SRI International

Technical Report 446R • April 1991

HIERARCHIC AUTOEPISTEMIC THEORIES FOR NONMONOTONIC REASONING: Preliminary Report

Kurt Konolige

Artificial Intelligence Center
Center for the Study of Language and Information
Computer and Engineering Sciences Division

This research was supported by the Office of Naval Research Contract N00014-85-C-0251, by subcontract from Stanford University under Defense Advanced Research Projects Administration Contract N00039-84-C-0211, and by a gift from the System Development Foundation.

333 Ravenswood Avenue • Menlo Park, CA 94025-3493 • (415) 326-6200 • FAX: (415) 326-5512 • Telex: 334486

Abstract

Nonmonotonic logics are meant to be a formalization of nonmonotonic reasoning. However, for the most part they fail to embody two of the most important aspects of such reasoning: the explicit computational nature of nonmonotonic inference, and the assignment of preferences among competing inferences. We propose a method of nonmonotonic reasoning in which the notion of inference from specific bodies of evidence plays a fundamental role. The formalization is based on autoepistemic logic, but introduces additional structure, a hierarchy of evidential spaces. The method offers a natural formalization of many different applications of nonmonotonic reasoning, including reasoning about action, speech acts, belief revision, and various situations involving competing defaults.

Contents

1 Introduction

The nonmonotonic character of commonsense reasoning in various domains of concern to AI is well established. Recent evidence, especially the work connected with the Yale Shooting Problem (see [4]) has illuminated the often profound mismatch between nonmonotonic reasoning in the abstract and the logical systems proposed to formalize it. This is not to say that we should abandon the use of formal nonmonotonic systems; rather, it argues that we should seek ways to make them model our intuitive conception of nonmonotonic reasoning more closely. Generally speaking, current formal nonmonotonic systems suffer from two shortcomings:

1. They have no computationally realizable implementation.

2. They have only limited means for adjudicating among competing nonmonotonic inferences.

Reviewing the current major formalisms in this regard: Circumscription [11] and related model-preference systems [17], default logic [15], and autoepistemic (AE) logics [13; 7] are computationally intractable. The various proposals based on the notion of *defeasible rules* (see, for example, [14]) have yet to be given an implementation. The standard means of arriving at an implementation is to restrict the formal language, which restricts the expressivity of the resulting system, often to a rather severe extent.

The importance of having a flexible means for deciding among competing nonmonotonic inferences has become clear in the recent debate over the Yale Shooting Problem. It also arises in other contexts, such as taxonomic hierarchies [2] or speech act theory [1]. Prioritized circumscription [9] gives circumscription the capability of assigning priorities to various default assumptions. To some extent, preferences among default inferences can be encoded in AE and default logics by introducing auxiliary information into the statements of defaults. This method, however, does not always correspond satisfactorily with our intuitions. The most natural statement of preferences is with respect to the multiple *extensions* of a particular theory; that is, we prefer certain extensions because the default rules used in them have a higher priority than those used in alternative extensions.

Hierarchic autoepistemic logic, or HAEL, is a modification of autoepistemic logic [13] that addresses issues of implementation and priorities. In

HAEL, the primary structure is not a single uniform theory, but rather, a collection of spaces linked in a hierarchy. Spaces represent different sources of information available to an agent, and the hierarchy expresses the way in which this information is combined. For example, in representing taxonomic defaults, more specific information would take precedence over more general attributes. HAEL thus permits a natural expression of preferences among defaults and, in general, a more natural translation of our informal conception of nonmonotonic reasoning into a formal system. Further, given the hierarchic nature of the relation among spaces, there is a well-founded constructive semantics for the autoepistemic operator, in contrast to the usual self-referential fixedpoints. We can then easily arrive at computational realizations that make use of resource-bounded inference methods.

HAEL has been implemented and integrated with KADS, a resolution theorem-proving system for commonsense reasoning [3]. We have developed axiomatizations for reasoning about action, a preliminary form of belief revision, and speech-act theory. Currently, the resolution system and speech-act axiomatization are being employed in a natural-language generation system [1].

This paper has the following outline. The deficiencies of AE logic are pointed out in Section 2. In Section 3, we present an informal overview of HAEL, its relation to AE logic, and its applicability for nonmonotonic reasoning. Section 4 contains the formal characterization of HAEL, including its semantics and characterization in terms of *stable sets*. In Section 5, we list several extensions and problematic issues for HAEL.

2 Autoepistemic Logic

Autoepistemic logic was originally developed by Moore [13] as a means of formalizing reasoning about one's own beliefs. Other research, especially [7; 5], suggests that default reasoning might also be represented in terms of reasoning about self-belief. For example, consider the default statement

> Bats normally fly. (1)

Recasting this in terms of self-belief, we might say something like

> If I know that x is a bat, then I'll assume x flies unless I have information to the contrary. (2)

The key phrase here is the "unless" clause, which makes the rule inapplicable in the presence of conflicting information.

In AE logic, self-belief is represented by using a modal operator L. The construction $L\phi$ is intended to mean: "the sentence ϕ is one of my beliefs." The AE logic version of sentence (2) above is the schema

$$LBx \wedge \neg L \neg Fx \supset Fx \; . \tag{3}$$

The predicate B is the property of being a bat, and the predicate F the property of flying. The astute reader will note that there is no L operator on the conclusion of the implication; the reason for this is part of the technical subtleties of AE logic, and may be found in [5].

The formalization of AE logic is a good approximation to default reasoning: in [6], we show how the correct representational scheme can embody some of the main features of this reasoning. However, as we mentioned in the introduction, there are still significant problems, which we now discuss.

