STRESZCZENIE

Artykuł stanowi dogłębne studium semantyki logiki autoepistemicznej wykorzystujące wyniki wieloletnich badań autora w tym przedmiocie. Rozpoczyna się ono od skrótowego przeglądu semantyk powszechnie stosowanych wzorców niemonotonicznej dedukcji biorącą się z braku wiedzy, włączając w to dedukcję autoepistemiczną, w terminach równania stałopunktowego \( \Phi(T,E) = E \). Następnie bada ono wąsko semantykę minimalnych ekspansji dla zdaniowej logiki autoepistemicznej oraz jej operację \( Cn_{AE} \) odpowiadającą schematowi wnioskowania opartemu na założeniu “wiedząc tylko”. W szczególności następujące założenie \( \text{MKA} \) o minimalności wiedzy:

\[
\varphi \in \text{MKA}(T) \text{ wtedy, i tylko wtedy, gdy } \varphi \text{ nie dokłada modalnie pozytywnych } S5\text{-konsekwencji do } T
\]

jest używane w celu syntaktycznego charakteryzowania operacji \( Cn_{AE} \) przy pomocy stosownego twierdzenia o pełności.

Artykuł oferuje również dowód, że operacja konsekwencji \( Cn_{S5} \) logiki modalnej \( S5 \) jest maksymalną monotonną operacją konsekwencji spełniającą \( Cn_{S5}(T) \subseteq \text{MKA}(T) \) dla każdej teorii modalnej \( T \).

ABSTRACT

This paper comprises an in-depth study of semantics of autoepistemic logic that is based on author’s may years of research in the subject matter. It begins with a brief review of semantics of common patterns of nonmonotonic deduction arising from a lack of knowledge, including autoepistemic deduction, in terms of the fixed-point equation \( \Phi(T,E) = E \). Then it narrows the investigation of minimal expansion semantics for autoepistemic propositional logic and its “only knowing” consequence operation \( Cn_{AE} \). In particular, the following minimal-knowledge assumption \( \text{MKA} \):

\[
\varphi \in \text{MKA}(T) \text{ iff } \varphi \text{ does not add modally positive } S5\text{-consequences to } T
\]

is used to syntactically characterize the operation \( Cn_{AE} \) by means of suitable completeness theorem.

The paper also offers a proof that the consequence operation \( Cn_{S5} \) of modal logic \( S5 \) is the maximal monotonic consequence operation satisfying \( Cn_{S5}(T) \subseteq \text{MKA}(T) \) for every modal theory \( T \).

KEY WORDS: autoepistemic logic, non-monotonic logic, modal logic, Kripke models, reasoning from a lack of knowledge, “only knowing” semantics, minimal-knowledge assumption.

AMS CLASSIFICATION: 68G99, 03C40, 03C52.
INTRODUCTION

Formal methods of knowledge representation are a topic of central interest in theoretical research in Artificial Intelligence. Adequate assessment of the information contained in knowledge bases, and - in particular - deriving secondary knowledge from a lack of certain information or knowledge, is one of the most challenging problems in this area. One of the concepts that has been used as a theoretic framework for such assessment is introspection. It allows one to formally express what does the knowledge base in question know and what it does not.

This paper addresses certain logical aspects related to introspection in the said context, providing characterization of nonmonotonic deduction based on such introspection.

The paper is organized as follows.

Section 1 provides a brief review of major patterns of nonmonotonic inferences that arise from a lack of knowledge.

Section 2 introduces the language and semantics of autoepistemic logic.

Sections 3, 4, and 5 constitute a brief review of fundamental concepts involved in construction of autoepistemic logic, and their basic properties known from the literature. Section 3 introduces the concept of stable theory, and relates the consequence operation induced by that concept to modal system S5. Section 4 introduces universal Kripke structures and proves that they are completely axiomatized by S5. Section 5 defines the autoepistemic semantics based on the concept of minimal expansion, and proves that this semantics is equivalent to the semantics of maximal universal Kripke models.

Sections 6 and 7 consist of technicalities needed to prove the completeness theorem for autoepistemic logic. Section 6 shows the S5-reducibility of autoepistemic theories to certain normal and clausal forms. Section 7 translates two classic results from first-order model theory into the language of autoepistemic logic: a version of Lyndon homomorphism theorem and Bossu-Siegel minimal modelability lemma.

Section 8 defines the minimal knowledge assumption (MKA), proves that MKA completely characterizes $Cn_{AE}$, and discusses certain consequences of that fact. The characterization of $Cn_{AE}$ is used to assess limitations of the proving power of $AE$ logic. Although, indeed, $Cn_{AE}$ cannot prove any new modally positive sentences from $T$, which in my opinion is exactly what it should not do, a stratified application of MKA can. I will comment on that in Section 9.

Section 10 briefly relates the presented approach to relevant results of others, and points out certain undesirable properties of some of them.

Section 11 mentions some other major schemes of nonmonotonic deduction.
1. A BRIEF REVIEW OF NONMONOTONIC DEDUCTION

Interest in nonmonotonic reasoning originated from a need to formalize methods of drawing conclusions from the absence of information or a lack of knowledge. To that end, in addition to axioms and (monotonic) rules of inference, one may adopt rules of the form

\[ \not \vdash \varphi \mid \cdots \vdash \psi \]

(meaning: if \( \varphi \) is not provable\(^2\), and \( \cdots \), then infer \( \psi \)). Rules of this kind are nonmonotonic in that, unlike in classical logic, they can prove more from less premises. However, because in systems that admit nonmonotonic rules of inference the notion of provability is potentially dependent on non-provability, defining a deductive closure for such systems that would avoid falling into a pitfall of circular definition was considerably more involved than the similar task in monotonic logic. This has usually been accomplished by having the deductive closure implicitly defined by the fixed-point equation

\[ \Phi(T, E) = E \]  

(1)

that is expressed in terms of consequence operator \( \Phi \) applied to two sets of statements:

- \( T \) - a set of premises, the subject of deductive closure, and
- \( E \) - a set of assumptions for verification of provability/non-provability assertions, e.g., for verification of \( \not \vdash \varphi \).

The operator \( \Phi \) is monotonic with respect to \( T \) (if \( T \subseteq T' \) then \( \Phi(T, E) \subseteq \Phi(T', E) \)), which feature allows one to avoid circularity, but not necessarily monotonic with respect to \( E \).

Set \( E \) is called a deductive closure of \( T \) if, and only if, in addition to containing classical (e.g., propositional, modal, etc.) consequences of \( T \) and, perhaps, satisfying some other constrains, it is a solution of the fixed-point equation (1). Although such a set obeys all the nonmonotonic rules of inference of the system, not necessarily every set closed under those rules will form a solution of the equation. In particular, all the elements of a solution set \( E \) are supported in that they are either monotonically derivable from \( T \) or follow from \( E \) by means of the prescribed (usually, nonmonotonic) rules of inference. (Note that deriving \( \varphi \) from \( \varphi \), an obviously valid inference, is generally considered not nonmonotonic.) Hence, the notion of supportedness may be viewed as a generalization of provability.

