon which we have drawn.

We make one departure from standard situation calculus in allowing propositions without a time index to stand for facts that are true at all times; this is because in the Map Task domain, the facts of the world do not change, only agents' beliefs regarding them. In this respect, the combination of modal and temporal operators follows Appelt and Konolige [1] although we have a longer-handed notation for beliefs.

It may appear that we are duplicating some effort, as the properties of HAEL spaces have several elements in common with the belief spaces in weak S5. However, there are important differences, including the ubiquity of propagation of information from lower to higher HAEL spaces which is not a property of our model of belief.

2.1 Syntax

Our alphabet consists of the following: At the FOPC level,

- A set of individual constants C,
- A set of individual variables V,
- A set of functions F of fixed arity,
- A set of predicates P, also of fixed arity.

T is a term if $T \in C$ or $T \in V$. If $T_1 \ldots T_n$ are terms and $f \in F$ has arity n then $f(T_1 \ldots T_n)$ is a term. If $T_1 \ldots T_n$ are terms and $p \in P$ has arity n then $p(T_1 \ldots T_n)$ is a wff. If $\wp_1$ and $\wp_2$ are wffs then so are $\wp_1 \wedge \wp_2$, $\wp_1 \vee \wp_2$ and $\neg \wp_1$. If $v \in V$ and $\wp$ is a wff, then $\forall v \wp$ is a wff.

At the situation calculus level,

- A set of time constants TC,
- A set of time variables TV,
- A set of time functions TF of fixed arity,
- A set of time predicates TP, also of fixed arity.

T is a temporal term if $T \in TC$ or $T \in TV$. If $T_1 \ldots T_n$ are temporal terms and $f \in TF$ has arity n then $f(T_1 \ldots T_n)$ is a temporal term. If $T_1 \ldots T_n$ are temporal terms and $p \in TP$ has arity n then $p(T_1 \ldots T_n)$ is a wff. If $v \in TV$ and $\wp$ is a wff, then $\forall v \wp$ is a wff.

At the weak S5 level,

- A set of agent constants AC,
- A set of agent variables AV,
- A set of utterance types U,
- The belief operator bel,
- The action operator do

A is an agent term if $A \in AC$ or $A \in AV$. If A is an agent term, TT a temporal term, T a term and $\wp$ a wff then $bel(A, TT, \wp)$ and $do(A, TT, U, \wp)$ are wffs. If $v \in AV$ and $\wp$ is a wff, then $\forall v \wp$ is a wff.

At the HAEL level: wffs are divided into a series of evidential spaces $tau_0 \ldots \tau_n$. If n is an integer and $\wp$ is a sentence (no free variables), then $L_n \wp$ is a sentence in evidential spaces tau_m where $m > n$. We shall refer to the set of all wffs so defined as $\mathcal{L}$, and to the set of wffs allowable in evidential space n (i.e., excluding the L_m operator where $m > n$) $\mathcal{L}_n$. We will allow the usual syntactic sugar of, e.g., putting a single quantifier for a whole lot of variables, and writing equality as an infix.

2.2 Semantics

A world model $\mathcal{W}$ for this language consists of

- - a set $\mathcal{I}$ of individuals,
  - a fully ordered set $\mathcal{T}$ of discrete points in time,
  - a set $\mathcal{A}$ of agents
- a mapping $\mathcal{M}$ which maps:
  - C into $\mathcal{I}$,
  - F into $\bigcup_k \mathcal{I}^k \Rightarrow \mathcal{I}$
  - P into $\bigcup_k \mathcal{I}^k \Rightarrow \{true, false\}$
  - TC into $\mathcal{T}$
  - TF into $\bigcup_k \mathcal{T}^k \Rightarrow \mathcal{T}$
  - TP into $\bigcup_k \mathcal{T}^k \Rightarrow \{true, false\}$
  - AC into $\mathcal{A}$
  - U into $\mathcal{A} \times \mathcal{T} \times S \Rightarrow \{true, false\}$, where S is the set of wffs in the language
- A *variable assignment* is a mapping $\mathcal{V}$ which maps
  - V into $\mathcal{I}$,
  - TV into $\mathcal{T}$
  - AV into $\mathcal{A}$

The denotations of terms, and the truth values of sentences, are independent of time since object-level sentences are always invariant in time, whereas belief sentences always include the time at which they apply.

The denotation of a term t in world $\mathcal{W}, \mathcal{D}(t, \mathcal{W})$ is defined as follows:

- if $t \in C$ then $\mathcal{D}(t, \mathcal{W}) = \mathcal{M}(T)$
- if $t \in V$ then $\mathcal{D}(t, \mathcal{W}) = \mathcal{V}(T)$
- if $t = f(t_1 \ldots t_n)$ and $f \in F$ then $\mathcal{M}(t, \mathcal{W}) = \mathcal{M}(f)(\mathcal{D}(t_1, \mathcal{W}), \ldots \mathcal{D}(t_n, \mathcal{W}))$

The denotation of temporal and agent terms is defined similarly.

The truth value for any wff with respect to a world $\mathcal{W}$ is determined by the semantic rules given below. Note that while the semantics of beliefs is defined over a Kripke structure of possible worlds, that of utterances is defined model theoretically within a world. This reflects the intuition that what an agent actually believes is some relation betwen a proposition and the way the world is, whereas what he utters is merely a token consisting of a sentence in a particular mood.

- $\mathcal{W} \models p(t_1 \ldots t_n)$ iff $\mathcal{M}(p)(\mathcal{D}(t_1, \mathcal{W}), \ldots \mathcal{D}(t_n, \mathcal{W})) = true$ (and, or, implication are done in the usual way)
- $\mathcal{W} \models do(a, t, u, \wp)$ iff $\mathcal{M}(u)(\mathcal{D}(a), \mathcal{D}(t), \wp) = true$
- $\mathcal{W} \models \forall \theta \wp$ if for every $\mathcal{W}'$ matching $\mathcal{W}$ in every aspect except possibly the assignment of the variable $v, \mathcal{W}' \models \wp$

The truth values of belief sentences are defined in relation to a Kripke structure, or set of possible worlds with accessibility relations corresponding to what each agent considers to be possible at each point in time. For weak S5 modal logic, the accessibility relationship does not have any particular properties, and in this respect the possible-worlds semantics is less elegant than a situation semantics. However, we shall also use Kripke structures to present a semantics for HAEL structures.

A Kripke structure $\mathcal{K}$ consists of a set of possible worlds $\mathcal{P}$, the actual world $\mathcal{W} \in \mathcal{P}$ in which $\mathcal{I}, \mathcal{T}$ and $\mathcal{A}$ have their earlier-defined meanings, and the accessibility relationship $\mathcal{R}$ which maps $\mathcal{A} \times \mathcal{T}$ into $\mathcal{P}^2 \Rightarrow \{true, false\}$. Note that our sets of agents and time instants are the same in every possible world; we cannot express a state of affairs in which agents doubt their own, or others', existence, or think the time might be other than what it is.

- $< \mathcal{P}, \mathcal{W}, \mathcal{R} > \models bel(a, t, \wp)$ iff $\{\mathcal{W}' \in \mathcal{P} \mid \mathcal{R}(\mathcal{D}(a, \mathcal{W}), \mathcal{D}(t, \mathcal{W}))(\mathcal{W}, \mathcal{W}') = true\} \subseteq \{\mathcal{W}'' \in \mathcal{P} \mid < \mathcal{P}, \mathcal{W}'', \mathcal{R} > \models \wp\}$

In other words, a Kripke structure entails belief in a proposition if any world that is consistent with the beliefs of the agent at the time would, if it were the real world, entail that proposition.

At the HAEL level, we must cope with the fact that a theory expressed in our language is actually a set of theories corresponding to the set of evidential spaces. Konolige interprets the L operator, both in autoepistemic logic and in HAEL, as denoting self-belief. However, doubt has been cast on the validity of this interpretation [11, Ch. 3]; since L_n operators can only appear in higher evidential spaces than those to which they refer, this would mean that agents were never as certain about their own beliefs as they were about the subjects of those beliefs. Instead we take belief sentences occuring in different evidential spaces as representing different degrees of certainty, and treat L_n as denoting a sentence to be treated as background to the reasoning in progress; should a conclusion contradict such a sentence, one of its premises must be questioned.

Since we have already introduced an accessibility relationship between possible worlds for the belief operators, we can extend it to provide a semantics for the L operator. We therefore re-specify the Kripke structure such that it contains not one but n belief mappings or accessibility relations, $\mathcal{R}_n$, each corresponding to the worlds agents consider admissible at least some of the time in each evidential space. The semantics for belief sentences remain the same, except that now a model entails a belief in evidential space n iff it holds in all worlds accessible to the agent at the time under $\mathcal{R}_n$. Now unlike Konolige, we can only provide a semantics for the L_n operators when their argument is a sentence starting with a belief operator, but since the only things that can be uncertain in our models are beliefs, this is enough. The logic can be used to specify what happens if the certain beliefs of one agent are placed in direct contradiction with the less certain beliefs of another, although none of our theories will ever include direct relationships between different agents' beliefs.

The propagation of belief sentences up the hierarchy of evidential spaces corresponds to the restriction that the accessibility relationship in any space must include the relationship in the preceeding space. Sentences other than belief sentences will have the same truth value for a given model in any evidential space, as they do not depend on the accesibility relationship for their interpretation. Using $\mathcal{K} \models_n \wp$ to say that Kripke structure $\mathcal{K}$ is a model for $\wp$ in evidential space n, we have:

- $< \mathcal{P}, \mathcal{W}, \mathcal{R} > \models_n L_{n-1} bel(a, t, \wp)$ iff $\{\mathcal{W}' \in \mathcal{P} \mid \mathcal{R}_{n-1}(\mathcal{D}(a, \mathcal{W}), \mathcal{D}(t, \mathcal{W}))(\mathcal{W}, \mathcal{W}') = true\} \subseteq \{\mathcal{W}'' \in \mathcal{P} \mid < \mathcal{P}, \mathcal{W}'', \mathcal{R} > \models_{n-1} bel(a, t, \wp)\}$

This is similar to the definition that we have already given for beliefs, at least in the case of self-belief; not surprising as the intention of autoepistemic logic was to capture the notion of self-belief. The difference is that the conclusion goes in a different evidential space, whereas $bel(A, T, bel(A, T, \wp))$ (if we were at all interested in it) would go in the same space as $bel(A, T, \wp)$. Now in general, $\mathcal{K} \models_n \neg\wp$ iff $\mathcal{K} \not\models_n \wp$, so for any sentence $\wp$ and any space n, a model must entail either $L_n\wp$ or $\neg L_n\wp$ in τ_{n+1}.