2.1 Priorities

Often defaults will conflict, and it is an important part of default reasoning to be able to decide which of a conflicting pair should dominate. For example, we might have a default that normally mammals do not fly. Since bats are mammals, an individual bat x has defaults for both flying and not flying. In this case, it is clear that we should choose the default that relies on the more specific information, since bats as a subclass of mammals generally do fly.

In AE logic, these defaults are represented as

$$\begin{aligned} &LBx \wedge \neg L\neg Fx \supset Fx \\ &LMx \wedge \neg LFx \supset \neg Fx\ , \end{aligned} \tag{4}$$

where M is the property of being a mammal. The usual way in which priorities are specified for defaults is to modify the mammal default by adding an extra part to the $\neg LFx$ atom (see [16]). This becomes:

$$LMx \wedge \neg L(Fx \vee Bx) \supset \neg Fx\ . \tag{5}$$

The presence of Bx under the self-belief operator means that if an individual x is known to be a bat, the default rule for mammals will not be applicable.

In cases where the conflict between defaults is evident, as in this example, this modification will work. However, consider the case of complicated domain knowledge in which it is difficult to predict whether or not two properties (call them P and Q) conflict. Suppose that bats normally have property P, and mammals property Q:

$$\begin{aligned} &LBx \wedge \neg L\neg Px \supset Px \\ &LMx \wedge \neg L\neg Qx \supset Qx\ . \end{aligned} \tag{6}$$

Now we have a dilemma. If P and Q actually do conflict for an individual bat x, then we should rewrite the second sentence to disable the default for x, or else there is an unresolved conflict. But if we do this, and the properties do not conflict, then we give up concluding that x has property Q. So without explicit knowledge about the relationship of P and Q, it is impossible to tell how the default for property Q should be written.

So far, we have been unable to find a simple solution to this problem, although by complicating the representation of defaults it may be possible. But priorities on default rules are only part of the problem: more generally, we will have priorities on the evidence that we use as the input to defaults. Consider, for example, the classic Nixon diamond:

Republicans are normally Hawks.
$LRx \wedge \neg L\neg Hx \supset Hx$

Quakers are normally Pacifists.
$LQx \wedge \neg L\neg Px \supset Px$ (7)

Nixon is a Quaker and Republican.
$Qn \wedge Rn$

It is a common intuition that these defaults truly conflict, that is, it is impossible to decide on this basis whether an individual like Nixon is a hawk or a pacifist.

Now suppose that we have uncertain information about Nixon: we are absolutely sure that he is a Republican, but have only weak positive support for him being a Quaker. In this case, we would have a preference for applying the default for Republicans, since its inputs are so much more certain. What this example suggests is that preferences or support among different bodies of input evidence is just as important as the defaults themselves in deciding which of a conflicting pair of defaults to apply. In fact, we can treat the taxonomic example in terms of priorities among evidence: we prefer applying a rule whose evidence is more specific. Several systems have been proposed in which general rules of evidence preference are used to guide the adjudiction of conflicting defaults [10; 14]. However, there is no explicit formal mechanism in AE logic (or in any of the other main nonmonotonic formalisms) to support even a basic theory of evidence.

2.2 Computational Issues

Autoepistemic logic, when based on a first-order language, is not even semidecidable [8]. Given this, there are two ways in which we might find an automatable proof theory for the logic.

1. Use a weaker language.

2. Use an incomplete inference procedure.

The first solution, weakening the language, has been investigated only to the extent that the propositional case is known to be decidable [12]. However, this alternative is undesireable because it can severly limit the expressivity of the language, and the ability to formalize default statements in a simple manner.

The second alternative, a sound but incomplete inference procedure, is one that has been used with success in many automated reasoning systems for AI that employ a first-order language. Generally we are interested in only a small fraction of the consequences of a set of proper axioms, and we can use heuristic procedures to guide the construction of the appropriate proofs.

Heuristic proof methods for first-order logic depend on the fact that sound local rules of inference exist. By *local* we mean that the input to these rules is a fixed number of sentences that have been established as provable. For example, *modus ponens* is a local inference rule which takes two input sentences of the form A and $A \supset B$, and returns the sentence B. For AE logic, however, there cannot be any local rules for many sentences involving the L operator. The reason is that the definition of an AE theory (called an *extension*) involves a fixedpoint construction. In effect, it is generally impossible to tell whether an arbitrary sentence of the form $\neg L\phi$ is a theorem without having already determined the complete set of theorems. In effect, this precludes us in the general case from having local rules of inference that are sound.

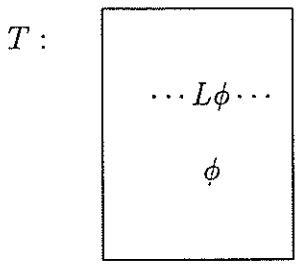

Figure 1: Autoepistemic semantics

3 Hierarchic Autoepistemic Theories

Hierarchic AE logic is derived from AE logic by splitting a uniform belief set into components, called *spaces*. In AE logic, an agent is assumed to have an initial set of premise sentences, A. The language of A contains an operator L for talking about self-belief: $L\phi$ is intended to mean that ϕ is one of the agent's beliefs. A belief set T that is derivable from the premises A by an ideal agent will be *stable* with respect to self-belief, that is, a sentence ϕ is in T if and only if $L\phi$ is in T. This interpretation of L is clearly self-referential, since it refers to the theory in which L itself is embedded (see Figure 1).