As a result, the fixed-point style definition of the deductive closure does not allow for inferences that would actually remedy the very lack of information the nonmonotonic rules were introduced to deal with in the first place. For example, intuitively straightforward nonmonotonic rule

\[ \not \vdash \varphi \mid \not \vdash \neg \varphi \]

(2)

(meaning: if \( \varphi \) is independent from what has been proved so far then infer \( \varphi \), by default) cannot produce from \( T = 0 \) any solutions \( E \) of the fixed-point equation (although the set \( Cn(\varphi) \) is closed under both rule (2) and propositional consequence) because if \( \varphi \in E \) then there is no support for \( \varphi \) in \( E \), at least as far as (2) is concerned\(^3\); otherwise (2) is obviously violated.

To its all natural appeal (for a logician), such a definition introduces two major complications. Firstly, the fixed-point equation (1), as opposed to classical definition of deductive closure via the concept of proof, may provide no clue how to compute its solution. Secondly, the equation may have more than one solution, which - as a matter of fact - increases the expressive power of the logic in question, or no solutions at all, which creates obvious problems of ontological nature. Moreover, being an implicit definition, it may render the explicit meaning of the deductive properties of the system, as well as the actual semantics of its language, unclear.

Autoepistemic logic, introduced by Moore in [Moo85] incorporates axioms and rules of inference expressed in a language \( L_K \) of modal logic \( S5 \) (whose axioms and rules of inference are listed in Section 3). In Moore’s formalization (according to its characterization in [Suc00b]), autoepistemic logic involves two sets: \( T \) (the subject of closure) and \( E \) (the context), and fixed-point equation (1). In addition to axioms and rules of \( S5 \), it admits the following nonmonotonic rule of inference, which

\(^2\)Adjective “provable” and phrase “known to be true” are synonymous in this context.

\(^3\)In this case, \( \varphi \not\in \Phi(0, E) \) because \( \vdash \varphi \), so (2) is not applicable and \( \varphi \) cannot be nonmonotonically derived from \( E \).
I refer to as anti-necessitation:

\[ \phi \not\vdash K\phi \]  

(some authors use symbol \(\Box\) instead of \(K\)) whose meaning is formally described by:

\[ \phi \not\in E \quad \neg K\phi \in \Phi(T, E) \]  

The necessitation rule of \(S5\)

\[ \phi \quad \vdash K\phi \]  

is interpreted similarly to anti-necessitation, that is by

\[ \phi \in E \quad K\phi \in \Phi(T, E) \]  

The operator \(\Phi(T, E)\) of deductive closure of \(T\) under necessitation (5) and anti-necessitation (3) relative to \(E\) is defined as the closure of \(T\) under (4), (6), and propositional consequence. Set \(E\) is called an expansion (a nonmonotonic deductive closure, that is) of \(T\) iff it satisfies the fixed-point equation (1). As a result, \(E\) is supported and stable, the latter attribute meaning that it satisfies the so-called stability conditions:

\[ \phi \in E \text{ iff } K\phi \in E \text{ iff } \neg K\phi \not\in E. \]  

**Example 1.1.** Every theory \(T\) without occurrences of modal operator \(K\) has exactly one expansion: the only set \(E(T)\) closed under the rule

\[ \phi \not\in E(T) \quad \vdash \neg K\phi \in E(T) \]  

and \(S5\)-consequence whose modal-free part \(E_0(T)\) coincides with \(T\). Theory \(T = \{Kp\}\), paradoxically, has no expansions, while theory \(T = \{\neg Kp \supset q, \neg Kq \supset p\}\) has two expansions: \(E(\{p\})\) and \(E(\{q\})\).

Because \(S5\) may be characterized by:

\[ T \vdash_{S5} \phi \text{ iff } \phi \text{ belongs to every stable theory } E \text{ that contains } T, \]

Moore’s autoepistemic logic is a strengthening of \(S5\). In particular, the stability conditions (7) fix the meaning of operator \(K\) of knowledge as nonmonotonic and somewhat self-referential provability, so that in the semantics of \(L_K\) defined by expansions, “\(K\phi\)” means “\(\phi\) is supported”.

An alternative approach to formalization of nonmonotonic logic that is free of the paradoxical properties indicated above begins with restricting semantics of its language \(L\) to selected models for \(L\) by means of certain context-sensitive assumption, as, e.g., minimality assumption. In this approach, a partial order \(<\) is superimposed on the class of semantical structures for \(L\). It allows for formulation of the \(<\)-minimality assumption: a structure \(M\) is a \(<\)-minimal model of \(T\) iff it is a model of \(T\) and for no model \(N\) of \(T\), \(M < N\) holds. This, in turn, yields a definition of \(<\)-minimal entailment \(\vdash_{<}\):

\[ T \vdash_{<} \phi \text{ iff } \phi \text{ is true in every } <\text{-minimal model of } T. \]

Examples of \(<\) in the class of first-order structures contain those defined by proper inclusions \(\subset\) between models’ respective relations, and by proper inclusions \(\subseteq\) between models’ universes, as well as their variants and combinations.

In this paper, I will pursue that latter approach.

Other general schemes of nonmonotonic deduction are briefly discussed in Section 11.

### 2. AUTOEPISTEMIC LOGIC

Since its introduction by Moore in [Moo85], autoepistemic logic has found numerous applications in the areas of research which deal with formally represented knowledge. They include: the theory of knowledge, automated reasoning, and logic programming. Autoepistemic logic is based upon the concept of introspection, that is, the awareness of one’s own knowledge and ignorance to the extent which would allow for self-assessment of what one actually knows and what one doesn’t. The language \(L_K\) of autoepistemic logic consists of a propositional language \(L\), and, in addition to constants
\( \top \) (true), \( \bot \) (false), propositional variables \( p_1, p_2, p_3, \ldots \), logical connectives \( \lor, \land, \neg, \supset, \) etc., contains a modal operator \( K \) for expressing facts that are known to the knower in question. Because I am mainly concerned with the modal properties of \( L_K \), I will follow the general trend and do not dwell on the nature of propositional variables of \( L_K \), e.g. whether or not they possess any internal structure, or how they are possibly related to one another, although since Henkin discovered that the predicate calculus is reducible to propositional logic, this might be a very interesting topic to pursue.