Unfortunately we have not yet ruled out the case where the accesibility relationship avoids all worlds in which a default is applicable. We need to constrain the models for a theory to include only those in which the accessibility relationship in each space includes everything not disproved in a lower space. This can be done by adding the restriction that if there exists a model for theory T that entails $\neg\wp$ in τ_n then we must rule out any model for T which is identical with respect to the accessibility relationship in evidential spaces $\tau_0 \ldots \tau_n$, but which entails $L_n\wp$ in τ_{n+1}. This provision gives the semantics the nonmonotonic property that adding a sentence to a theory gives it extra models, allowing it not to entail consequences which previously it did. Note that it is not a model preference criterion; models that do not comply are ruled out entirely. This would be a problem for theories containing default conclusions about the world if they turned out to be wrong, but since ours contains only default conclusions about agents' beliefs the restriction on models described above only rules out models in which agents believe things they have *no cause* to believe – models which are inconsistent with our concept of agency.

2.3 Proof theory

Because defaults in HAEL are encoded within the object language, it would normally be the case that the only rule of inference we need is *modus ponens (MP)*. However, the L_i operator refers to the provability of a sentence in evidential space τ_i, so in order to infer $\neg L_i\wp$ we have to attempt to prove $\wp$ in τ_i, and add either $L_i\wp$ or $\neg L_i\wp$ depending on the result. This has the unfortunate consequence that even if the object language were semidecidable (which it is for some restricted versions with which we have experimented) the language as a whole is undecidable.

Furthermore, the necessitation principle (to be introduced in section 2.4 allows conclusions to be drawn according to what are theorems of the logic. This means that from the point of view of the proof procedure, axioms need to be marked as such, and inferences drawn based on proofs starting only with the axioms.

2.4 The logic of belief

Weak S5 modal logic is also known as KD45, because it incorporates the principle of necessitation and the K, 4 and 5 axiom schemata from full S5, but has the T axiom schema replaced by the weaker D schema. Where this is interpreted as a logic of belief, the necessitation principle states that all theorems of the logic are known to all agents, and as such knowledge of theorems therefore itself constitutes a further set of theorems, it too is known to all agents. Theorems are therefore mutually known. However, since there is nothing in the notation to distinguish theorems from sentences true in a particular model, the necessitation principle cannot itself be written as an axiom. In derivations that follow, Θ and Φ stand for wffs in schemata for axioms that refer to all wffs, and we shall refer to beliefs generated through the necessitation principle on theorems Th by $\mathrm{N}(Th)$.

The K axiom schema states that agents know the tautological consequences of their beliefs:

$$\text{K axiom} \qquad \vdash \forall A, T\; bel(A, T, \Theta) \wedge bel(A, T, (\Theta \rightarrow \Phi)) \rightarrow bel(A, T, \Phi) \qquad (1)$$

The T axiom schema would state that all beliefs of all agents are true, and would thus specify a logic of knowledge rather than one of belief. Instead the D axiom schema states that agents cannot hold inconsistent beliefs.

$$\text{*T axiom} \quad \vdash \forall A, T\; bel(A, T, \Theta) \rightarrow \Theta \tag{2}$$
$$\text{D axiom} \quad \vdash \forall A, T\; \neg(bel(A, T, \Theta) \wedge bel(A, T, \neg\Theta)) \tag{3}$$

The schemata generally known as 4 and 5 correspond in terms of belief to positive and negative introspection, and are written as follows:

$$\text{4 axiom} \quad \vdash \forall A, T\; bel(A, T, \Theta) \rightarrow bel(A, T, bel(A, T, \Theta)) \tag{4}$$
$$\text{5 axiom} \quad \vdash \forall A, T\; \neg bel(A, T, \Theta) \rightarrow bel(A, T, \neg bel(A, T, \Theta)) \tag{5}$$

We can easily derive

$$\text{Contrapositive of 5} \quad \neg bel(A, T, \neg bel(A, T, \Theta)) \rightarrow bel(A, T, \Theta) \tag{6}$$
$$\text{Substituting into 3} \quad bel(A, T, bel(A, T, \Theta)) \rightarrow \neg bel(A, T, \neg bel(A, T, \Theta)) \tag{7}$$
$$\text{Combining 6 and 7} \quad bel(A, T, bel(A, T, \Theta)) \rightarrow bel(A, T, \Theta) \tag{8}$$

From these it follows that at any moment agents know exactly what they do and do not believe at that moment, nested self-beliefs being equivalent to the beliefs themselves. This corresponds to the principle in the ATMS implementation by which agents' beliefs about their own beliefs are not represented save implicitly in the beliefs themselves. Because this model is a history, we can derive agents' beliefs that their own beliefs were incomplete at earlier times, though these play no role in the system. We can also derive two useful theorems, M and C:

$$\text{M theorem} \quad \vdash \forall A, T\; bel(A, T, \Theta \wedge \Phi) \rightarrow bel(A, T, \Theta) \wedge bel(A, T, \Phi) \tag{9}$$
$$\text{C theorem} \quad \vdash \forall A, T\; bel(A, T, \Theta) \wedge bel(A, T, \Phi) \rightarrow bel(A, T, \Theta \wedge \Phi) \tag{10}$$

To these axiom schemata are added those relating to the domain in which we are interested. Each disjunctive domain axiom schema at the object level will give rise to a separate axiom schema at the first nesting level for each of its disjuncts, as follows: Suppose we have an object level axiom schema

$$\vdash \Theta_1 \vee \ldots \vee \Theta_n \tag{11}$$

By the necessitation principle this is part of agents' beliefs:

$$\text{N}(11) \vdash \forall A, T\; bel(A, T, \Theta_1 \vee \ldots \vee \Theta_n) \tag{12}$$

and so is the result of every possible way of rewriting it as an implication, i.e., for each k,

$$\vdash \forall A, T\; bel(A, T, (\neg\Theta_1 \wedge \ldots \wedge \neg\Theta_n \text{ (excluding } \neg\Theta_k)) \rightarrow \Theta_k) \tag{13}$$

These resolve with the K axioms to give:

$$13,\ 1 \quad \vdash \forall A, T\; bel(A, T, (\neg\Theta_1 \wedge \ldots \wedge \neg\Theta_n) \text{ (excluding } \neg\Theta_k)) \rightarrow bel(A, T, \Theta_k) \tag{14}$$

Applying the necessitation principle to each axiom schema of the above form results in n new, distinct axiom schemata as follows:

$$\text{N}(14) \quad \vdash \forall A_1, A_2, T\; bel(A_2, T, (bel(A_1, T, (\neg\Theta_1 \wedge \ldots \wedge \neg\Theta_n) \text{ (excluding } \neg\Theta_k)) \rightarrow bel(A_1, T, \Theta_k)) \tag{15}$$

These are distinct in that they do not imply one another, so for any theorem prover they have to be generated or loaded separately. Applying the same set of operations again for the addition of another level of nesting, another n axiom schemata are generated for each of these; it follows that when implementing a theorem prover in a domain in which there is no limit on belief nesting, we have to generate such axiom schemata when they are required rather than loading them in advance.

It is worth noting that we do not need equality for modelling beliefs in domains where there is no ambiguity. Thus we do not need to supply the axioms characterizing the identity predicate for this system.

3 Details of possible HAEL formalizations

There are a number of ways in which defeasible conclusions drawn by agents in the Map Task can be mapped onto a set of evidential spaces. Firstly, there is the question of what counts as an uncertainty at all. Obviously enough the uncertainty about whether or not the partner's map corresponds to one's own in various details can be represented by allowing the default conclusion that it does to be derived in a particular space. But there are other regularities which could be represented as defaults. Two examples are persistence of beliefs and the non-occurrence of utterances.

3.1 Only use single-place default axioms

In Konolige's original exposition of HAEL, the form of default rules is a generalized version of their form in AE logic or in Reiter's default logic. Applications of these logics are generally restricted to domains in which only normal default rules apply, i.e., where conclusions are drawn conditionally on their being consistent with what is already believed. Such defaults appear in HAEL as follows:

$$\text{in } \tau_{n+1} : L_n\Theta(X) \wedge \neg L_n\neg\Phi(X) \rightarrow \Phi(X) \tag{16}$$

where $\Theta(X)$ and $\Phi(X)$ are arbitrary sentences in which individual X occurs. It is always possible to define a simple default $D(X)$ which is true in evidential space τ_{n+1} in the absence of evidence to the contrary, and which will allow the original conclusion to be drawn by adding an axiom in the strongest evidential space as follows:

$$\text{in } \tau_0 : \Theta(X) \wedge D(X) \rightarrow \Phi(X) \tag{17}$$

If there is evidence supporting $\Theta(X)$ and contradicting $\Phi(X)$ in a stronger evidential space than τ_{n+1} then the application of *Modus Tollens* to 17 (or its inherited version in τ_n) will generate $\neg D(X)$ in τ_n, preventing the default from firing. If the default is expressed as in 16, the rule is prevented from firing by the direct presence of $L_n\neg\Phi(X)$ in τ_{n+1}, resulting in the same overall conclusions. However, this type of default rule can give unsatisfactory conclusions because we cannot get useful propositions by applying *Modus Tollens* to it; we would have to add its reverse form explicitly.

To illustrate this, look what happens when a premise gives rise to a weak default conclusion, which itself gives rise to a strong default inference which is contradicted. What we want to do is throw out the weak conclusion, as the failure of the strong one calls it into question. Suppose we have the weak default that anything with two legs is a bird (most two-legged species are birds), and a stronger default rule that birds have beaks. If we know that Jules has two legs, we want to conclude that Jules is a bird, but if we know in addition that Jules has no beak we want to revise this opinion, allowing e.g., tht Jules is a human. Using defaults such as 17, this goes through nicely, as follows:

Axioms in τ_0 :	$\vdash \forall A\, legs(A) = 2 \wedge D_1(A) \rightarrow Bird(A)$	(18)
	$\vdash \forall A\, Bird(A) \wedge D_2(A) \rightarrow Beaked(A)$	(19)
Premises:	$legs(Jules) = 2$	(20)
	$\neg Beaked(Jules)$	(21)
Sub $Jules$ into 19	$Bird(Jules) \wedge D_2(Jules) \rightarrow Beaked(Jules)$	(22)
21, 22 by MT :	$\neg(Bird(Jules) \wedge D_2(Jules))$	(23)
23, by de Morgan	$\neg Bird(Jules) \vee \neg D_2(Jules)$	(24)
Sub $Jules$ into 18	$legs(Jules) = 2 \wedge D_1(Jules) \rightarrow Bird(Jules)$	(25)
24, 25 by MT :	$\neg(legs(Jules) = 2 \wedge D_1(Jules)) \vee \neg D_2(Jules)$	(26)
26, by de Morgan	$\neg(legs(Jules) = 2) \vee \neg D_1(Jules) \vee \neg D_2(Jules)$	(27)
Resolve 27 with 20	$\neg D_1(Jules) \vee \neg D_2(Jules)$	(28)

Defaults in τ_1 :	$\forall A\, \neg L_0 \neg D_2(A) \rightarrow D_2(A)$	(29)
Self-awareness of τ_0 :	$\neg L_0 \neg D_2(Jules)$	(30)
Sub $Jules$ into 29	$\neg L_0 \neg D_2(Jules) \rightarrow D_2(Jules)$	(31)
30, 31, by MP	$D_2(Jules)$	(32)
Resolve 28 with 32	$\neg D_1(Jules)$	(33)
Resolve 24 with 32	$\neg Bird(Jules)$	(34)

Having negated the default $D_1(Jules)$ in τ_1, we cannot conclude in any space that Jules is a bird. If the defaults are written as in 16, there is no way of inferring Jules' non-birdhood, and the conclusion is drawn in τ_2 that Jules is a bird. This would have to be fixed by adding a further default rule to the effect that any unbeaked entity is not a bird unless we happen to know that it is.