In the hierarchic modification of AE logic, the dependence of L on T is broken by dividing T into a hierarchy of spaces, and indexing L so that it refers to spaces beneath it in the hierarchy. For example, we might divide T into two spaces, T_0 and T_1, with T_1 succeeding T_0 in the hierarchy (see Figure 2). Space T_1 may contain atoms of the form $L_0\phi$, which refer to the presence of ϕ in the space T_0. The interpretation of L is constructive as long as the hierarchy is well-founded (no infinite descending chains) and every space contains only modal operators referring to lower spaces.

HAEL is still an *auto*epistemic logic, because the spaces together comprise the agent's belief set. In fact, HAEL could be considered a more natural formalization of autoepistemic reasoning than AE logic, because of its hierarchic structure. In AE logic, we found it necessary to characterize extensions in terms of the groundedness of inferences used in their construction (see [5]), in order to exclude those containing circular reasoning. No such device is necessary for HAEL; circularity in the derivation of beliefs is impossible

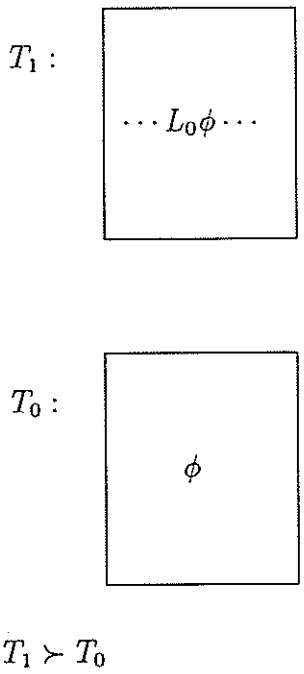

Figure 2: Hierarchic autoepistemic semantics

by the nature of the logic.

Breaking the circularity of AE logic has other advantages. Given a fairly natural class of closure conditions, every HAEL structure has exactly one "extension," or associated theory. So HAEL, although a nonmonotonic logic, preserves many of the desirable properties of first-order logic, including a well-defined notion of *proof* and *theorem*, and a well-founded, compositional semantics. Computationally, HAEL is still not even semi-decidable in the general case; unlike AE logic, however, it lends itself readily to proof-theoretic approximation.

The spaces of HAEL are meant to serve as bodies of evidence, as discussed in Section 2. Spaces lower in the hierarchy are considered to be stronger evidence, and conclusions derived in them take precedence over defaults in spaces higher in the hierarchy. Priorities among defaults and bodies of evidence are readily expressed in the structure of the hierarchy. Many different domains for nonmonotonic reasoning can be fruitfully conceptualized in this fashion. The most natural case is taxonomic hierarchies with

exceptions, because the structure of the spaces mimics the taxonomy (we give an informal encoding of the bat in HAEL example just below). Speech act theory is a very complex and interesting application domain, since the sources of information (agents' mental states, the content and linguistic force of the utterance) interact in complicated ways to induce belief revision after the utterance. In this case, we model the structure of the belief revision process with spaces that reflect the relative force of the new information on old beliefs (see [1]).[1]

To illustrate the main features and use of the evidence hierarchy, we present the default rules about bats and mammals flying before introducing all of the necessary mathematical machinery in the next section. Figure 3 gives the basic HAEL structure for this example. There are three evidence spaces, ordered $\tau_0 \prec \tau_1 \prec \tau_2$. The sentences in τ_0 are stronger evidence than those in the other spaces, and τ_1 is stronger than τ_2. In informal terms, τ_0 is a base level of known facts, τ_1 contains knowledge about bats, and τ_2 about mammals.

The information in the left-hand side of each space is the initial proper axioms that are supplied for the domain. It is known that the individual a is a bat, that bats are mammals, and so on. Note that the default rules are placed in their appropriate evidence spaces, and that they can refer to the contents of spaces underneath themselves in the heirarchy.

On the right-hand side of each space is a list of some sentences that can be concluded in the space. These conclusions come from three sources:

1. The proper axioms of the space.

2. The conclusions of any space below. All sentences about the world which are asserted in a space are automatically inherited by all superior spaces

3. Information about what is contained in other spaces. For example, in τ_1 it is possible to conclude that $\neg Fa$ is not in τ_0.

As can be seen, the correct conclusion that a flies is inferred in τ_1, and is passed up to τ_2, preventing the derivation of a not flying.

[1]It should be noted that this is the first formalization of speech-act theory in a non-monotonic system that attempts to deal with a nontrivial belief-revision process.