Having stated the above as my point of departure, I identify a knowledge base with a set \( T \) of sentences of \( L_K \), and yes-or-no queries to \( T \) with sentences of \( L_K \). The knowledge base \( T \) constitutes a formal description of one’s knowledge. Its intensional meaning is “\( T \) is all that is known” (on a certain subject or a group of subjects), which I will refer to as the “only knowing” interpretation.

Modal-free sentences of \( T \), the members of \( L \), represent the objective knowledge, the so called facts. Operator \( K \) allows for expressing the self-assessing statements or queries, e.g., \( K(p_1 \lor p_2) \) has the meaning of “it is known that \( p_1 \) or \( p_2 \)”, \( Kp_1 \lor Kp_2 \) means “it is known that \( p_1 \) or it is known that \( p_2 \)”, and \( \neg Kp_1 \) is interpreted as “it is not known that \( p_1 \)”.

Given a knowledge base \( T \), the central question which remains to be answered is this:

“What is the proper answer to a yes-or-no query \( \varphi \) addressed to \( T \)?”

In this paper I reduce it to:

“Is \( \varphi \) an autoepistemic consequence of \( T \)?”

In light of such a settlement, the autoepistemic consequence operation \( Cn_{AE} \), which assigns to each \( T \) the set \( Cn_{AE}(T) \) of all autoepistemic consequences of \( T \), becomes a key element in a system of knowledge representation. Its characterization by purely proof-theoretic means in the usual form of a completeness theorem is the main goal of my investigation.

The nonmonotonicity of autoepistemic logic “in which - as McDermott and Doyle put it in [MD80] - the introduction of new axioms can invalidate old theorems” makes the said goal a challenging problem. In particular, a natural candidate for characterization of \( Cn_{AE} \), Kripke’s modal system S5, cannot be fully adequate because it is monotonic. Indeed, \( \neg Kp_2 \) should belong to autoepistemic consequences of \( \{p_1\} \) (since one doesn’t know \( p_2 \) if \( p_1 \) is all one knows), although \( \{p_1\} \) does not prove \( \neg Kp_2 \) within modal system S5.

Moore’s original attempt of defining autoepistemic consequences of a set of sentences in terms of the intersection of its so called expansions yielded a nonmonotonic operation \( Cn_{Exp} \), which captured properly certain cases of \( T \) (e.g. the modal-free case), but revealed a pathological behavior in some other cases. For example, autoepistemic theory \( \{Kp_1 \lor Kp_2\} \) has no expansions in the sense of [Moo85]. Therefore, depending on interpretation of this fact, \( \{Kp_1 \lor Kp_2\} \) proves everything (including the false sentence \( \bot \)) or nothing (not even \( Kp_1 \lor Kp_2 \) itself). This counterexample shows that \( Cn_{Exp} \) can hardly be recognized as admissible (“the”, if you will) autoepistemic consequence operation because it does not preserve consistency.

The system I present here constitutes a nonmonotonic strengthening of logic S5. Unlike other attempts to formalize autoepistemic reasoning, this one does not depend on any fixed-point equation and is based solely on the concepts of minimal expansion and maximal Kripke structure introduced in [HM85], instead. These concepts allow for adequate and free of paradoxes and unnecessary convolution semantic definition of \( Cn_{AE} \): \( \varphi \) is an autoepistemic consequence of \( T \) iff \( \varphi \) is true in all maximal Kripke models of \( T \), or equivalently, iff \( \varphi \) belongs to all minimal expansions of \( T \). In this paper I characterize \( Cn_{AE} \) in terms of minimal-knowledge assumption MKA formulated as follows:

A sentence \( \varphi \) of \( L_K \) is derivable from \( T \) under MKA if and only if \( \varphi \) does not add modally positive S5-consequences to \( T \).

Because modally positive sentences of \( L_K \) express the knowledge (as opposed to ignorance), MKA articulates, in fact, the “only knowing” interpretation of \( T \) in terms of S5-provability.

The main result of this paper has the form of a completeness theorem which says that for every set \( T \) of sentences of \( L_K \),

\[
Cn_{AE}(T) = MKA(T).
\]
3. STABLE THEORIES AND INTROSPECTIVE KNOWERS

Following [Moo85], I adopt the concept of fully rational and introspective knower (called agent in op. cit.) in order to provide the language $L_K$ with its autoepistemic semantics. To that end, Moore resorted, albeit implicitly, to the fixed-point equation (1) constrained by (4) and (6). Below, I reintroduce that semantics, using More’s presentation, which is going to be the point of departure for my study of autoepistemic logic. However, later I will drop any references to (1) and its solutions $E$.

Every knower possesses certain knowledge which is completely characterized by a set $E$ of sentences of $L_K$, that is, the knower knows that a sentence is true if and only if it belongs to $E$. Modal-free elements of $E$ constitute knower’s objective knowledge $E_{Obj}$. The modal operator $K$ is understood as an abbreviation of “the knower knows that ...”, so that the knower’s knowledge can also reflect his awareness of what he actually knows. Rationality of the knower is represented by the requirement that $E$ is closed under the propositional consequence, while his introspectiveness is expressed by two conditions: $\varphi \in E$ implies $K\varphi \in E$, and $\varphi \notin E$ implies $\neg K\varphi \in E$. All three of them constitute the so called stability conditions, usually attributed to Stalnaker. The knower is a model for a knowledge base $T$ if, perhaps in addition to other requirements, all sentences of $T$ are known to him.

It is convenient to distinguish consistent stable theories, i.e. the ones which do not contain both $\varphi$ and $\neg \varphi$. Consistent stable theory $E$ may be equivalently defined by the following three conditions:

Sta 1. $E$ is closed under the propositional consequence.
Sta 2. $\varphi \in E$ iff $K\varphi \in E$.
Sta 3. $\varphi \notin E$ iff $\neg K\varphi \in E$.

Because there is only one inconsistent theory closed under the propositional consequence, namely the one which contains all the sentences of $L_K$, one can define the stability as inconsistency or satisfiability of Sta 1 ... Sta 3 (it is an easy exercise to prove that this definition is equivalent to the Stalnaker’s one). Let us note that the $or$ in the above formulation is exclusive one since the inconsistent theory closed under the propositional consequence does not satisfy Sta 3.

The adoption of fully rational and introspective knower as a model for autoepistemic language $L_K$ imposes a constrain on interpretation of phrase “only knowing”. In fact, any fully rational knower who knows $T$ must also know all the logical consequences of $T$. Therefore, the adjective “only” in “only knowing” should not be understood literally. On the other hand, as a result of knower’s full introspectiveness this adjective should not be understood too inclusively, either. For instance, knowing only $\{Kp_1 \lor Kp_2\}$ is not the same a knowing all the logical consequences of $\{Kp_1 \lor Kp_2\}$. To the contrary, it means knowing $p_1$ but not knowing $p_2$, or vice versa. This is more than only knowing $p_1 \lor p_2$.