Representing default rules as axioms plus single-place default inferences also simplifies the modelling of nested reasoning. As we saw in section 2.4, we need to distinguish theorems from other sentences in the database in order that we can ascribe them to agents. Sticking to the schema of 17 enables the axiomatic component of an inference rule to be clearly separated from the default component, allowing us to choose suitable rules for the sharing of the latter amongst multiple agents, for instance a default of belief ascription.

3.2 Persistence as a default in HAEL

The idea that persistence of beliefs should be represented as a default is a common one in AI. In particular it is seen as a way round the frame problem, the necessity to derive from scratch all the facts that remain unchanged betwen successive states of the world. However, representing persistence as a default gives rise to a conceptual problem due to confusion of stronger evidence with newer evidence. For example, let us create a HAEL simulated agent Nigel, who leaves Tulip, his car, parked outside his house when he goes out. He has very strong evidence from which to conclude that he will find it there when he returns. One should model this by placing the persistence axiom in a strong evidential space. Now before he returns he sees a car identical to Tulip pass at great speed, pursued by the police. He doesn't see the number plate but recognizes the possibility that this is Tulip, stolen, and that therefore it won't be outside his house when he gets back. The weak evidence for the theft theory requires that it depend on a default in a weaker evidential space than that containing the persistence default, yet it would be wrong to throw out the theft theory on the grounds that it makes predictions that contradict those of the theory in which Tulip's location persists. Problems conceptually similar to this do occur when modelling the Map Task, e.g., an utterance might cast possible doubt on a conclusion that was drawn strongly from a previous utterance.

HAEL can represent such problems satisfactorily by having a series of evidential spaces representing persistence states immediately below each evidential space into which non-temporal default conclusions can be placed. In addition, the persistence axioms must be dependent upon the absence of an 'abnormal' condition, which is asserted when evidence of a change is manifest. In the following examples, we shall use $succ(T)$ to denote the time following the completion of whatever action was started at T. Note that in this world, unlike in the Map Task, facts about the external world (namely, Tulip's location) are subject to change over time; to represent such facts we use two-place predicates in which one place is a temporal term.

Nigel's beliefs in the situation in which there is no evidence to suggest that Tulip has been moved are represented as follows (for brevity's sake we omit the belief operator and its other arguments):

in τ_0 : $home(Tulip, T_0)$ (35)

in τ_1 : Defaults at this level

$L_0 home(Tulip, T) \wedge \neg L_0 ab(home(Tulip, succ(T))) \rightarrow$
$home(Tulip, succ(T))$ (36)

Self-awareness of τ_0 :

$L_0 home(Tulip, T_0)$ (37)

$\neg L_0 ab(home(Tulip, T_0[succ(T_0), succ(succ(T_0)) \ldots]))$ (38)

Conclusions:

$home(Tulip, T_0[succ(T_0), succ(succ(T_0)) \ldots])$ (39)

The sight of the obviously stolen car asserts the abnormality of the assumption that Tulip stayed where it was. This allows the inference that Tulip is not where it was to be made as follows:

in τ_0 : $home(Tulip, T_0)$ (40)

$stolen(car123, succ(T_0))$ (41)

$ab(home(Tulip, succ(T_0)))$ (42)

$stolen(Tulip, succ(T_0)) \rightarrow \neg home(Tulip, succ(T_0))$ (43)

in τ_1 : Defaults at this level

$L_0 home(Tulip, T) \wedge \neg L_0 ab(home(Tulip, succ(T))) \rightarrow$
$home(Tulip, succ(T))$ (44)

Self-awareness of τ_0 :

$L_0 home(Tulip, T_0)$ (45)

Conclusions: none

in τ_2 : Defaults at this level

$\neg L_1 car123 \neq Tulip \rightarrow car123 = Tulip$ (46)

Self-awareness of τ_1 :

$\neg L_1 car123 \neq Tulip$ (47)

Inheritance from τ_1 :

$$stolen(car123, succ(T_0)) \tag{48}$$

$$stolen(Tulip, T_1) \rightarrow \neg home(Tulip, succ(T_0)) \tag{49}$$

Conclusions:

$$car123 = Tulip \tag{50}$$

$$stolen(Tulip, succ(T_0)) \tag{51}$$

$$\neg home(Tulip, succ(T_0)) \tag{52}$$

On the other hand, if Nigel sees a car identical to Tulip parked at a meter, but again fails to notice the number plate, there is no implied assertion of abnormality. In this case, the weak evidence that it is Tulip at the meter is overridden by the strong persistence of the belief that Tulip is where he left it, which allows the deduction of the fact that the car he saw was not Tulip, as follows:

in τ_0 :

$$home(Tulip, T_0) \tag{53}$$

$$elsewhere(car123, succ(T_0)) \tag{54}$$

$$elsewhere(tulip, succ(T_0)) \rightarrow \neg home(Tulip, succ(T_0)) \tag{55}$$

in τ_1 : Defaults at this level

$$L_0 home(Tulip, T) \wedge \neg L_0 ab(home(Tulip, succ(T))) \rightarrow home(Tulip, succ(T)) \tag{56}$$

Self-awareness of τ_0 :

$$L_0 home(Tulip, T_0[succ(T_0), succ(succ(T_0)) \ldots]) \tag{57}$$

$$\neg L_0 ab(home(Tulip, T_0[succ(T_0), succ(succ(T_0))) \ldots]) \tag{58}$$

Inheritance from τ_0 :

$$elsewhere(car123, succ(T_0)) \tag{59}$$

$$elsewhere(Tulip, succ(T_0)) \rightarrow \neg home(Tulip, succ(T_0)) \tag{60}$$

Conclusions:

$$home(Tulip, succ(T_0)[succ(succ(T_0)) \ldots]) \tag{61}$$

$$home(Tulip, succ(T_0)) \rightarrow \neg elsewhere(Tulip, succ(T_0)) \tag{62}$$

$$elsewhere(car123, succ(T_0)) \tag{63}$$

$$\neg elsewhere(Tulip, succ(T_0)) \tag{64}$$

$$car123 \neq Tulip \tag{65}$$

The problem here is that if persistence is abnormal, this fact must be asserted in the lowest evidential space, even though Nigel only believes in a higher space that the car he saw was Tulip. What is required is a rule that allows any conclusion in a higher space that the value of a particular proposition has changed, to allow the assertion in the lowest space that its persistence is abnormal.

In these examples, only one extra evidential space is shown to illustrate the dimension of time, but there could be more, as would be required for example to reason about the time at which Tulip was stolen. However, one extra evidential space indicating persistence for each space in which new conclusions are drawn would be adequate for the Map Task.

3.3 Persistence in the Map Task

In the last section we saw that we can use an axiom for persistence of beliefs as part of a simulated agent drawing conclusions in the light of new information.

However, as we are attempting to specify a history rather than a simulated agent, we also need to represent the way the agent's own beliefs change over time. As a first experiment along these lines, let us incorporate the belief axioms of weak S5 as described in section 2.4. We will now express beliefs with a three-place operator bel, taking an agent, a time and a sentence representing a proposition, e.g., $bel(Nigel, T_0, home(Tulip, T_1))$. Nigel's belief that his car is at home is asserted when he leaves it there; simultaneously he applies the persistence axiom[1] to generate the belief that it will still be there later. If we take a time between when Nigel leaves his car and when he sees the stolen car, we find both his belief that it was there to start with and his belief that it is still there, persisting in parallel. After he has seen the stolen car, he believes that it is abnormal for the car's presence at home to persist to the current time. But the persistence of the belief he derived earlier that the car would now still be at home must also be defeated.

There is no easy way to specify this. In the Tulip example, Nigel's beliefs did not persist at all; at the points in time which were discussed, all beliefs were derived from the axioms relating to the events under consideration. In order to derive a set of valid beliefs about a moment in time, one can either do this or use a truth maintenance system. In the later case beliefs do persist from earlier states, and dependency information is explicitly maintained to allow the beliefs that have changed to be identified and updated. As the whole motivation for the HAEL theory is due to the difficulty of logically characterizing such a process, we must abandon any persistence of beliefs and resign ourselves to generating Nigel's beliefs from first principles each time we are interested in them. This begs the question: how do we assign any beliefs at all to Nigel regarding times at which nothing is happening to him? Well, we can distinguish memory from persistence and say that anything in the strongest evidential space is an incontrovertible fact as far as our agent is concerned, and his belief in it persists without the need for a default inference. The persistence default axiom is only used for domain circumstances which have a good chance of staying the same; beliefs about such circumstances are derived anew whenever queried, so as to take into account all possible conflicts of evidence.

In fact, the purposes for which we are constructing this formalization do not require a persistence axiom at all. In the Map Task, the only beliefs that need to persist are those relating directly to what an agent has perceived: his instructions for the task, the map before him and the utterances that have gone before. All his beliefs can be derived from these fundamental ones.

3.4 Belief ascription and adoption as defaults

In a world in which an objective truth exists, and agents' beliefs constitute a model of that truth which their constitution causes them to keep as accurate as possible, an agent may make the assumption that his own beliefs are likely to be shared by other agents, and that anything he knows another agent to believe is worth adopting as a belief himself. The effects of utterances in the model described here rely on the latter principle of belief adoption, and these can be taken as axiomatic because utterances reflect agents' own states of mind, about which they are never mistaken. On the other hand, belief ascription can be applied in the absence of certain knowledge of the other agent's state of mind, and therefore seems to be a principle worthy of defining an evidential strength in terms of HAEL. If it is used as a general principle, it must be overriden in specific instances in which either the partner's utterances or the conditions imposed on the domain contradict the ascription. But at this stage we must keep in mind that all theorems are shared

[1] That is to say, the nested version of the persistence axiom created by the necessitation principle applies in respect of his beliefs

between all agents anyway, so only premises that do not follow from the axioms need to be ascribed.