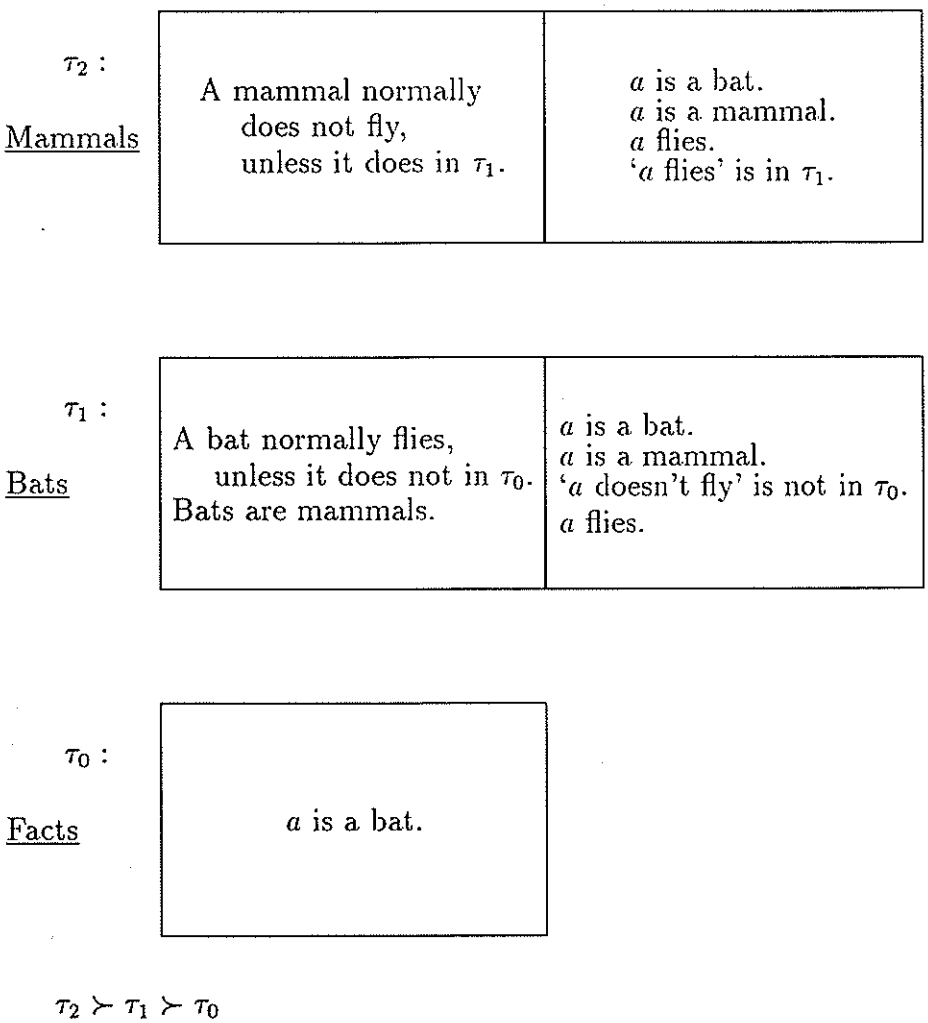

Figure 3: A taxonomic example in HAEL

4 HAEL Structures and their Semantics

We now present the formal definition of HAEL structures and two independent semantics for these structures. The first is based on the notion of a stable set, an idea introduced by Stalnaker [18] and used extensively in the development of AE logic [13; 5]. Stable sets are defined using closure conditions that reflect the end result of introspection of an ideal agent on his own beliefs. The second semantics is a classical approach: first-order valuations modified to account for the intended interpretation of the L_i-operators. This semantics is taken directly from AE logic and shows many of the same properties. However, the hierarchical nature of HAEL structures produces some significant differences. In AE logic, a belief set that follows from a given assumption set A via the semantics is called an *extension* of A. There may be no, one, or many mutually conflicting extensions of A. HAEL structures always have exactly one extension, and thus a well-defined notion of theorem.

There is also a mismatch in AE logic between stable-set semantics and autoepistemic valuations. A stable set for A which is minimal (in an appropriate sense) is a good candidate for a belief set; yet minimal stable sets for A exist that are not extensions of A. In HAEL, we show that the two semantics coincide: the unique minimal stable set of an HAEL structure is the extension of that structure given by its autoepistemic valuations.

4.1 HAEL Structures

In AE logic, one starts with a set of premise sentences A, representing the initial beliefs or knowledge base of an agent. The corresponding object in HAEL is an *HAEL structure*. A structure τ consists of an indexed set of evidence spaces τ_i, together with a well-founded, irreflexive partial order on the set. We write $\tau_i \prec \tau_j$ if τ_i precedes τ_j in the order. The partial order of spaces reflects the relative strength of the conclusions reached in them, with preceding spaces having stronger conclusions. The condition of well-foundedness means that there is no infinite descending chain in the partial order; the hierarchy always bottoms out.

Each space τ_i contains an initial premise set A_i, and also an associated first-order deduction procedure I_i. The deduction procedures are sound (with respect to first-order logic) but need not be complete. The idea behind parameterizing HAEL structures by inference procedures in the spaces is that

ideal reasoning can be represented by complete procedures, while resource-bounded approximations can be represented by incomplete but efficient procedures. In the rest of this paper, we shall assume complete first-order deduction in each space; HAEL structures of this form are called *complete*.

The language $\mathcal{L}$ of HAEL consists of a standard first-order language, augmented by a indexed set of unary modal operators, L_i. If ϕ is any sentence (no free variables) of the first-order language, then $L_i\phi$ is a sentence of $\mathcal{L}$. Note that neither nesting of modal operators nor quantifying into a modal context is allowed. Sentences without modal operators are called *ordinary*. The intended meaning of $L_i\phi$ is that the sentence ϕ is an element of space τ_i.

Within each space, inferences are made from the assumption set, together with information derived from spaces lower in the hierarchy. Because spaces are meant to be downward-looking, the language $\mathcal{L}_i \subseteq \mathcal{L}$ of a space τ_i need contain only modal operators referring to spaces lower in the hierarchy. We formalize this restriction with the following statement:[2]

$$\textit{The operator } L_j \textit{ occurs in } \mathcal{L}_i \textit{ if and only if } \tau_j \prec \tau_i. \tag{8}$$

Here is the bat example from the last section formalized as an HAEL structure.

$$\begin{aligned} &\tau_0 \prec \tau_1 \prec \tau_2 \\ &A_0 = \{B(a)\} \\ &A_1 = \{\forall x.Bx \supset Mx, \\ &\qquad L_0B(a) \wedge \neg L_0\neg F(a) \supset F(a)\} \\ &A_2 = \{L_1M(a) \wedge \neg L_1F(a) \supset \neg F(a)\}\,. \end{aligned} \tag{9}$$

There are three spaces, with a strict order (heritable) between them. Space τ_0 is the lowest in the hierarchy, and contains the most specific information (based on the taxonomy). In the assumption set A_1, there is a default rule about bats flying: if it is known in τ_1 that a is a bat, and unknown in τ_0 that a does not fly, then it will be inferred that a flies. The assumption set A_2 is similar to A_1; it also permits the deduction that bats are mammals. The

[2]We can relax this restriction so that L_i can occur in $\mathcal{L}_i$ under certain circumstances. Because it is simpler to present the semantics without this complication, we will not defer considering it until Section 5.

information that a is a bat and a mammal will be passed up to τ_2, along with any inferences about its ability to fly.