As I indicated Section 2, my goal is to characterize, by purely proof-theoretic means, the autoepistemic consequence operation $Cn_{AE}$. Before I do that I have to spell out a scheme of semantic definition of $Cn_{AE}$:

$\varphi \in Cn_{AE}$ iff each knower who only knows $T$ also knows $\varphi$.

Obviously, the condition on the right hand side of the above scheme is not the same as “each knower who knows $T$ also knows $\varphi$ because the latter does not capture the “only knowing” interpretation of $T$. Indeed, the latter condition would lead to the following consequence operation $Cn_{Sta}$:

$$Cn_{Sta}(T) = \bigcap \{E \mid T \subseteq E \text{ and } E \text{ is stable}\}. \quad (8)$$

It turns out that $Cn_{Sta}$ coincides with monotonic consequence operation $Cn_{SS}$, and therefore cannot coincide with nonmonotonic $Cn_{AE}$. In the next section I will formulate a satisfactory definition of $Cn_{AE}$. Here I present an argument that $Cn_{Sta}$ coincides with $Cn_{SS}$.

$Cn_{SS}(T)$ is defined for every $T \subseteq L_K$ as the least set satisfying the following postulates.

AXIOMS. For every $\varphi, \psi \in L_K$,

1. $K\varphi \supset \varphi$
2. $\neg K\varphi \supset K\neg K\varphi$
3. $K(\varphi \supset \psi) \supset (K\varphi \supset K\psi)$
are in \( Cn_{S5}(T) \). Moreover, for every propositional tautology \( \tau \in L_K \),

\[
\text{Pl. } \tau
\]
is in \( Cn_{S5} \).

**RULES OF INFERENCE.** For every \( \varphi, \psi \in L_K \),

- MP, if \( \varphi \in Cn_{S5}(T) \) and \( \varphi \lor \psi \in Cn_{S5}(T) \) then \( \psi \in Cn_{S5}(T) \),
- RN, if \( \varphi \in Cn_{S5}(T) \) then \( K\varphi \in Cn_{S5}(T) \).

(Cf. [Che80] for details on system S5 with Rules of inference.

For example, \( p_2 \supset Kp_1 \in Cn_{S5}(\{p_1\}) \), because by RN \( Kp_1 \in Cn_{S5}(\{p_1\}) \), \( Kp_1 \supset (p_2 \supset Kp_1) \) is a propositional tautology in \( L_K \), and \( p_2 \supset Kp_1 \in Cn_{S5}(\{p_1\}) \) follows from \( Kp_1 \in Cn_{S5}(\{p_1\}) \) and \( Kp_1 \supset (p_2 \supset Kp_1) \in Cn_{S5}(\{p_1\}) \) by application of MP.

Operation \( Cn_{S5} \) is monotonic, that is, if \( T \subseteq T' \) then \( Cn_{S5}(T) \subseteq Cn_{S5}(T') \). This fact is a direct result of the *inclusive* character of its both rules of inference.

The completeness theorem of modal system S5 (cf. [Che80]) yields the following straightforward, nevertheless useful consequence known to autoepistemic logicians from some time (an early proof was published in [Kon88]; see also [MaT91] p. 594).

**Theorem 3.1.** For every set \( T \) of sentences of \( L_K \),

\[
Cn_{S5}(T) = Cn_{Sta}(T).
\]

**Proof postponed to the end of Section 4.**

We have already noted that \( Cn_{S5} \) is monotonic. Therefore, Theorem 3.1 affirms that \( Cn_{Sta} \) is monotonic too. Although it can be easily shown that all the axioms of S5 are valid in autoepistemic semantics of \( L_K \), and that its rules of inference never derive invalid conclusions from valid premises, the monotonicity of \( Cn_{S5} \) and \( Cn_{Sta} \) eliminates them from consideration as candidates for complete characterization of \( Cn_{AE} \).

### 4. UNIVERSAL MODELS FOR STABLE THEORIES

In this section I define a special case of Kripke structures which will be subsequently used to define the autoepistemic semantics for language \( L_K \).

A *universal Kripke structure* for \( L_K \) is a Kripke structure of the form

\[
\mathfrak{M} = (M, P),
\]

where

- \( M \) is a set of possible worlds of \( \mathfrak{M} \), and
- \( P = \langle P_1, P_2, ... \rangle \) is an infinite sequence of subsets of \( M \).

The satisfaction relation \( \models \), with \( \mathfrak{M} \models \varphi[m] \) meaning: \( \mathfrak{M} \) satisfies \( \varphi \) in a possible world \( m \in M \), is defined by induction as follows.

**Definition 4.1.**

1. \( \mathfrak{M} \not\models \bot[m] \); \( \mathfrak{M} \models T[m] \)
2. \( \mathfrak{M} \models p_i[m] \) iff \( m \in P_i \)
3. \( \mathfrak{M} \models \neg \varphi[m] \) iff \( \mathfrak{M} \not\models \varphi[m] \)
4. \( \mathfrak{M} \models (\varphi \lor \psi)[m] \) iff \( \mathfrak{M} \models \varphi[m] \) or \( \mathfrak{M} \models \psi[m] \)
5. \( \mathfrak{M} \models K\varphi[m] \) iff for each \( n \in M \), \( \mathfrak{M} \models \varphi[n] \)

* (other connectives are treated as appropriate abbreviations). Moreover,


\[
\mathfrak{M} \models \varphi \text{ iff for each } m \in M, \mathfrak{M} \models \varphi[m]
\]

and

\[
\mathfrak{M} \models T \text{ iff for each } \varphi \in T, \mathfrak{M} \models \varphi.
\]

\( \square \)
In particular, \( M \models K\varphi \) iff \( M \models \varphi \), and \( M \models \neg K\varphi \) iff \( M \not\models \varphi \).

The semantics for \( L_K \) defined above is completely characterized by the axioms and rules of inference of modal system S5 introduced in Section 3.

**Theorem 4.2.** For every set \( T \) of sentences of \( L_K \), and every sentence \( \varphi \) of \( L_K \), \( \varphi \in Cn_{S5}(T) \) iff for each universal Kripke structure \( M \models T \), \( M \models \varphi \) holds.