When modelling the map task, we take the features of the agents' reasoning processes as axiomatic, e.g.,, their ability to tell where something is on hearing a description of it in terms if things they know. The only information that they are party to which is not represented by sentences in the HAEL model is the positions of features on the maps and the instructions they receive at the beginning of the task. So the IG ascribes all his knowledge of the map features to the IF, but in the case of the route this is overridden, as he has been told directly that the IF does not have it. We do not need to ascribe axiomatic knowledge to other agents as agents credit each other with such knowledge anyway in accordance with the principle of necessitation. However, we must have some way of ascribing map information whilst picking out knowledge relating to route sections so that its ascription can be blocked.

The default rule for ascription is written as a separate axiom and default clause as described in section 3.1. What we do is assert in the initial conditions that information relating to the route sections is private to the IG at any time. This does not prevent the IG from subsequently teaching the IF the position of route sections, as privacy only defeats the default implication that they know where they are, and any other implication to the same effect, e.g., from a description, will go through. It may be found during the task that some other landmarks are private as well, in which case the privacy operator at the appropriate points in time will be implied for them.

$$\text{in } \tau_0 : \quad \forall A_1, A_2, T_1, T_2, C\; bel(A_1, T_1, \text{degree}(C)) \wedge \neg bel(A_1, T_1, priv(C)) \rightarrow bel(A_1, T_1, bel(A_2, T_2, \text{degree}(C))) \tag{66}$$

$$\forall A_1, A_2, T_1, T_2, C_1, C_2\; bel(A_1, T_1, desc(C_1, \text{prep}, C_2)) \wedge \neg bel(A_1, T_1, priv(C)) \rightarrow bel(A_1, T_1, bel(A_2, T_2, desc(C_1, \text{prep}, C_2))) \tag{67}$$

$$\text{in } \tau_2 : \quad \forall A, T, C\; \neg L_1 bel(A, T, priv(C)) \rightarrow \neg bel(A, T, priv(C)) \tag{68}$$

No constraint is placed on the relationship between the time at which an agent holds a belief and the time another agent believes it according to him as a result of ascription. Thus, if an agent comes to believe at any point that one of his landmarks is private, he must be aware that this is the case throughout the dialogue.

$$\text{in } \tau_0 : \quad \forall A, T_1, T_2, C\; bel(A, T_1, priv(C)) \rightarrow bel(A, T_2, priv(C)) \tag{69}$$

It is important that agents who believe that other agents believe things can reason that the other agents will in turn ascribe those beliefs. The ascription axiom above will itself be shared by the necessitation principle to give a nested ascription axiom. However, the default clause itself is not an axiom and is therefore not shared, so a separate default clause for the nested case must be added. The ascribed rules and separate default clause are as follows:

$$\text{in } \tau_0 : \quad \forall A_1, A_2, A_3, T_1, T_2, T_3, C\; bel(A_1, T_1, bel(A_2, T_2, \text{degree}(C))) \wedge \neg bel(A_1, T_1, bel(A_2, T_2, priv(C))) \rightarrow bel(A_1, T_1, bel(A_2, T_2, bel(A_3, T_3, \text{degree}(C)))) \tag{70}$$

$$\forall A_1, A_2, A_3, T_1, T_2, T_3, C_1, C_2\; bel(A_1, T_1, bel(A_2, T_2, desc(C_1, \text{prep}, C_2))) \wedge \neg bel(A_1, T_1, bel(A_2, T_2, priv(C_1))) \rightarrow$$

$$bel(A_1, T_1, bel(A_2, T_2, bel(A_3, T_3, desc(C_1, \text{prep}, C_2)))) \quad (71)$$

in τ_2 : $$\forall A_1, A_2, T_1, T_2, C \; \neg L_1 bel(A_1, T_1, bel(A_2, T_2, priv(C))) \rightarrow \neg bel(A_1, T_1, bel(A_2, T_2, priv(C))) \quad (72)$$

This is the simplest logical way of expressing ascription, but for practical implications the most efficient way is to introduce a separate predicate $mapped(C)$ referring explicitly to the presence of a landmark on an agent's map, and to define the ascription axiom in terms of this predicate as follows: $bel(A_1, T, priv(C)) =_d efbel(A_1, T, mapped(C)) \wedge \neg bel(A_1, T, bel(A_2, T, mapped(C)))$. This prevents unnecessary use of the ascription axiom. beliefs about privacy, so that if the sharing of a belief is blocked at any time by a contradiction that results from it, it cannot be assumed to be shared at any other time.

3.5 Separating direct from indirect evidence

From the foregoing it might appear that only two evidential spaces are required here; one for conclusions relying only on the assumption that well-formed communication has taken place, and one for conclusions which also make unsupported assumptions about correspondences between the maps. However, a model constructed in this way will tend to require the use of chains of reasoning to support the assertions that various utterances have not occurred. In order to capture the intuition that if an utterance has not occurred the agents don't need to deduce this from the failure of its effects to be manifest, we can make the following changes:

1. Weaken all the axioms (without making them defeasible) by placing them in a weaker evidential space, except for those relating to perception and memory, i.e., axioms 73 and 74,

2. Supply a default rule in the space into which the weakened axioms have been placed, asserting that agents typically do not believe facts which would be directly perceived or remembered.

Thus evidential space τ_1 contains defaults to the effect that no utterances are believed to have happened by either agent. These seem like unreasonable defaults, but they are only required to speed up implementation of the system rather than to change the states it generates. As an example, if it is required to prove that Fred can assume that Doris shares one of his landmarks, we must check that Fred does not have strong evidence that the landmark is not shared. One set of circumstances that would count as such evidence is the case in which that landmark plays a role in a description Doris has been given, but the described object is not vivid to Doris. To disprove such a possibility we must prove that no such descriptions are known to Doris. This requires that Doris has heard no descriptive utterances, and it is easiest to assert directly that she only heard such utterances that are known to have occurred. Since the axioms corresponding to reasoning have all been moved to τ_1, the defaults underlying this assertion can only be overriden directly by an utterance that is known to have occurred.

4 HAEL characterization of ATMS model behaviour

At this point we shall examine the sort of agent model we are leading up to. In the absence of considerations stemming from the requirements of implementing the model, and providing no detailed treatment of evidential reasoning for the time being, the problem is fairly simple, and a solution needs to be expounded at this point to enlighten the following discussion. We shall refer to this as the 'naïve solution' as it ignores some aspects of human reasoning which can be captured by the ATMS implementation but for which further extensions to the logical model would be required.

4.1 Axioms characterizing utterance effects

The effects of utterances in this model are the same as those in Perrault [10] and Appelt and Konolige [1] except that in this domain we can discount the possibility of agents either being mistaken or not telling the truth. In order to create a single description of the relationship between utterances and agents' states of belief, we will describe the preconditions of utterances, and show that the modelled effects follow from the addressee's perception of the utterance combined with the necessitation principle.

There are five utterance moods, assertions, queries, positive and negative responses and closures. The occurrence of an utterance is represented by an expression of the form $do(agent, time, mood, proposition)$. All utterances have the effect that when they occur, all agents in the domain instantly believe that they do so, and since perceptual events cannot be subject to doubt they go on so believing. These axioms are expressed as follows:

$$\begin{array}{rl} \text{in } \tau_0: & \\ \text{Perception:} & \forall A_1, A_2, T\; do(A_1, T, act, content) \rightarrow \\ & bel(A_2, T, do(A_1, T, act, content)) \qquad (73) \\ \text{Memory:} & \forall A_1, A_2, T_1, T_2\; bel(A_2, T_1, do(A_1, T_2, act, content)) \rightarrow \\ & bel(A_2, succ(T_1), do(A_1, T_2, act, content)) \qquad (74) \end{array}$$

The condition of making an assertion is that the speaker knows the proposition asserted; an agent hearing the utterance can deduce this knowledge on the part of the speaker, but because of the conditions of honesty and correctness the addressee can in fact derive the proposition itself as well. The effects of an assertion are therefore written:

$$\text{in } \tau_1: \forall A_1, A_2, T_1, T_2, P\; bel(A_1, T_1, do(A_2, T_2, assert, P)) \rightarrow bel(A_1, T_1, P) \qquad (75)$$

Example 1 *"The first section goes left of the swamp."*

leads the addressee to believe that $desc(1S, left\_of, swamp)$.

Note that a speaker's acquaintance with a concept — what she actually knows about it that constitutes having it vivid — cannot be put into words, and hence the utterance $do(A, T, assert, vivid(C))$ must be glossed as the expression of the belief rather than the proposition believed: "I have the C". The addressee in this case comes to believe the speaker has C vivid, but cannot herself get C vivid because the utterance does not contain the necessary information.

Queries divide into wh-questions, where the query relates to the actual whereabouts of a landmark (which we are treating as a proposition, vividness), and yes/no questions where the query relates to an ordinary proposition, such as a description. We use a single underlying utterance classification for all queries, the distinction between the surface forms being decided solely on the basis of the type of proposition being queried. In any case the precondition is that the speaker does not already have this information.

$$\begin{array}{l} \text{in } \tau_1: \forall A_1, A_2, T_1, T_2, P\; bel(A_1, T_1, do(A_2, T_2, query, P)) \rightarrow \\ \qquad bel(A_1, T_1, \neg bel(A_2, T_2, P)) \qquad (76) \end{array}$$

Example 2 *"Where is the swamp?"*

— causes addressee to believe $\neg bel(speaker, T, vivid(swamp))$.

Example 3 *"Is the swamp left of the palm beach?"*

— causes addressee to believe $\neg bel(speaker, T, desc(swamp, left\_of, palm\_beach))$.

Example 4 *"Do you have the swamp?"*

— causes addressee to believe $\neg bel(speaker, T, bel(addressee, T, vivid(swamp)))$.

Yes/no questions also have the precondition that the speaker does not believe the negation of P either, but this result is not used in the examples. This precondition makes no sense for wh-questions as vividness has no inverse.