Because the partial order of an HAEL structure is well-founded, we can perform inductive proofs using it. At times we will need to refer to unions of sets derived from the spaces preceding some space τ_n; to do this, we use $\bigcup_{j \prec n} X_j$, where j ranges over all indices for which $\tau_j \prec \tau_n$.

4.2 Complex Stable Sets

Stalnaker considered a belief set Γ that satisfied the following three conditions:

1. Γ is closed under first-order consequence.[3]
2. If $\phi \in \Gamma$, then $L\phi \in \Gamma$.
3. If $\phi \notin \Gamma$, then $\neg L\phi \in \Gamma$.

He called such a set *stable*, because an agent holding such a belief set could not justifiably deduce any further consequences of his beliefs. In HAEL, these conditions must be modified to reflect the nature of the L_i-operators, as well as the inheritance of sentences among spaces.

DEFINITION 4.1 *A* complex stable set *for a structure τ is a sequence of sets of sentences $\Gamma_0, \Gamma_1, \ldots$, corresponding to the spaces of τ, that satisfies the following five conditions:*

1. *Every Γ_i contains the assumption set A_i.*
2. *Every Γ_i is closed under the inference rules of τ_i. In the case of an ideal agent, the closure is first-order logical consequence.*
3. *If ϕ is an ordinary sentence of Γ_j, and $\tau_j \prec \tau_i$, then ϕ is in Γ_i.*
4. *If $\phi \in \Gamma_j$, and $\tau_j \prec \tau_i$, then $L_j\phi \in \Gamma_i$.*
5. *If $\phi \notin \Gamma_j$, and $\tau_j \prec \tau_i$, then $\neg L_j\phi \in \Gamma_i$.*

[3]Stalnaker considered propositional languages and so used tautological consequence.

To illustrate complex stable sets, consider the previous example of flying bats. Let $\mathrm{Cn}_i[X]$ stand for the first-order closure of X using language $\mathcal{L}_i$, and define the set $S = S_0, S_1, \cdots$ by

$$\begin{aligned} S_0 &= \mathrm{Cn}_0[B(a)] \\ S_1 &= \mathrm{Cn}_1[B(a), L_0B(a), \neg L_0\neg B(a), \neg L_0\neg F(a), \\ &\qquad \forall x.Bx \supset Mx,\ M(a),\ F(a), \cdots] \\ S_2 &= \mathrm{Cn}_2[B(a), M(a), F(a), L_0B(a), L_1B(a), \\ &\qquad L_1F(a), \cdots]\,. \end{aligned} \tag{10}$$

The set S is a complex stable set for the HAEL structure τ as defined in Equations (9). The lowest set S_0 contains just the first-order consequences of $B(a)$. S_1 inherits this sentence, and has the additional information $M(a)$ from its assumption set. Modal atoms of the form $L_0\phi$ and $\neg L_0\phi$ are also present, reflecting the presence or absence of sentences in S_0; the sentence $F(a)$ is derived from these plus the assumption set. Finally, S_2 inherits all ordinary sentences from S_1, as well as $L_1F(a)$.

The subsets S_i of S are *minimal* in the sense that we included no more than we were forced to by the conditions on complex stable sets. For example, another stable set S' might have $S'_0 = \mathrm{Cn}_0[B(a), \neg F(a)]$, with the other spaces defined accordingly. The sentence $\neg F(a)$ in S'_0 is not justified by the original assumption set A_0, but there is nothing in the definition of complex stable sets that forbids it from being there. So, a complex stable set is a candidate for the extension of an HAEL structure only if it is minimal. But what is the appropriate notion of minimality here? For simple stable sets, minimality can be defined in terms of set inclusion of the ordinary sentences of the stable sets. Complex stable sets have multiple spaces, and the definition of minimality must take into account the relative strength of information in these spaces.

DEFINITION 4.2 *A stable set S for the HAEL structure τ is* minimal *if for each subset S_i of S, there is no stable set S' for τ that agrees with S on all $\tau_j \prec \tau_i$, and for which $S'_i \subset S_i$.*

A complex stable set for τ is minimal if each of its subsets is minimal, given that the preceding subsets (those of higher priority) are considered fixed.

There is exactly one minimal complex stable set for an HAEL structure. We now prove this fact, and give an inductive definition of the set.

PROPOSITION 4.1 *Every HAEL structure τ has a unique minimal complex stable set, which can be determined by the following inductive procedure:*

Define

$$\begin{array}{lcl} \mathrm{Cn}_i[X] & = & \textit{the first-order closure of } X \textit{ in } \mathcal{L}_i \\ \mathrm{Ord}(X) & = & \textit{the ordinary sentences of } X \\ \mathrm{L}_i(X) & = & \{L_i\phi \,|\, \phi \in X \textit{ and } \phi \textit{ ordinary}\} \\ \mathrm{M}_i(X) & = & \{\neg L_i\phi \,|\, \phi \notin X \textit{ and } \phi \textit{ ordinary}\} \end{array}$$

For leaf τ_i (that is, there is no τ_j such that $\tau_j \prec \tau_i$), let

$$S_i = \mathrm{Cn}_i[A_i] \,.$$

For nonleaf τ_n, define

$$S_n = \mathrm{Cn}_n[A_n \cup \bigcup_{\tau_j \prec \tau_n} \mathrm{Ord}(S_j) \cup \mathrm{L}_j(S_j) \cup \mathrm{M}_j(S_j)] \,.$$

S is the unique minimal complex stable set for τ.