**Proof.** A direct proof of this fact may be found in [Che80], thm. 5.15. Here I present a simple argument pointed out in [MaT91]. Clearly, all the axioms \( T \), 5, \( K \) and \( Pl \) are valid in any universal Kripke structure, and rules MP and RN lead from true sentences to true sentences (proof by inspection). This gives the soundness. Proof of completeness requires a trick. Every standard Kripke structure for S5 (i.e. one of the form \( M = \langle M, R, P \rangle \), where \( R \) is an equivalence relation in \( M \)) may be split onto the union of universal Kripke structures of the form

\[
\mathbb{M}^{(i)} = \langle M^{(i)}, M^{(i)} \times M^{(i)}, P^{(i)} \rangle, i \in I,
\]

where \( M^{(i)} \)'s are the equivalence classes of \( R \), and \( P^{(i)}_j \) is defined as \( P_j \cap M^{(i)} \), for each \( j = 1, 2, ... \). It is a matter of routine induction to show that for every \( m \in M^{(i)} \), and every sentence \( \varphi \) of \( L_K \), \( \mathbb{M} \models \varphi[m] \) iff \( \mathbb{M}^{(i)} \models \varphi[m] \). Consequently, \( \mathbb{M} \models \varphi \) holds iff for each \( i \in I \), \( \mathbb{M}^{(i)} \models \varphi \). Therefore, if a sentence \( \varphi \) is true in all universal Kripke models of \( T \) then it is true in all Kripke models of \( T \), and hence, by completeness of modal logic S5 with respect to standard Kripke semantics (cf. [Che80], thm. 5.14), provable by means of S5.

**Proof**. Universal Kripke structures constitute a semantic counterpart of consistent stable theories in the following sense.

Let \( Th(\mathbb{M}) = \{ \varphi \in L_K \mid \mathbb{M} \models \varphi \} \). It is a matter of straightforward verification to show that for every universal Kripke structure \( \mathbb{M} \), \( Th(\mathbb{M}) \) is stable. Obviously, \( Th(\mathbb{M}) \) is consistent as well. The converse is also true: if \( E \) is consistent and stable then there is a universal Kripke structure \( \mathbb{M} \) with \( E = Th(\mathbb{M}) \). Indeed, let \( M \) be the set of all propositional models of \( E_{Obj} \), and for every propositional variable \( p_i \), let \( P_i \) be the set of all those elements of \( M \) which satisfy \( p_i \). Put \( \mathbb{M} = \langle M, (P_1, P_2, ...) \rangle \). We have \( E_{Obj} = (Th(\mathbb{M}))_{Obj} \). Routine induction shows that \( E = (Th(\mathbb{M})) \). Thus the stable consistent theories are exactly the theories of universal Kripke structures.

Now, the proof of Theorem 3.1 becomes an easy exercise. By Theorem 4.2, \( \varphi \in Cn_{S5}(T) \) iff \( \varphi \) is true in every universal Kripke model of \( T \), that is, if for every universal Kripke structure \( \mathbb{M} \) with \( T \subseteq Th(\mathbb{M}) \), \( \varphi \in Th(\mathbb{M}) \) holds. By the previous observation this means that \( \varphi \) belongs to all stable theories which contain \( T \), that is, \( \varphi \in Cn_{Sta}(T) \).

### 5. MAXIMAL MODELS FOR MINIMAL EXPANSIONS

The adoption of the concept of fully rational introspective knower whose knowledge is faithfully expressed by a stable theory \( E \) as a model for a knowledge base represented by a set of sentences of \( L_K \) restricts the area of consideration of candidates for operation \( Cn_{AE} \) to those which satisfy the following equation:

\[
Cn_{AE}(T) = \bigcap \{ E \mid T \subseteq E \text{ and } E \text{ is stable, and } ... \} \tag{9}
\]

where “...” represents some extra requirement imposed on \( E \). If “...” is interpreted as the empty requirement then equation (9) defines the operation \( Cn_{Sta} \), which coincides with monotonic consequence operation \( Cn_{S5} \) of system S5, and therefore is not an adequate characterization of \( Cn_{AE} \). If “...” is understood as “\( E \) satisfies fixed-point equation (16) of Section 10” then (9) defines \( Cn_{Exp} \). As we have seen, operation \( Cn_{Exp} \), although nonmonotonic, does not seem appropriate for the formalization of \( Cn_{AE} \), either. So, we have to look for another interpretation of “...”.

I based my choice on the concept of minimal expansion introduced in [HM85]. It comes from an observation that for purely objective consistent \( T \), the inductive definition of a unique stable theory \( E \) which contains \( T \) captured properly the introspectiveness of \( Cn_{AE} \).

In the simplest case when \( T \) contains exclusively sentences of \( L \) (i.e. modal-free sentences), constructing \( Cn_{AE}(T) \) is a matter of induction: put \( Cn_{AE}^{(0)}(T) = Cn(T); Cn_{AE}^{(n+1)}(T) = Cn(K(Cn_{AE}^{(n)}(T)) \cup \neg K(L^n_K \setminus Cn_{AE}^{(n)}(T))) \); and \( Cn_{AE}(T) = \bigcup_{n \in \omega} Cn_{AE}^{(n)}(T) \); where \( Cn \) is the operation of propositional consequence, \( K \Phi \) and \( \neg K \Phi \) abbreviate \( \{ K\varphi \mid \varphi \in \Phi \} \) and \( \{ \neg K\varphi \mid \varphi \in \Phi \} \).
respectively, and $L_K^{(n)}$ is the set of sentences of $L_K$ whose nesting depth of operator $K$ is not greater than $n$ (for instance, $\neg K(p_1 \lor p_2) \in L_K^{(2)} \setminus L_K^{(1)}$). In particular, for every $\psi \in L_K^{(n)}$, the following negation clause holds:

$$\text{if } \psi \notin Cn_{AE}^{(n)}(T) \text{ then } \neg K\psi \in Cn_{AE}^{(n+1)}(T).$$

Quite obviously, clause (10) causes the defined above fragment of $Cn_{AE}$ to be nonmonotonic. Indeed, if $p_1$ and $p_2$ are propositional variables then $\neg Kp_2 \notin Cn_{AE}^{(1)}(\{p_1\})$, but obviously $\neg Kp_2 \notin Cn_{AE}^{(1)}(\{p_1, p_2\})$.

In the case of $T$ containing arbitrary sentences of $L_K$, determining the operation $Cn_{AE}$ turned out to be a considerably more difficult task (although still relatively easy for certain special cases, as e.g., honest theories investigated in [HM85]).

If we look closer at that inductive definition then we figure out that an instance of its negation clause (10) infers $\neg K\psi \in Cn_{AE}^{(1)}(T)$, which is responsible for the uniqueness of $E$, reveals our original intention about the meaning of the operator $K$: we want to maximize the set of those sentences $\varphi$ of $L$ for which $\neg K\varphi$ is true. (McDermott and Doyle had, most likely, a similar desire when they proposed in [MD80] their famous nonmonotonic rule of inference). Now, the choice for “...” in equation (9) becomes obvious: we fill the dots with “the objective part $E_{Obj}$ of $E$ is minimal”. This gives rise to the following definition of [HM85].