A positive response to an assertion indicates that the dialogue component of which it forms a part is complete. It thus indicates that the original speaker's purpose in making the assertion has been fulfilled. Normally the speaker's purpose is to bring about mutual belief in the propositional content of her utterance, but in a task-oriented domain this is inevitable, so the addressee instead looks at the speaker's supervening goal, i.e., to achieve mutual vivid acquaintance with the map feature that is the subject of the description. So in response to utterance 1 the addressee will reply "OK" if she subsequently has the first section vivid.

$$\text{in } \tau_1 : \forall A_1, A_2, T_1, T_2, C_1, C_2\ bel(A_1, T_1, do(A_2, T_2, OK, desc(C_1, \text{prep}, C_2))) \rightarrow$$
$$bel(A_1, T_1, bel(A_2, T_2, vivid(C_1))) \tag{77}$$

A positive response to a yes/no question has the precondition that the agent making it believes the propositional content of the utterance that initiated the response. Thus the hearer of such a response infers the speaker's belief directly.

$$\text{in } \tau_1 : \forall A_1, A_2, T_1, T_2, C\ bel(A_1, T_1, do(A_2, T_2, OK, vivid(C))) \rightarrow$$
$$bel(A_1, T_1, bel(A_2, T_2, vivid(C))) \tag{78}$$

In both these cases, the proposition which the speaker of the positive response indicates she believes is in fact a degree of acquaintance. Thus the hearer of the response cannot come to believe it herself because the utterance does not contain the information that actually constitutes the acquaintance. A positive response to a yes/no question involving a real proposition, such as utterance 3, would allow the hearer to infer the proposition itself via the correctness principle.

To respond positively to a wh-question, the addressee asserts the proposition, a move that has the same effects as an initial assertion. In fact, such a move would also serve as a positive response to a yes/no question, as follows:

Example 5 *IG: "Do you have the swamp?"*

Example 6 *IF: "I have the swamp".*

An assertion such as this would have the same effect on the original speaker's beliefs as the answer "yes", but would also require a further reply.

Negative responses simply indicate the lack of any belief on the part of the responder that would enable her to make a positive response. The various English glosses, examples of which follow, are actually distinct speech acts, but their effects follow a single pattern, so they all have the same utterance type.[2]

[2] A system for providing English glosses of dialogues represented by utterances characterized in this way would have to refer to the type of the initializing utterance to generate the best English gloss for each negative response.

A negative response to an initialization whose content is a typical proposition indicates the responder's inability to make the expected use of that proposition, e.g.,, to get the described landmark vivid.

$$\text{in } \tau_1 : \forall A_1, A_2, T_1, T_2, C_1, C_2\; bel(A_1, T_1, do(A_2, T_2, not\_OK, desc(C_1, \text{prep}, C_2))) \rightarrow$$
$$bel(A_1, T_1, \neg bel(A_2, T_2, vivid(C_2))) \tag{79}$$

Where the propositional content of the initiating utterance is a degree of acquaintance, then it is the responder's vividness itself about this landmark that is denied:

$$\text{in } \tau_1 : \forall A_1, A_2, T_1, T_2, C\; bel(A_1, T_1, do(A_2, T_2, not\_OK, vivid(C))) \rightarrow$$
$$bel(A_1, T_1, \neg bel(A_2, T_2, vivid(C))) \tag{80}$$

As is the case for queries, there is only one underlying categorization for negative responses, the different surface forms corresponding to the different proposition forms knowledge of which can be denied.

Finally we need to characterize the "close" utterance type that appears in the JAM examples. This is used by IFs to indicate that they now have vivid acquaintance with a landmark they did not have earlier, and thus any plan to acquaint them with that landmark can be marked as successfully concluded. It is glossed as "I am done talking about the X", and has the precondition that X is vivid for the speaker. The effects are thus the same as those of saying "I have the X", the representations in the two cases being, respectively,

$$do(A, T, close, vivid(C)) \tag{81}$$
$$do(A, T, assert, vivid(C)) \tag{82}$$

4.2 Axioms characterizing the Map Task domain

The reasoning behaviour of the agents is specified according to the standard logic of belief, weak S5, with the proviso that all logical inferences are made in evidential space τ_1, with the strongest evidential space being reserved for beliefs resulting directly from perceptual evidence; this allows the expression of the autoepistemic rule that if an agent does not believe a perceptual event to have occurred, she can apply deductions from her resulting knowledge that that event did not occur. The defaults of things not being on agents' maps and of utterances not happening are not really domain defaults, though, merely logical conveniences to save representing explicitly the absence of all the things that agents perceive are not in their environments. The default rules are as follows:

$$\text{Silence:} \quad \text{in}\tau_1 : \quad \forall A_1, A_2, T, P \neg L_0 \neg bel(A_1, T, \neg do(A_2, T, mood, P)) \rightarrow$$
$$bel(A_1, T, \neg do(A_2, T, mood, P)) \tag{83}$$
$$\text{Absence:} \quad \text{in}\tau_1 : \quad \forall A, T, C \neg L_0 \neg bel(A, T, \neg mapped(C)) \rightarrow$$
$$bel(A, T, \neg mapped(C)) \tag{84}$$

Note that the conditions of evoking the defaults of silence and absence are expressed as lack of disbelief, in order that the defaults can take the form described in section 3.1. Negations of these default conditions (disbelief in silence or map absence) must be derived from beliefs that an utterance *has* happened or a landmark *is* mapped. Thus the D axiom asserting consistency of beliefs has to be placed in evidential space τ_0 to allow these negations to be derived in τ_0.

We also need two domain axioms to describe the special behaviour of the predicate *vivid*. Axiom 85 characterizes the propagation of vividness via descriptions, i.e., that if an agent knows a description of a landmark in terms of other landmarks that are vivid, then she has that vivid as well. Axiom 86 expresses the fact that if some landmark appears on an agent's map then they have that landmark vivid.

$$\text{Vividness:} \quad \forall A, T, C_1, C_2\, bel(A, T, desc(C_1, \text{prep}, C_2)) \wedge bel(A, T, vivid(C_2)) \rightarrow bel(A, T, vivid(C_1)) \tag{85}$$

$$\text{Presence:} \quad \forall A, T, C\, bel(A, T, mapped(C)) \rightarrow bel(A, T, vivid(C)) \tag{86}$$

Finally we need to characterize agents' knowledge that typically, landmarks that appear on one map appear in the same form on the other. This is expressed as a default in τ_2 as such conclusions can be overridden by the results of reasoning with the other axioms. described at the end of section 3.4, we will have a separate predicate for belief in privacy, allowing the time parameter to be dropped from it.

$$\text{Sharing axiom} \quad \forall A_1, A_2, T, C\, bel(A_1, T, mapped(C)) \wedge \neg bel(A_1, T, priv(C)) \rightarrow bel(A_1, T, bel(A_2, T, mapped(C))) \tag{87}$$

$$\text{Above ascribed} \quad \forall A_1, A_2, A_3, T_1, T_2, C\, bel(A_1, T_1, bel(A_2, T_2, mapped(C))) \wedge bel(A_1, T_1, \neg bel(A_2, T_2, priv(C))) \rightarrow bel(A_1, T_1, bel(A_2, T_2, bel(A_3, T_2, mapped(C)))) \tag{88}$$

$$\text{Durable privacy} \quad \forall A, T_1, T_2, C\, bel(A, T_1, priv(C)) \rightarrow bel(A, T_2, priv(C)) \tag{89}$$

$$\text{Above ascribed} \quad \forall A_1, A_2, T_1, T_2, T_3, C\, bel(A_1, T_1, bel(A_2, T_2, priv(C))) \rightarrow bel(A_1, T_1, bel(A_2, T_3, priv(C))) \tag{90}$$

$$\text{Base default} \quad \text{in } \tau_2 : \forall A, T, C \neg L_1 bel(A, T, priv(C)) \rightarrow \neg bel(A, T, priv(C)) \tag{91}$$

$$\text{Nested default} \quad \text{in } \tau_2 : \forall A_1, A_2, T, C \neg L_1 bel(A_2, T, bel(A_1, T, priv(C))) \rightarrow \neg bel(A_2, T, bel(A_1, T, priv(C))) \tag{92}$$

4.3 Worked examples

Let us examine how this scheme models what takes place in a simple dialogue generated by JAM. It was stated in section 1.1 that the logic describes a *history* in that all actions and belief states throughout the dialogue are represented concurrently in the scheme. We can indeed create such a model; we include the axiom schemata described so far, plus premises corresponding to what agents see on their maps and the utterances that occur throughout the dialogue. We can then generate the closure of this theory under inference, and read from it what any agent believes at any time throughout the dialogue.

Thanks to the nonmonotonic properties of the logic and the defaults of silence and absence mentioned above, the same initial state can be used to generate a *rolling history;* this is a complete history but generated using the object-level events only up to a certain time, and updated by the addition of premises corresponding to further events in the order in which they happen. When a new event is added, future belief states represented in the model may change, but consistency is maintained. Since the model can be used in this way, an implementation of it could be married to the JAM planner and used in the actual simulation; it would only require the implicit assumption in JAM that its belief state queries refer to the present to be made explicit at the interface. For times before the rolling present, the history will

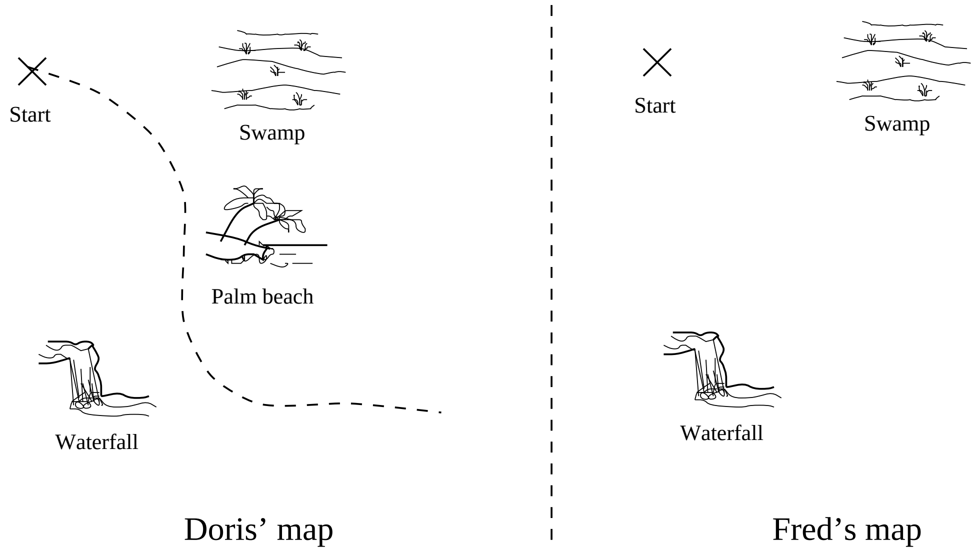

Figure 1: The relevant parts of the maps for the example

include exactly the same beliefs as the rolling history; this follows from the facts that agents cannot be aware of what utterances will happen in the future,[3] nor can they draw any conclusions about the past or present from their assumptions that no utterances will happen in the future.