Proof. All of the conditions of Definition 4.1 that S is a complex stable set. To show that S is minimal, we show that the conditions of Definition 4.2 are satisfied for S.

Assume that S' is a stable set for τ with a subset S'_i differing from S_i, such that S' agrees with S on all $\tau_j \prec \tau_i$. Because S' is a complex stable set, it must contain A_i, and for each $\tau_j \prec \tau_i$, it must also contain $\mathrm{Ord}(S_j)$, $\mathrm{L}_j(S_j)$, and $\mathrm{M}_j(S_j)$. It is also closed under first-order consequence. By the definition of S_i, we must therefore have $S_i \subseteq S'_i$, and since these two subsets differ, $S_i \subset S'_i$. Hence S must be minimal. Further, S' cannot itself be a minimal stable set, because this last subset relation violates the conditions of minimality for S'. Because S' was chosen arbitrarily, there can be no other minimal stable sets than S.

The existence of a unique minimal complex stable set for every HAEL structure gives us a means of defining the theorems of a structure. Let S be the complex stable set for τ. We say that a sentence ϕ is *derivable in the space* τ_i if and only if it is an element of S_i, and write $\tau \vdash_i \phi$ if this holds. For the bat example, the following derivations exist, where τ is the HAEL structure of 9):

$$\begin{aligned}
&\tau \vdash_0 B(a) \\
&\tau \not\vdash_0 \neg F(a) \\
&\tau \vdash_1 B(a) \wedge M(a) \wedge \neg L_0 \neg F(a) \wedge F(a) \\
&\tau \vdash_2 B(a) \wedge M(a) \wedge L_1 F(a) \wedge F(a) \; .
\end{aligned} \tag{11}$$

4.3 HAEL Semantics

We have used complex stable sets to give a proof-theoretic notion of theorem to HAEL structures. An alternative approach is to develop a semantics for these structures and define a notion of validity with respect to the semantics. As with autoepistemic logic, the semantic picture is complicated by the presence of self-referential elements, and validity must be determined by use of a fixedpoint equation. Happily, validity turns out to be equivalent to derivability for HAEL structures, so that the sentences that are valid logical consequences of a structure are exactly those given by its minimal complex stable set.

We start with the notion of a *valuation* of an HAEL structure τ. In classical logic, a valuation assigns **true** or **false** to each sentence of the language, and a valuation is said to *satisfy* a theory if all the sentences of the theory are assigned **true**. If the valuation v assigns **true** to the sentence ϕ, we write $v \models \phi$. Restrictions on valuations single out the intended semantics of the theory, e.g., first-order valuations must respect the intended meaning of the quantifiers and boolean operators.

In autoepistemic logic, the interpretation of the modal operator L adds an additional complication to valuations. Since the intended interpretation of $L\phi$ is that ϕ be in the belief set of the agent, an AE valuation consists of a first-order valuation, v, and a set of sentences (the belief set), Γ (see [13]). We call Γ the *modal index* of the valuation. The interpretation rules for AE valuations are as follows (we let ϕ stand for an arbitrary ordinary sentence):

$$\begin{array}{ll} \langle v, \Gamma \rangle \models \phi & \text{iff} \quad v \models \phi \\ \langle v, \Gamma \rangle \models L\phi & \text{iff} \quad \phi \in \Gamma \, . \end{array} \tag{12}$$

The interpretation of the L-operator is completely decoupled from the first-order valuation.

The *autoepistemic extension* of an assumption set A is a set of sentences T that are the logical consequences of A under AE valuations. Because the intended interpretation of L is *self*-belief, only those AE valuations that respect this interpretation can be used. Let $A \models_\Gamma \phi$ mean that every AE valuation with modal index Γ that satisfies the set A also satisfies ϕ. An extension T of A is defined by the following equation (see [5]):

$$T = \{\phi \mid A \models_T \phi\} \, . \tag{13}$$

By fixing the modal index as T, we are assured that the interpretation of L is with respect to the belief set T itself. Of course, the equation defining extensions is self-referential and, as we have pointed out, self-reference creates problems from a computational point of view.

The semantics of HAEL structures is similar to AE assumption sets, but is complicated by the presence of multiple spaces. The interpretation of the indexed operators L_i must be with respect to a sequence of belief subsets, instead of a single belief set Γ. So an HAEL valuation $\langle v, \Gamma_1, \cdots, \Gamma_n, \cdots \rangle$ consists of a first-order valuation v, together with the indexed belief subsets Γ_i, which we call a *complex belief set*. The interpretation rules for HAEL valuations are similar to those for AE valuations (again, ϕ stands for an arbitrary ordinary sentence).

$$\begin{array}{ll} \langle v, \Gamma_1, \cdots, \Gamma_n, \cdots \rangle \models \phi & \text{iff} \quad v \models \phi \\ \langle v, \Gamma_1, \cdots, \Gamma_n, \cdots \rangle \models L_i\phi & \text{iff} \quad \phi \in \Gamma_i \end{array} \tag{14}$$

The interpretation of each L_i is with respect to the appropriate belief subset. Note that there is no necessary relation in valuations among the interpretations of the modal operators, or between the modal operators and the first-order valuation.

An autoepistemic extension of an HAEL structure, τ, consists of a sequence of a complex belief set, $T = T_1, \cdots, T_n, \cdots$, corresponding to the spaces of the structure. Again, we require that extensions be defined using only those valuations that respect the nature of the L_i-operators as *self*-belief. Also, because each space inherits the ordinary sentences of preceding subsets, the assumption set must be augmented appropriately.