**Definition 5.1.** A stable theory $E$ is called a minimal expansion of $T$ if and only if:

1. $T \subseteq E$, and
2. for every stable theory $E'$ with $T \subseteq E'$, if $E' \sqsubseteq E$ then $E' = E$,

where the relation $\sqsubseteq$ between stable theories is defined by:

$$E' \sqsubseteq E \text{ iff } E^\text{Obj} \subseteq E^\text{Obj}.$$  

The relation $\sqsubseteq$ compares stable theories with respect to the contents of their objective parts, i.e. $E' \subseteq E$ means that $E'$ contains no more objective knowledge than $E$. Thus a minimal expansion of $T$ is a stable expansion of $T$ with possibly minimal content of objective knowledge.

The concept of minimal expansion translates (9) into the following definition of the autoepistemic consequence operation $Cn_{AE}$:

$$Cn_{AE}(T) = \bigcap \{E \mid E \text{ is a minimal expansion of } T\}. \quad (11)$$

It turns out that minimal expansions have their semantical counterparts within the class of universal Kripke structures for $L_K$, namely: the maximal models introduced in [HM85]. They are defined as follows.

Let for any two universal Kripke structures $\mathfrak{M}$ and $\mathfrak{M}'$ $\mathfrak{M} \subseteq \mathfrak{M}'$ mean: $M \subseteq M'$ and for every $i = 1, 2, ..., P_i = P'_i \cap M$.

**Definition 5.2.** [HM85] The relation $\triangleleft$ between universal Kripke structures is defined by:

$\mathfrak{M} \triangleleft \mathfrak{M}'$ iff $\mathfrak{M} \subseteq \mathfrak{M}'$ and there exists $n \in N \setminus M$ such that for every $m \in M$ there is a (modal-free) sentence of $L$ with $\mathfrak{M} \models \varphi[m]$ and $\mathfrak{M} \models \neg \varphi[n]$.

$\mathfrak{M}$ is a maximal model of $T$ iff $\mathfrak{M}$ is a universal Kripke structure for $L_K$ with $\mathfrak{M} \models T$, such that for no $\mathfrak{M}' \triangleright \mathfrak{M}$, $\mathfrak{M}' \models T$. \hfill \Box

I call the semantics of $L_K$ restricted to maximal models a maximal semantics. Maximal semantics defines its consequence operation $Cn_{max}$ by:

$$Cn_{max}(T) = \{ \varphi \in L_K \mid \text{ for every maximal model } \mathfrak{M} \text{ of } T, \mathfrak{M} \models \varphi \}.$$  

It is a matter of straightforward inspection to figure out that for any two universal Kripke structures $\mathfrak{M}$ and $\mathfrak{M}'$ with $\mathfrak{M} \triangleleft \mathfrak{M}'$, and every modal-free sentence $\varphi$ of $L_K$, if $K\varphi$ is true in $\mathfrak{M}'$ then it is also true in $\mathfrak{M}$. Therefore, for any two stable theories $E$ and $E'$, $E \sqsubseteq E'$ iff there exist universal Kripke structures $\mathfrak{M}$ and $\mathfrak{M}'$ with $\mathfrak{M} \triangleleft \mathfrak{M}'$, such that $E = Th(\mathfrak{M})$ and $E' = Th(\mathfrak{M}')$. This means that the minimal expansions of $T$ are exactly the theories of maximal models of $T$.

The above observation yields the following fact (due to [HM85]) which articulates the completeness of operation $Cn_{AE}$ with respect to maximal semantics.
Theorem 5.3. For every set $T$ of sentences of $L_K$,

$$Cn_{\text{max}}(T) = Cn_{\text{AE}}(T).$$

It should be noted that although axioms of modal system S5 were not involved in the definition of $Cn_{\text{AE}}$, Theorem 5.3 implies that for every $T$, $Cn_{\text{S5}}(T) \subseteq Cn_{\text{AE}}(T)$, that is, $Cn_{\text{AE}}$ is closed under the S5-consequence (because $Cn_{\text{max}}$ obviously is). I will demonstrate in Section 8 that $Cn_{\text{S5}}$ is a maximal monotonic consequence operation with this property, and therefore may be considered the monotonic fragment of $Cn_{\text{AE}}$.

6. NORMAL AND CLAUSAL FORMS

In this technical section, I prove certain normal and clausal form lemmas which state that Boolean combinations of sentences of the form $K\phi$, where $\phi \in L$, called here $K_1$-sentences, possess the expressive power of the entire language $L_K$. This property is a necessary prerequisite for Section 7. To that end I introduce two translations: $h_x$, from $L_K$ into a first-order language $L_x$ with one variable $x$, and $f_x$, from the class of first-order structures for $L_x$ onto a certain class of Kripke structures. They are defined as follows.

Definition 6.1. Let $L_x$ be a first-order language with one variable $x$ and infinitely many unary predicate symbols $P_i$, $i = 1, 2, \ldots$. Mapping $f_x$ from the class of first-order structures for $L_x$ onto the class of universal Kripke structures for $L_K$ is defined by

$$f_x(\mathfrak{M}) = (\{m \in M \mid \mathfrak{M} \models P_i(x) \mid i = 1, 2, \ldots\})_{\text{PC}}$$

where $\models$ denotes the first-order satisfaction relation, and $M$ is the domain of the first-order structure $\mathfrak{M}$ for $L_x$.

Mapping $h_x$ from language $L_K$ onto language $L_x$ is defined by induction:

1. $h_x(\bot) = \bot, h_x(\top) = \top$
2. $h_x(p_i) = P_i(x)$
3. $h_x(\neg \phi) = \neg h_x(\phi)$
4. $h_x(\phi \lor \psi) = h_x(\phi) \lor h_x(\psi)$
5. $h_x(\forall x \phi) = \forall x h_x(\phi)$

(other connectives are treated as appropriate abbreviations).

Mappings $f_x$ and $h_x$ are “1-1” and “onto”, therefore, the inverse functions $f_x^{-1}$ and $h_x^{-1}$ are unambiguously determined. All four of them allow us to “translate” certain well-known properties of first-order logic back to $L_K$, using the following theorem.

Theorem 6.2. Let $\phi$ be a sentence of $L_K$, $\mathfrak{M}$ a first-order structure for $L_x$, and $m \in M$.

$$f_x(\mathfrak{M}) \models \phi[m] \iff \mathfrak{M} \models_{\text{PC}} h_x(\phi)[m].$$

Proof. If $\phi$ is a propositional variable (say, $p_i$) then $f_x(\mathfrak{M}) \models p_i[m] \iff \mathfrak{M} \models_{\text{PC}} P_i(x)[m] \iff \mathfrak{M} \models h_x(p_i)[m]$. Cases of $\bot$ and $\top$ are trivial. The rest of the proof is a routine induction. □

In particular, the above theorem has two useful consequences.