Let us consider what happens in each case when negotiating the initial stages of a typical map, that in figure 1. Firstly, the following premises describe the agents' beliefs resulting directly from their maps: $1S$ represents the first section of the route, going left of the palm beach on the IG's map, $SW1$ is the swamp on both maps and $PB1$ is the palm beach.

$$\text{in } \tau_0: \quad \forall T\, bel(IG, T, mapped(1S)) \tag{93}$$
$$\forall T\, bel(IG, T, mapped(PB1)) \tag{94}$$
$$\forall T\, bel(IG, T, mapped(SW1)) \tag{95}$$
$$\forall T\, bel(IG, T, desc(1S, left\_of, PB1)) \tag{96}$$
$$\forall T\, bel(IG, T, desc(PB1, below, SW1)) \tag{97}$$

$$\forall T\, bel(IF, T, mapped(SW1)) \tag{98}$$

In addition to these, the agents have beliefs resulting from the instructions in which they are told that they can assume that the landmarks, with the exception of the route, are shared:

$$\text{in } \tau_1: \quad bel(IG, T_1, priv(1S)) \tag{99}$$
$$\text{in } \tau_2: \quad \neg bel(IG, T_1, priv(PB1)) \tag{100}$$
$$\forall T_1, T_2\, bel(IG, T_1, bel(IF, T_2, mapped(PB1))) \tag{101}$$
$$\neg bel(IG, T_1, priv(SW1)) \tag{102}$$
$$\forall T_1, T_2\, bel(IG, T_1, bel(IF, T_2, mapped(SW1))) \tag{103}$$

[3] Not until we develop a planner that uses the same belief database for its plans at least.

$$\text{in } \tau_2 : \quad \neg bel(IF, T_1, priv(SW1)) \tag{104}$$
$$\forall T_1, T_2\; bel(IF, T_1, bel(IG, T_2, mapped(SW1))) \tag{105}$$

As the model stands, we have not included a third-level default clause which would allow an agent thinking that the other has a landmark to assume that the other thinks they have it; we argue that such deeply nested beliefs are in any case unnecessary.

Finally, prior to any utterances, agents will draw inferences that what is mapped is vivid (using axiom 86) both for themselves and their partners.

$$\text{in } \tau_1 : \quad \forall T\; bel(IG, T, vivid(1S)) \tag{106}$$
$$\forall T\; bel(IG, T, vivid(PB1)) \tag{107}$$
$$\forall T\; bel(IG, T, vivid(SW1)) \tag{108}$$

$$\forall T\; bel(IF, T, vivid(SW1)) \tag{109}$$

$$\text{in } \tau_2 : \quad \forall T_1, T_2\; bel(IG, T_1, bel(IF, T_2, vivid(PB1))) \tag{110}$$
$$\forall T_1, T_2\; bel(IG, T_1, bel(IF, T_2, vivid(SW1))) \tag{111}$$

$$\forall T_1, T_2\; bel(IF, T_1, bel(IG, T_2, vivid(SW1))) \tag{112}$$

One of the possible dialogues generated by JAM in this situation runs as follows:

Example 7 *:*
IG: "The first section goes left of the palm beach."
IF: "I do not understand. Where is the palm beach?"
IG: "The palm beach is below the swamp".
IF: "OK. I am done talking about the first section.
IG: "OK."

Each sentence (including "OK") is represented as a separate utterance, each occurring at the time instant following the preceeding one. We will call these time instants $T_1, T_2 = succ(T_1), \ldots, T_6 = succ(T_5)$. The HAEL theory corresponding to the dialogue above is:

$$do(IG, T_1, assert, desc(1S, left\_of, PB1)) \tag{113}$$
$$do(IF, T_2, not\_OK, desc(1S, left\_of, PB1)) \tag{114}$$
$$do(IF, T_3, query, vivid(PB1)) \tag{115}$$
$$do(IG, T_4, assert, desc(PB1, below, SW1)) \tag{116}$$
$$do(IF, T_5, OK, desc(PB1, below, SW1)) \tag{117}$$
$$do(IF, T_6, close, vivid(1S)) \tag{118}$$

We shall examine the rolling-history case here as this is simpler than the full history which produces the same results anyway. In the initial situation, the IG believes weakly that the IF has the palm beach, but in this case decides on that basis to describe part of the route in terms of it. At T_1, the IG's beliefs behave as follows:

in τ_0 :

$$113, 73 \quad bel(IG, T_1, do(IG, T_1, assert, desc(1S, left\_of, PB1))) \tag{119}$$

$$119, N(73) \quad bel(IG, T_1, bel(IF, T_1, do(IG, T_1, assert, desc(1S, left\_of, PB1)))) \tag{120}$$

in τ_1 :

$$120, N(75) \quad bel(IG, T_1, bel(IF, T_1, desc(1S, left\_of, PB1))) \tag{121}$$

in τ_2 :

$$121, 110, N(85) \quad bel(IG, T_1, bel(IF, T_1, vivid(1S))) \tag{122}$$

After making the utterance, the IG believes that she has succeeded in conferring vividness of the first section upon the IF, but only has this belief in the weakest evidential space. As an effect of the utterance, ther IF will come to believe the proposition expressed but, not having the palm beach vivid, will be unable to draw any other conclusions, and thus fulfil the conditions for a negative reply. In this case, this is followed by a query resulting from the adoption of the IG's plan by the IF.

At T_2, the IG's beliefs behave as follows:

in τ_0 :

$$114, 73 \quad bel(IG, T_2, do(IF, T_2, not_OK, desc(1S, left\_of, PB1))) \tag{123}$$

in τ_1 :

$$123, N(79) \quad bel(IG, T_2, \neg bel(IF, T_2, vivid(PB1))) \tag{124}$$

$$124, N(86) \quad bel(IG, T_2, \neg bel(IF, T_2, mapped(PB1))) \tag{125}$$

$$125, 94, 87 \quad bel(IG, T_2, priv(PB1)) \tag{126}$$

The arrival at the belief that her palm beach is not shared by the IF causes the IG to abandon her belief that it is at any time vivid to the IF, and hence also her belief that the first section is vivid to the IF.

In this case the IF's subsequent question, "Where is the palm beach?", does not convey any extra information to the IG, who has already deduced that the IF does not have it. When the IG replies "The swamp is below the palm beach", she deduces in τ_2 that the IF then does have the palm beach vivid, by a proof isomorphic to that relating to the first section in formuale 119 to 122. Although the IG's belief that the palm beach is "private" holds at any time, her positive belief that the IF *does not* have it is restricted to times T_2 and T_3, so at T_4 she can believe the IF has it, and hence the first section as well.

For the IF, the IG's second utterance sets off the following chain of reasoning:

in τ_0 :

$$113, 73 \quad bel(IF, T_2, do(IG, T_2, assert, desc(1S, left\_of, PB1))) \tag{127}$$

$$127, 74 \quad bel(IF, T_3, do(IG, T_2, assert, desc(1S, left\_of, PB1))) \tag{128}$$

$$128, 74 \quad bel(IF, T_4, do(IG, T_2, assert, desc(1S, left\_of, PB1))) \tag{129}$$

$$116, 73 \quad bel(IF, T_4, do(IG, T_4, assert, desc(PB1, below, SW1))) \tag{130}$$

in τ_1 :

$$130, 75 \quad bel(IF, T_4, desc(PB1, below, SW1)) \tag{131}$$

$$131, 109, 85 \quad bel(IF, T_4, vivid(PB1)) \tag{132}$$

$$129, 75 \quad bel(IF, T_4, desc(1S, left\_of, PB1)) \tag{133}$$

$$133, 132, 85 \quad bel(IF, T_4, vivid(1S)) \tag{134}$$

The IF's response, "OK", indicates directly that she now has the palm beach vivid (as this was the subject of the description that elicited the response) and the IG now acquires this belief in τ_1. She can also deduce in τ_1 that the IF has the first section vivid at this point, so the IF's final utterance (which conveys this information explicitly) is redundant.

5 Reasons for wanting an implemented HAEL model

The original reason for wanting to implement the HAEL model by means of a theorem prover was simply to automate the tedious process of generating proofs and going through worked examples. It was subsequently found that the amount of work involved in determining and specifying the HAEL model would be fairly large, and the implementation started to appear as a necessity, principally for the following reasons.

5.1 Detection of mistakes resulting from nonmonotonicity

Generating a result in a monotonic logic is simply a matter of starting with the axioms and the premises of the model, and generating further sentences until the desired ones come up. If each stage is free of mistakes, the result will be valid. However HAEL like all nonmonotonic logics allows conclusions to be drawn from defaults which may be invalidated by the existence of a contradictory conclusion, even if this is not part of the proof under discussion. The problem that this causes for designing HAEL schemata is exacerbated by the very features that make it such a useful system. In particular, any contradiction should allow of resolution by following the reasoning that produced it back to all the default inferences on which it rests and cancelling the one in the weakest evidential space. Thus results in weaker spaces are susceptible to overturning if any of their premises lead to a potential contradiction elsewhere.

5.2 Ability to test different systems quickly

Various features of the logical model described in section 4 were decided on with the help of an implemented system. In particular, the rejection of general axiom schemata for the persistence and ascription of all types of belief was due to observation of the performance of implemented models using such defaults. We were in fact led to the discovery of the theoretical problems described in sections 3.2 and 3.4 by experiments with the implemented system.

However, it is not evident that the exact HAEL characterization we have described, or even its general form, is the only possible one. One alternative form that promises more flexibility could be derived by abandoning the three-place belief operator. Instead, we would have two types of proposition: those whose value is a function of time, and those formed by supplying the **holds** relation with one of the first type and a time identifier to produce a statement whose truth value is fixed. Beliefs would then be two-place operators, relating an agent to a proposition of the second type but counting as the first type themselves. Here are some examples from the stolen car problem:

$$home(Tulip) \qquad \text{(type 1)} \qquad (135)$$

$$\mathbf{holds}(home(Tulip), T_0) \qquad \text{(type 2)} \qquad (136)$$

$$bel(Fred, \mathbf{holds}(home(Tulip), T_0)) \qquad \text{(type 1)} \qquad (137)$$

$$\mathbf{holds}(bel(Fred, \mathbf{holds}(home(Tulip), T_0)), T_1) \qquad \text{(type 2)} \qquad (138)$$

While the formalism we used in the example could be translated straightforwardly to this form, the latter offers the possibility of abbreviating descriptions of aspects of the world that are true at individual times but which never change over time (such as beliefs in privacy – see page 19). For an exploration of such possibilities, an implemented system is essential, as any number of minor design choices will affect the behaviour of the entire system.