DEFINITION 4.3 *The complex belief set T is an* extension *of τ if it satisfies the equations*

$$T_i = \{\phi \in \mathcal{L}_i \mid A_i \cup \bigcup_{\tau_j \prec \tau_i} \mathrm{Ord}(T_j) \models_T \phi\} .$$

As with AE logic, the definition of extensions for HAEL appears to be self-referential, since T_i appears on both sides of the equation. However, this self-reference is illusory from the point of view of the individual spaces, because they contain L_i-operators referring only to spaces lower in the hierarchy. In fact, every HAEL structure has a unique extension, and that extension is the minimal complex stable set.

PROPOSITION 4.2 *Every HAEL structure τ has a unique extension T, which is the complex stable set for τ.*

> *Proof.* By induction over the structure of τ, we can show that there is a unique solution to these equations, and that this yields the minimal stable set. For the base case, let τ_i be a leaf space. The language $\mathcal{L}_i$ has no modal operators, and T_i must be $\mathrm{Cn}_i[A_i]$. For nonleaf τ_i, assume that all T_j with $\tau_j \prec \tau_i$ are uniquely defined and equal to the corresponding S_j of the minimal stable set. Then it is easy to show from the definitions that S_i and T_i must coincide.

Having a single extension is a nice feature of HAEL structures, because there is a single notion of *theorem*, and the problem of choosing among competing multiple extensions (as in AE logic) does not exist. However, there is a price to pay. In AE logic, multiple extensions arise because there are conflicting defaults: the classic Nixon diamond is a well-known example, where the default that Republicans are not pacifists conflicts with the default that Quakers are. In HAEL, if both these defaults are placed in the same space, an inconsistency will occur (there will still be a single extension, but the space will consist of all sentences because of closure under logical consequence). Thus the HAEL structure must be constructed so that conflicts of this sort within the same space are avoided.

4.4 Proof Theory

Proposition 4.1 is important in that it makes the notion of theorem well-founded for HAEL structures. It also is the basis for proof methods on HAEL structures. Consider the previous example of the bat taxonomy [Equation (9)]. We want to know whether a flies, that is, whether $F(a)$ or $\neg F(a)$ is provable in T_2. Suppose we set $\neg F(a)$ as a goal in T_2. There is only one axiom which applies, and this gives the subgoal $L_1 M(a) \wedge \neg L_1 F(a)$. To establish the first conjunct, we set up $M(a)$ as a goal in T_1. Using the universal implication, we arrive at the subgoal $B(a)$, which matches with $B(a)$ in T_0. Hence we have shown that $L_1 M(a)$ holds in T_2.

In a similar manner, we set up $F(a)$ as a subgoal in T_1. Using the second axiom of A_1, we have the conjunctive goal $L_0 B(a) \wedge \neg L_0 \neg F(a)$. The first subgoal is easily proven, since $B(a)$ is in T_0. Now we try to prove $\neg F(a)$ in T_0. This is not possible, so $\neg L_0 \neg F(a)$ is proven in T_1. We have just shown $F(a)$ to be provable in T_1, so $\neg L_1 F(a)$ is not provable in T_2. Our attempt to prove $\neg F(a)$ in T_2 fails. On the other hand, along the way we have shown $F(a)$ to be provable in T_1; hence by inheritance it is also in T_2.

In this example, we used backward chaining exclusively as a proof method. Other methods are also possible, e.g., mixtures of forward and backward chaining. Whenever there is a question as to the provability of a modal atom, an appropriate subgoal is set up in a preceding space, and the proof process continues.

It should be noted that no proof process can be complete when the non-modal language is undecidable, because the inference of $\neg L_i \phi$ requires that we establish ϕ to be not provable in T_i. However, a proof method can readily approximate the construction of Proposition 4.2 by assuming that ϕ cannot be proven after expending a finite amount of effort in attempting to prove it. Given enough resources, a proof procedure of this sort will converge on the right answer.

We have implemented HAEL on a resolution theorem-proving system, modified to accept a belief logic of the sort described in Geissler and Konolige [3]. The implementation was straightforward, and involved adding a simple negation-as-failure component to the prover. The implementation has been successfully applied to reasoning about speech acts in a natural-language understanding project [1].

5 Extensions and Limitations

5.1 Extensions

There are several ways in which we can extend the utility of HAEL. The first of these is to relax the restriction on the modal operator appearing in the same space as its index. We note that it is only reasoning about the *non*-provability of a sentence in its own space which is problematical. Hence we may allow reasoning within a space about what is provable. We modify the restriction on the language of spaces (Equation 8) to be:

$$\textit{If } L_j \textit{ occurs in } \tau_i, \textit{ then } \tau_j \preceq \tau_i. \tag{15}$$

We also restrict the occurrence of L_j in the assumption set:

$$\textit{If } L_j \textit{ occurs positively (negatively) in } A_i, \textit{ then } \tau_j \prec \tau_i \ (\tau_j \preceq \tau_i). \tag{16}$$

The distinction between positive and negative occurrences is important when the operator L_j is in its own theory τ_j. If L_j occurs *positively* in the assumption set A_j, it could be used to make inferences based on what is not in the space. For example, in the sentence $\neg L_j P \supset Q$, L_j occurs positively, and the intended meaning of this sentence (when in the assumption set A_j) is that Q is in τ_j if P is not in τ_j. If Q itself implies P through some chain of inference, we get just the kind of self-referential reasoning we are trying to avoid.