Corollary 6.3. For every set $T$ of sentences of $L_K$, and for every sentence $\phi$ of $L_K$,

$$\phi \in Cn_{\text{S5}}(T) \iff h_x(\phi) \in Cn_{\text{PC}}(h_x(T)),$$

where $Cn_{\text{PC}}$ denotes the first-order consequence operation in $L_x$.

Proof. By completeness of S5, $\phi \in Cn_{\text{S5}}(T)$ iff $\phi$ is true in each universal Kripke model of $T$, which by Theorem 6.2 yields equivalently: $h_x(\phi)$ is true in each first-order model of $h_x(T)$. Completeness of first-order propositional calculus completes the proof. □
Corollary 6.4. If \( \varphi \) is a modally closed sentence of \( L_K \) (i.e. every occurrence of propositional variable in \( \varphi \) belongs to the scope of operator \( K \) in \( \varphi \)) then
\[
\varphi = K \varphi
\]
is a theorem of S5.

Proof. If \( \varphi \) is modally closed then \( h_x(\varphi) \) is a sentence of \( L_x \). Therefore, \( M \models^P C h_x(\varphi)[m] \) iff \( M \models \forall x h_x(\varphi)[m] \). Applying Theorem 6.2 we obtain: \( f_x(M) \models \varphi[m] \) iff \( f_x(M) \models K \varphi[m] \), i.e. \( f_x(M) \models (\varphi \equiv K \varphi)[m] \). Because \( f_x \) is “onto”, all universal Kripke structures are of the form \( f_x(M) \). Hence \( \varphi \equiv K \varphi \) is valid.

I need these results to prove the S5-reducibility of the sentences of \( L_K \) to normal and clausal forms.

Normal Form Lemma 6.5. For every sentence \( \varphi \) of \( L_K \) there exists a \( K_1 \)-sentence \( \psi \) of \( L_K \) with
\[
Cn_{S5}(\varphi) = Cn_{S5}(\psi).
\]

Proof. Let \( \varphi \) be a sentence of \( L_K \). Because \( h_x(\varphi) \) is a formula of \( L_x \), we have \( Cn_{PC}(h_x(\varphi)) = Cn_{PC}(\forall x h_x(\varphi)) \). Because \( L_x \) has only one variable, there is a sentence \( \chi \) of \( L_x \) such that every occurrence of its only variable \( x \) belongs to the scope of exactly one quantifier \( \forall \) (we assume that \( \exists \) is an abbreviation for \( \neg \forall \neg \)). With \( Cn_{PC}(\chi) = Cn_{PC}(\forall x h_x(\varphi)) \). By corollary 6.3 we obtain \( Cn_{S5}(h_x^{-1}(\chi)) = Cn_{S5}(\varphi) \). Observation that \( h_x^{-1}(\chi) \) is a \( K_1 \)-sentence completes the proof.

Clausal Form Lemma 6.6. For every set \( T \) of sentences of \( L_K \) there exists a set \( U \) of sentences of the form
\[
K \varphi_1 \lor \ldots \lor K \varphi_n \lor \neg K \psi_1 \lor \ldots \lor \neg K \psi_m,
\]
where all \( \varphi_i \)'s and \( \psi_j \)'s are modal-free, with
\[
Cn_{S5}(T) = Cn_{S5}(U).
\]

Proof. By Theorem 6.5 every sentence \( \varphi \) of \( L_K \) is S5-equivalent to a \( K_1 \)-sentence of \( L_K \), and therefore is S5-equivalent to a \( K_1 \)-sentence in conjunctive normal form with respect to atoms of the form \( K \psi \), where \( \psi \) is modal-free. Let \( \varphi_1 \land \ldots \land \varphi_n \) be such a \( K_1 \)-sentence in conjunctive normal form. Let \( \kappa(\varphi) = \{ \varphi_1, \ldots, \varphi_n \} \), and let \( U = \bigcup \{ \kappa(\varphi) \mid \varphi \in T \} \). Observation that \( Cn_{S5}(T) = Cn_{S5}(U) \) completes the proof.

Other normal forms of autoepistemic sentences were investigated in [MaT91].

The latter result seems particularly interesting from the point of view of uniform representation of autoepistemic theories in a form of sets of clauses. This form of representation allows for transfer of methods and results of logic programming (cf. [Apt90]) into autoepistemic logic.

7. Preservation Properties

In this section I interpret in modal language \( L_K \) two classic theorems which turned out exceptionally useful in study of minimal model semantics of deductive data bases and logic programs. For this purpose I map \( K_1 \)-sentences of \( L_K \) into a first-order language \( L^H \). This mapping allows us to reflect expressible properties of structures for \( L^H \) into the language \( L_K \) and its semantics. Theorems 6.5 and 6.6 guarantee that translating just \( K_1 \)-sentences is enough to cover entire \( L_K \). Quite naturally, part of terminology of this section comes from theory of minimal models (cf. [McCS80; Min82; BS84; Sue90]).

Language \( L^H \) is defined as one without \( = \), with function symbols, constants, and a unary predicate \( R \), whose terms are the modal-free sentences of \( L_K \) (formally, we must include \( \bot \), \( \top \) and propositional variables of \( L_K \) as constants of \( L^H \), and propositional connectives of the modal-free fragment \( L \) of \( L_K \) as function symbols of \( L^H \)). Herbrand structures for \( L^H \) are defined as first-order structures of the form
\[
M = \langle L, R \rangle,
\]
where \( L \) is the domain of \( M \) and consists of all constant terms of \( L^H \) (i.e. of all modal-free sentences of \( L_K \)) and \( R \) is a subset of \( L \). As it is usually the case with Herbrand models, all constant terms of \( L^H \) are interpreted in \( M \) by themselves, e.g. term \( p_3 \lor \bot \) always the value “\( p_3 \lor \bot \)” in \( M \).
Satisfaction relation $\models$ between the Herbrand structures and formulas of $L^H$ is defined in a usual way (see [Bar78] for details).

Partial ordering relation $\leq$ between Herbrand structures is defined by

$$\mathcal{M} \leq \mathcal{M}' \text{ iff } R \subseteq R'.$$

If $T$ is a set of sentences of $L^H$ then a Herbrand structure $\mathcal{M}$ for $L^H$ is called a minimal model of $T$ $^PC$ iff $\mathcal{M} \models T$, and for every Herbrand structure $\mathcal{N}$ with $\mathcal{N} \models T$, $\mathcal{N} \leq \mathcal{M}$ implies $\mathcal{N} = \mathcal{M}$.

A set $T$ of sentences of $L^H$ is minimally modelable iff for every Herbrand structure $\mathcal{M}$ with $\mathcal{M} \models T$, there is a minimal model $\mathcal{N}$ of $T$ with $\mathcal{N} \leq \mathcal{M}$.