6 Possible routes to an implementation

The obvious way to create a theorem-prover for this formalism is to write one in Prolog. The language is naturally suited to theorem-proving as it is itself an implementation of a depth-first Horn clause resolution theorem prover. Furthermore the default inferences in HAEL are handled neatly by Prolog's implementation of negation-as-failure. If something cannot be proved in one evidential space, a lack of self-awareness of it can be added to the next strongest space to enable appropriate default conclusions to be drawn.

There are a number of possibilities for such a system based on this principle, but all of them suffer from the same basic limitation, namely that since the logic of belief is only semidecideable, and default conclusions in one space require the absence of contradictory conclusions in the space below, the system as a whole is not even semidecideable. This means that if the implementation is organized in an arbitrary manner, there are propositions which are logically valid for which it will never find a proof. Fortunately we can attempt to organize the implementation in such a way that operations likely to give rise to a proof are carried out before equally valid ones which will more probably lead to endless search, so for the majority of typical queries we will get an answer.

We also have the problem that where there are many possible paths for deriving some clause that will be important in a proof, Prolog will by default explore them all, duplicating any subsequent reasoning for each path. The performance of Prolog-based theorem provers can be improved by keeping a log of all literals which have been refuted or for which the system has failed to find a refutation. When an attempt is made to find a resolvent with a literal, this log is checked and, if the literal is subsumed by one which has been refuted or subsumes one for which no refutation was found, the search is aborted and success or failure signalled as appropriate. Keeping such a log also enables the system to recognize when a search has come back to a literal that is congruent with one higher up the same resolution tableau. If this happens, continuing the search will yield no results, as the same literal will be repeatedly tested. However, allowing the search to fail in such circumstances renders the system unsound as pointed out in section 7.4.

6.1 Direct representation of axioms as Prolog clauses

Prolog is an implementation of a resolution theorem prover, restricted to Horn clauses and doing a depth-first search. In order to use the Prolog interpreter to attack problems in HAEL/weak S5 directly, we have to convert all our formulae into disjunctive normal form, and hope that they turn out to be expressible as Horn clauses.

For the system as it stands, all the domain axioms take the form of implications between nonnegated literals, so they do not cause a problem as they translate straightforwardly into Horn clauses, as does the D axiom. However, the K axiom schema as it stands (1) refers to agents' beliefs in the truth of implications, and if it were included in this form would require the inclusion of all the rewrite rules of propositional logic for manipulating beliefs.

Furthermore, it cannot be rewritten in any clausal form, as a belief in an implication does not imply or follow from an implication of beliefs. However, we can get a direct implementation in Prolog if we restrict the contents of beliefs to literals; we shall refer to the language incorporating this restriction as BNF (Belief Normal Form). If P is a term, A_n are agents and T_n are times, the literals we are allowed in BNF are the following:

$$
\begin{array}{l}
P \\
\neg P \\
\\
bel(A_1, T_1, P) \\
bel(A_1, T_1, \neg P) \\
\neg bel(A_1, T_1, P) \\
\neg bel(A_1, T_1, \neg P) \\
\\
bel(A_2, T_2, bel(A_1, T_1, P)) \\
bel(A_2, T_2, bel(A_1, T_1, \neg P)) \\
bel(A_2, T_2, \neg bel(A_1, T_1, P)) \\
bel(A_2, T_2, \neg bel(A_1, T_1, \neg P)) \\
\neg bel(A_2, T_2, bel(A_1, T_1, P)) \\
\neg bel(A_2, T_2, bel(A_1, T_1, \neg P)) \\
\neg bel(A_2, T_2, \neg bel(A_1, T_1, P)) \\
\neg bel(A_2, T_2, \neg bel(A_1, T_1, \neg P)) \\
\\
bel(A_3, T_3, bel(A_2, T_2, bel(A_1, T_1, P))) \\
\vdots
\end{array}
$$

To derive all the consequences of the K axiom and necessitation principle that can be expressed in Horn clauses of BNF, we need to combine their application to "jump over" the non-BNF formulae. If we start with a clausal axiom such as (11) we can, as in the example, derive the axioms (14). These can in turn be combined with the M theorem to give the following:

$$
\begin{array}{ll}
14, 9 & \vdash \forall A, T\; bel(A, T, \neg P_1) \wedge \ldots \wedge bel(A, T, \neg P_n) \\
& \text{(excluding } bel(A, T, \neg P_k)) \rightarrow bel(A, T, P_k)
\end{array}
\tag{139}
$$

These can be expressed as Horn clauses with literals in BNF. All we have to do now to implement the system is start with a list of axioms and assert Prolog clauses for all of them, plus all the axioms that are generated by applying the above procedure to them, also by applying it to these new axioms and so on. We present no formal proof that restricting inference to resolution plus the generation of axioms by the steps described here will not affect the completeness of deduction with respect to sentences in BNF, though we appear to have captured all the routes by which sentences in BNF can be derived. The problem of course is that we have to generate an infinite number of axioms. To avoid this we place a restriction on the depths of nested beliefs with which we will reason; the number of axioms grows exponentially with the depth of nesting we are allowing.

6.2 Implementing a resolution theorem prover

The last subsection described how we could express a problem in weak S5 directly as a Prolog program. We can instead implement an interpreter for weak S5 in

Prolog, with apparently little loss of efficiency. In doing this we stuck to BNF and depth-first search, but allowed non-Horn clauses. The main advantage in doing this is that rather than starting with all the axioms we will ever need for reasoning about belief, we can wait until we need to resolve a nested belief literal, then specifically generate clauses containing literals that will resolve with it.

The problem is that there are still many of them; for any weak inference relating to nested beliefs, there are many ways of making it from the same premises by combining the strong inferences that can be generated by the K axiom and necessitation principle with "weakening" ones derived from the D axiom. Unfortunately, blocking any combination of these renders the system incomplete.

6.3 Off-the-shelf systems

6.3.1 The LeanTaP theorem prover

LeanTaP (standing for Lean Tableau-based theorem-Prover) is a compact implementation of a first-order theorem-prover in Prolog, due to Bernhard Beckert and Joachim Posegga. It incorporates direct manipulation of conjunctions, disjunctions and quantifiers and thus allows the K axiom and M theorem to be entered directly, and the consequences of the necessitation principle to be represented fairly simply.

Unfortunately it incorporates a mixed breadth/depth first search strategy, which although successful with the simple problems at which it performs well, does not appear to be adequate for more complex problems. Indeed the authors state that for problems close to the practical limit of what can be achieved by theorem provers, a system with a complex control strategy involving meta-reasoning will out-perform LeanTaP.

6.3.2 OTTER

OTTER 3.0 [8] is a general-purpose first-order theorem-prover, designed for interactive use. It works in the domain of first-order formulae with equality, and presents the user with a vast array of weighting parameters, flags and rules of inference. It could be used for the problem described here, though would be very hard to interface to the planner the way the Prolog theorem provers could. The main considerations would be:

- How to formalize the weighting techniques built into the Prolog theorem prover,
- Incorporation of the distinction betwen HAEL evidential spaces.
- Implementation of the default reasoning system
- Possible use of OTTER's hyperresolution rules to implement the circumscription steps described in section 7.

Because it is such a complex, flexible and well supported system, success in all these areas would be virtually guaranteed, though achieving it would take quite a lot of work.

7 Beyond the naïve solution: the need for circumscription

Circumscription is the logical manoevure of asserting that the only individuals of which some predicate is true are those for which we can show that it is true. It is a form of nonmonotonic inference as it may subsequently become possible to show that the predicate is true for other individuals, but it is not the same as a default inference, since a circumscriptive axiom allows stronger conclusions to be drawn. For instance, if we know that birds usually fly but Tweety is an ostrich and ostriches don't fly, by default logic we can conclude that if Fred is a bird and is not an ostrich, then Fred flies. But circumscription allows us to conclude that for all X, if X is a bird and not an ostrich then X flies.

It appears that in 'toy' domains such as syllogisms, conclusions drawn from circumscriptive reasoning are more than defaults; they follow soundly from the assumption that the problem as stated includes all the information necessary for its solution. For instance, a person given the 'missionaries and cannibals' problem will feel cheated if subsequently told that the correct answer is that there is a bridge just down the river and they cross by that.

7.1 Conclusions requiring circumscription: the waterfall problem

It seems implausible that such an abstruse mode of reasoning should play a part in anything so mundane as the Map Task, but the very artificiality and closedness of the domain invite such reasoning. In particular, if Fred says, "I don't know where the swamp is", Doris replies "It's below the waterfall", and Fred concludes, "Fine, I have it now", then Doris can conclude that even though he has not said so, Fred has the waterfall, since he appears to have used it to work out where the swamp is. To reach this conclusion, Doris has circumscribed her knowledge of descriptions of the swamp which may be adequate for Fred and of which he is aware, to conclude that the only one is the one she gave him, i.e., below the waterfall. She knows Fred doesn't have the swamp on his map, or have any other useful description of it, or he would have known where it was to start with, so by applying a rule that states that vivid oblects must be on the map or be adequately described, she concludes that the description in terms of the waterfall is adequate and hence Fred must have the waterfall.

7.2 How the ATMS implements circumscription

The ATMS implementation embodies a sort of implicit circumscription in that when the effect of an utterance constituting a reply is added to the database, it has the effect of consolidating or negating assumptions rather than adding new information. Thus, if Doris tells Fred something about the swamp and Fred says he does not understand, Doris then contradicts all the groups of assumptions which at the time led to the belief that Fred did have the swamp vivid, which results in the retraction of that belief in the network. The belief itself cannot be contradicted, as contradictions cannot be removed and it would then be impossible to restore that belief in the case that Fred subsequently managed to get the swamp vivid.

Similarly, if Fred at any point asserts that he does have the swamp, and as far as Doris was concerned this was previously a matter of assumption, then Doris can deduce, not only that Fred has the swamp, but that one of her sets of assumptions which would have led to this conclusion is in fact true and can be elevated to the status of a premise. In the waterfall example above, Doris goes from assuming that Fred has the waterfall vivid to knowing that he does, and unless she has already described the waterfall to him in terms of other landmarks, this means making a premise of her belief that Fred has the waterfall on his map.

7.3 How circumscription can be represented in HAEL

The classical application of circumscription, due to Konolige, involves a second-order axiom stating that, eg. when circumscribing the property of redness, there is no predicate which is true of all the objects we know to be red but which is not true of all red objects. Circumscriptive axioms cannot simply be included in a formalism though, because a choice must be made as to which predicates to circumscribe at which points during reasoning. Furthermore, it is undesirable to introduce second-order logic into the description of what is happening as this reduces our credibility in calling it a standard logical framework.