On the other hand, a negative occurrence of L_j in A_j is not problematic, as long as we are careful about grounding all inferences. With the sentence $L_j P \supset Q$ in A_j, for example, we have a statement that the presence of P in τ_j allows the inference of Q. If P could be inferred from Q, then we would have a case of circular justifications, but only if we are allowed to assume P or $L_j P$ in the first place. As long as P must be inferred independently from the assumption set and information lower in the hierarchy, there is no problem of circularity.

A second extension is to specify inheritance of ordinary sentences only for a subset of the partial order. A subset of the partial order is distinguished as being *heritable*, and we write these as $\tau_i \prec_h \tau_j$. Heritable precedence is used in cases such as the taxonomic example, where all of the facts of the lower space τ_i are also meant to be facts inherited by the upper space τ_j. Nonheritable precedence is more appropriate when the information in the

spaces refers to different situations or incompatible views of the world, as we might do in an axiomatization of the situation calculus. In any particular structure, heritable and nonheritable relations could both be necessary, which is why inheritance $\prec_h$ is a subrelation of the partial order $\prec$.

The presence of nonheritable precedence causes a minor change in the definition of HAEL stable sets and extensions, and also Proposition 4.1. The changes are obvious: we just specify inheritance only for spaces which are in the heritable hierarchy.

5.2 Limitations

The encouraging result of having a single extension of HAEL structures is not without a penalty. In the face of conflicting defaults which are not prioritized, HAEL will yield an inconsistent extension, rather than two mutually incompatible extensions. The simplest example comes from representing the Nixon diamond. Suppose we encode the two defaults in the same space:

$$\begin{aligned} &\tau_0 \prec \tau_1 \\ &A_0 = \{Rn, Qn\} \\ &A_1 = \{L_0 Rx \wedge \neg L_0 \neg Hx \supset Hx\,, \\ &\qquad L_0 Qx \wedge \neg L_0 \neg Px \supset Px\} \\ &\qquad \forall x.Hx \supset \neg Px \end{aligned} \tag{17}$$

Both defaults will apply, since τ_0 satisfies their premisses. Since both Pn and $\neg Pn$ are derivable in τ_1, the stable set is inconsistent in this space.

There are two ways to remedy this problem: always prioritize conflicting defaults, or ameliorate the effects of contradicting defaults. The first choice is unpalatable, for the same reason we criticized AE logic at the beginning of the paper: we may not know when defaults conflict, and we shouldn't be forced to prioritize them if we don't know which should take precedence.

The second choice is more interesting, since we already have some notion of strength of evidence. If the evidence for a proposition is conflicting, then we should just ignore it, rather than forming a contradiction. This strategy would involve changing the truthvalue semantics of the logic, and we are starting to explore it.

Another problematic feature of the logic is that the structure of spaces must be given in a fixed form for any particular application. It would be nice

to specify the structure in a more flexible way, perhaps having conditional relations among the spaces. Such flexibility seems to require some sort of metalevel capability for HAEL.

References

[1] D. E. Appelt and K. Konolige, A nonmonotonic logic for reasoning about speech acts and belief revision, *Second Workshop on Non-Monotonic Reasoning* (1988).

[2] D. W. Etherington and R. Reiter, On inheritance hierarchies with exceptions, in: *Proceedings of the American Association of Artificial Intelligence* (1983).

[3] C. Geissler and K. Konolige, A resolution method for quantified modal logics of knowledge and belief, in: J. Y. Halpern, ed., *Conference on Theoretical Aspects of Reasoning about Knowledge* (Morgan Kaufmann, 1986) 309–324.

[4] S. Hanks and D. McDermott, Nonmonotonic logic and temporal projection, *Artificial Intelligence* **33** (3) (1987).

[5] K. Konolige, On the relation between default logic and autoepistemic theories, *Artificial Intelligence* **35** (3) (1988) 343–382.

[6] K. Konolige and K. Myers, Representing defaults with epistemic concepts, *Computational Intelligence* **5** (1989) 32–44.

[7] H. J. Levesque, A formal treatment of incomplete knowledge bases, Technical Report 614, Fairchild Artificial Intelligence Laboratory, Palo Alto, California (1982).

[8] H. J. Levesque, All I know: an abridged report, in: *Proceedings of the American Association of Artificial Intelligence*, Seattle, Washington (1987).

[9] V. Lifschitz, Some results on circumscription, in: *AAAI Workshop on Non-Monotonic Reasoning*, Menlo Park, California (1984).

[10] R. P. Loui, Defeat among arguments: A system of defeasible inference, *Computational Intelligence* **3** (2) (1987).

[11] J. McCarthy, Circumscription — a form of nonmonotonic reasoning, *Artificial Intelligence* **13** (1–2) (1980).

[12] R. C. Moore, Possible-world semantics for autoepistemic logic, Technical Note 337, SRI Artificial Intelligence Center, Menlo Park, California (1984).

[13] R. C. Moore, Semantical considerations on nonmonotonic logic, *Artificial Intelligence* **25** (1) (1985).

[14] D. Poole, On the comparison of theories: preferring the most specific explanation, in: *Proceedings of the International Joint Conference on Artificial Intelligence*, Los Angeles (1985) 144–147.

[15] R. Reiter, A logic for default reasoning, *Artificial Intelligence* **13** (1–2) (1980).

[16] R. Reiter and G. Criscuolo, Some representational issues in default reasoning, in: N. J. Cercone, ed., *Computational Linguistics* (Pergamon Press, Elmsford, New York, 1983) 15–27.

[17] Y. Shoham, *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence* (MIT Press, Cambridge, Massachusetss, 1987).

[18] R. C. Stalnaker, A note on nonmonotonic modal logic, Department of Philosophy, Cornell University, (1980).