We can avoid the need for second-order axioms by a combination of extra first-order axioms relating to the predicates we wish to circumscribe. In order to capture the waterfall example in an intuitively satisfying way, we must represent Doris' knowledge that Fred could not have found out anything about her map other than by her telling him it, and her certainty that she has not told him anything other than what she knows she has told him. In order to reason from knowing that Fred has the waterfall to knowing that he has the swamp, Doris would employ the following axioms (these are repeated from earlier sections):

$$\begin{array}{rl}
\text{in } \tau_0 & \\
\text{Perception (73)} & \forall A_1, A_2, T\; do(A_1, T, act, content) \rightarrow \\
& bel(A_2, T, do(A_1, T, act, content)) \\
\text{Memory (74)} & \forall A_1, A_2, T_1, T_2\; bel(A_2, T_1, do(A_1, T_2, act, content)) \rightarrow \\
& bel(A_2, succ(T_1), do(A_1, T_2, act, content)) \\
\text{in } \tau_1: & \\
\text{Assertion (75)} & \forall A_1, A_2, T_1, T_2, P\; bel(A_1, T_1, do(A_2, T_2, assert, P)) \rightarrow \\
& bel(A_1, T_1, P) \\
\text{Vividness (85)} & \forall A, T, C_1, C_2\; bel(A, T, desc(C_1, \text{prep}, C_2)) \wedge \\
& bel(A, T, vivid(C_2)) \rightarrow bel(A, T, vivid(C_1)) \\
\text{Presence (86)} & \forall A, T, C\; bel(A, T, mapped(C)) \rightarrow bel(A, T, vivid(C))
\end{array}$$

In order to follow this chain in the opposite direction, i.e., to show that if Doris knows Fred has the swamp this must be because he has the waterfall, these axioms need to be supplemented with their negative counterparts. Basically what we are doing is circumscribing the belief operator to say that the only beliefs an agent has are those that are well founded according to the axioms of belief we know about. So for any belief sentence, we take all the implication sentences for that belief and add a set of new implications to the effect that if *at least one* of the conditions for this belief in *each* implication is false, then the belief itself is false. We need only do this for the axioms listed above in order to eliminate situations in which something "mysteriously becomes vivid". The following sentences thus need to be added:

$$\begin{array}{rll}
\text{From 73, 74:} & \forall A_1, A_2, T_1, T_2\; bel(A_2, succ(T_1), do(A_1, T_2, act, content)) \rightarrow & \\
& (do(A_1, succ(T_1), act, content) \vee & \\
& bel(A_2, T_1, do(A_1, T_2, act, content))) & (140) \\
\text{From 75:} & \forall A_1, A_2, T_1, T_2, C_1, C_2\; bel(A_1, T_1, desc(C_1, \text{prep}, C_2) \wedge & \\
& \neg mapped(C_2)) \rightarrow & \\
& bel(A_1, T_1, do(A_2, T_2, assert, desc(C_1, \text{prep}, C_2))) & (141) \\
\text{From 85, 86:} & \forall A, T, C_1\; bel(A, T, vivid(C_1)) \rightarrow bel(A, T, mapped(C_1)) \vee &
\end{array}$$

$$(\exists C_2\, bel(A,T,desc(C_1,\text{prep},C_2)) \wedge bel(A,T,vivid(C_2)) \quad (142)$$

There are two problems with this. Firstly, as there is no earliest time point, agents can still have descriptive knowledge appear out of nowhere simply through having always had it. This is fixed simply enough by establishing an 'earliest time point' at which agents do not believe any utterances have occurred. The other problem is more pernicious. The object of the exercise is to show that there is no way a landmark could have become vivid for an agent as a result of a particular utterance other than through a particular other landmark already being vivid. However, the above axioms permit the landmark to become vivid in the case in which another landmark in terms of which it is already described is made vivid. This again ultimately allows a model in which vividness appears from nowhere, via a magical chain or loop of descriptions.

One way to stamp out this problem is to rewrite the vividness causality axiom as an axiom schema expressing the condition for a landmark being vivid nonrecursively, i.e., it must either be mapped, or described in terms of something that is mapped, or described in terms of something that is described in terms of something that is mapped, and so on with the proviso that none of the somethings are the same. With this schema added, the only way that a new description can make something vivid is by relating it to something else which must already be vivid, and the waterfall problem can be solved. Obviously in this case some upper limit must be set for the lengths of chains of descriptions connecting vivid objects to mapped objects; we are thus compromising the generality of the description.

To maintain generality we must instead add an extra axiom to the effect that *nothing* can become vivid for an agent except at a time at which he is offered a new description which refers exclusively to things he already has vivid. This is intuitively valid, but does not follow from the above axioms.

$$\forall A_1, T, C_1\, \neg bel(A_1,T,vivid(C_1)) \wedge bel(A_1,succ(T),vivid(C_1)) \rightarrow$$

$$\exists A_2, C_2, C_3\, do(A_2,succ(T),assert,desc(C_3,\text{prep},C_2)) \wedge bel(A,T,vivid(C_2)) \quad (143)$$

The addition of this axiom also makes the circumscription axioms derived above unnecessary for deriving the vividness of something that is mentioned with useful results, as it goes directly from the change in $A's$ acquaintance with a landmark in an interval to the applicability of some description offered in that interval.

7.4 Implementations of circumscription

The addition of any circumscription axioms has the immediate effect of slowing down the implementation, for the obvious reason that the theorem prover will now search from a consequent to its antecedent and thence back to the consequent, making a straightforward depth-first approach impossible. This is the case whether axiom 143 is used, or only the axioms that are derived from the domain rules by circumscribing the belief operator. There are a number of possible ways round this problem:

- Use some other search algorithm.
- Keep a list of intermediate conclusions and terminate searches that repeat themselves.
- Use different operators to represent beliefs derived from causal and evidential rules.

The first option results in extremely inefficient search in Prolog, as backtracking from chains that are terminated because they have grown too long involves continually undoing and repeating all inferences.

The second option has been tried with some success, but renders the implementation unsound; in particular, the waterfall problem *can* be solved without axiom 143 with such a system, as an attempt to falsify the conclusion that the waterfall is vivid will fail rather than repeatedly try to find another landmark that becomes vivid at the time the description involving the waterfall is given.

The third option has not yet been explored, but it has been used elsewhere [13, 9, pp. 502–507] to capture the notion that it makes sense to reason about other consequences of the possible antecedents of an observed event, but not about other antecedents of its possible consequences. It may be possible to represent the different types of belief involved by placing the same operator in different HAEL spaces, as belief derived from evidence can usually serve as the antecedent for causal rules, though how this would complicate the existing hierarchy of spaces is unclear.

8 Prospects for covering reasoning about ambiguous maps

8.1 Resolving conflicts

Recognizing an inconsistent interpretation most naturally means spotting conflicting defaults of a similar type. However, if these are in the same evidential space, the model dies.

There are many ways in which different interpretations of referring expressions can be put in different evidential spaces; completely arbitrarily, for instance. However, there does not seem to be a way of avoiding conflicts within a single space short of putting every interpretation in a different space. This may be workable, but is not intuitively good.

8.2 Coping with disjunctions

Vividness, disjunction and standard identity don't mix: given that identical terms can be substituted without changing the truth value of a sentence, we derive the rule $vivid(X) \land X = Y \to vivid(Y)$. With belief in this rule, an agent that knows $(A = B \lor A = C) \land vivid(B) \land vivid(C)$ will believe $vivid(A)$. This is a problem specifically with the possibility of mistaken beliefs of the sort subsumed by vividness; in each disjunct, the actual propositions subsumed by $vivid(A)$ are different, so they should not be represented by a single term.

Interestingly enough, this is precisely the sort of inference that can't be drawn if the axiom for deriving vividness is based on an autoepistemic rule; if τ_2 contains the rule $L_1 vivid(X) \land L_1 X = Y \to vivid(Y)$, this will require a belief in one certain identity to allow vividness to propagate from one concept to another. A logical characterization of the ambiguous version of the map task making use of this feature would require further exploration of the ways in which inferences are curtailed in such cases to make sure that they coincide with the restrictions on the sound inferences that can ve made with vividness.

9 Alternatives to HAEL

9.1 Probabilistic logic

Some form of probabilistic reasoning may be best suited to modelling beliefs, because

- It allows the distinction between predictive and diagnostic information to be expressed formally. This distinction is made implicitly by the ATMS, and it is modeling it in HAEL that gives rise to the necessity for an explicit closed world assumption.
- Probabilistic logic has a better defined model theory than nonmonotonic logic; since the ATMS deals in models and performs operations exclusively on them, there may be a better correspondence here than with HAEL.

On the down side, it appears that people working with uncertainty have less motivation to get the numbers exactly right every time. Thus doubt is cast on the well-foundedness of some probabilistic systems and hence on an ATMS-based system described in terms of them.

References

[1] Douglas Appelt and Kurt Konolige. A practical nonmonotonic theory for reasoning about speech acts. Technical Report 432, Artificial Intelligence Center, SRI International, April 1988.

[2] Philip R. Cohen and Hector J. Levesque. Persistence, intention and commitment. Technical Report 415, Artificial Intelligence Center, SRI International, February 1987. Reprinted as [3].

[3] Philip R. Cohen and Hector J. Levesque. Persistence, intention and commitment. In Phillip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 33–69. MIT Press, Cambridge, MA, 1990.

[4] Matther L. Ginsberg. Multivalued logics: A uniform approach to reasoning in artificial intelligence. *Computational Intelligence*, 1988.

[5] Joseph Y. Halpern and Yoram Moses. A guide to the modal logics of knowledge and belief: Preliminary draft. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, volume 1, pages 480–490, 1985.

[6] Kurt Konolige. Hierarchic autoepistemic theories for nonmonotonic reasoning: preliminary report. Technical Report 446, Artificial Intelligence Center, SRI International, August 1988.

[7] J. McCarthy and P.J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*. Edinburgh University Press, 1969.

[8] William W. McCune. OTTER 3.0 reference manual and guide. Technical Report ANL-94/6, Argonne National Laboratory, January 1994.

[9] Judea Pearl. *Probabilistic reasoning in intelligent systems.* Morgan Kaufmann, 1988.

[10] C. Raymond Perrault. An application of default logic to speech act theory. In Phillip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 161–185. MIT Press, Cambridge, MA, 1990.

[11] Yoav Shoham. *Reasoning about change.* MIT press, 1988.

[12] Jasper Taylor. An ATMS-based belief model for dialogue simulation. Technical Report HCRC/RP-45, Human Communication Research Centre, University of Edinburgh, October 1993.

[13] Jasper Taylor. *A Multi-Agent Planner for Modelling Dialogue.* PhD thesis, School of Cognitive and Computing Sciences, University of Sussex, 1994.