I distinguish two special subsets of sentences of $L^H$: the set of all quantifier-free sentences of $L^H$, denoted by $QF$, and the set of all positive (i.e. negation-free) sentences of $L^H$, denoted by $\text{Pos}$. Let us recall here that I treat all other connectives than $\lor, \land, \neg$ as appropriate abbreviations. In particular, $\varphi \lor \psi$ is an abbreviation for $\neg \varphi \land \psi$, therefore is not a positive sentence. Moreover, for any two sets $T, W$ of sentences of $L^H$, we use $T_W$ as an abbreviation of $\text{Cn}_{PC}(T) \cap W$, where $\text{Cn}_{PC}$ is the first-order consequence operation. For instance, $T_{Pos \cap QF}$ denotes the set of all positive, quantifier-free consequences of $T$. Similarly, for sets of sentences of $L_K, T_W$ abbreviates $\text{Cn}_{SS}(T) \cap W$.

Here are the results I want to translate from $L^H$ to $L_K$. The first one is an immediate corollary of a stronger theorem due to Bosu and Siegel ([BS84]).

**Lemma 7.1.** Every set of quantifier-free sentences of $L^H$ is minimally modelable. □

The second result, whose stronger form is due to Lyndon ([Lyn59]), relates $\leq$ and the truthfulness of the quantifier-free positive sentences.

**Lyndon Theorem 7.2.** Let $\mathcal{M}$ be a Herbrand structure and $T$ be a set of quantifier-free sentences of $L^H$.

$$\mathcal{M} \leq T_{Pos \cap QF} \text{ iff there exists } \mathcal{N} \leq \mathcal{M} \text{ with } \mathcal{N} \models T.$$

**Proof.** (⇒). If $\mathcal{N} \leq \mathcal{M}$ and $\mathcal{N} \models PC \mathcal{T}$ then by [Lyn59], thm 5, $\mathcal{M} \models PC T_{Pos}$, in particular, $\mathcal{M} \models PC T_{Pos \cap QF}$.

(⇐). Let $\mathcal{M} \models PC T_{Pos \cap QF}$. Because $\mathcal{M}$ is a Herbrand structure and $T$ is quantifier-free, $\mathcal{M} \models PC T_{Pos}$. Applying [Lyn59], thm 5 again, there are $\mathcal{A} \succeq \mathcal{M}$ and $\mathcal{B} \leq \mathcal{A}$ with $\mathcal{B} \models T$. Because $T$ is a universal theory, by Łoś - Tarski theorem ([Kei78], thm 3.11), $\mathcal{B} \cap M \models T$. $\mathcal{B} \leq \mathcal{A}$ yields $\mathcal{B} \cap M \leq \mathcal{A} \cap M = \mathcal{M}$. Letting $\mathcal{N} = \mathcal{B} \cap M$ gives $\mathcal{N} \leq \mathcal{M}$ and $\mathcal{N} \models PC T$.

Positive sentences were instrumental in formulations and proofs of several completeness theorems in first-order minimal model theory (cf. [Suc93]). To translate Theorem 7.2 from $L^H$ back to $L_K$ we need to distinguish their counterparts in $L_K$, namely modal positive sentences, which we define as the ones without occurrences of operator $K$ in a scope of negation. For instance, $K\neg p_1$ is a modally positive sentence, while $\neg Kp_1$ is not. I denote their class by $m\text{Pos}$.

The following mapping yields the previously mentioned translation.

**Definition 7.3.** The mapping $H$ from the set of $K_1$ sentences of $L_K$ onto set $QF$ of sentences of $L^H$ is defined inductively:

1. $H(K\varphi) = R(\varphi)$, where $\varphi$ is modal-free
2. $H(\neg \varphi) = \neg H(\varphi)$
3. $H(\varphi \lor \psi) = H(\varphi) \lor H(\psi)$
4. $H(\varphi \land \psi) = H(\varphi) \land H(\psi)$

(other connectives are treated as appropriate abbreviations). For any set $T$ of sentences of $L_K$:

$$H(T) = \{H(\varphi) \mid \varphi \in T\}.$$
Here are some basic properties of mapping $H$.

**Lemmas 7.4.**

1. $H$ is “1-1” and “onto” $QF$.
2. $H(mPos) = Pos \cap QF$.
3. For every $K_1$-sentence $\varphi$, $H(\varphi)$ is a propositional tautology iff $\varphi$ is a propositional tautology.
4. For every sequence $\varphi, \psi$ of $K_1$-sentences, $H(\varphi), H(\psi)$ is a propositional proof of $H(\psi)$ iff $\varphi, \psi$ is a propositional proof of $\psi$.
5. For every $K_1$-sentence $\varphi$ and every set $T$ of $K_1$-sentences of $L_K$,
   \[ H(\varphi) \in Cn(H(T)) \iff \varphi \in Cn(T), \]
   where $Cn$ denotes the propositional consequence operation.

**Proof.**

1. Routine induction.
2. $H(mPos) \subseteq Pos \cap QF$ is obvious. $H(mPos) \supseteq Pos \cap QF$ follows from 1 by induction based upon the inductive definition of positive quantifier-free formulas of $L^H$.
3. Propositional consequence operation treats as atoms both $R(\varphi)$’s and $K\varphi$’s, where $\varphi$ is modal-free. By virtue of 1, if $v$ is a propositional truth-assignment on $QF$ then $v \circ H$ defined by $(v \circ H)(\varphi) = v(H(\varphi))$ is a propositional truth-assignment on $K_1$, and vice versa, if $w$ is a propositional truth-assignment on $K_1$ then $v \circ H^{-1}$ is a propositional truth-assignment on $QF$. Hence, $\varphi$ is true under every propositional truth-assignment iff $H(\varphi)$ is.
4. Follows from 3.
5. Follows from 4.

$\square$

Mapping $H$ maps the relation $\leq$ onto $\sqsubseteq$, the next lemma states.

**Lemma 7.5.** For every stable theories $E$, $E'$ in $L_K$, and every Herbrand structures $\mathcal{M}, \mathcal{M}'$ for $L^H$ with $\mathcal{M} \vDash H(E_{K_1})$ and $\mathcal{M}' \vDash H(E'_{K_1})$,
\[ \mathcal{M} \preceq \mathcal{M}' \iff E \sqsubseteq E'. \]

**Proof.** ($\Rightarrow$). Assume $\mathcal{M} \preceq \mathcal{M}'$ and $E \not\subseteq E'$. Let $\varphi \in E_{Obj} \setminus E'_{Obj}$. We have: $\varphi \in E_{Obj}$ then $K\varphi \in E_{K_1}$ then $R(\varphi) \in H(E_{K_1})$ then $\mathcal{M} \vDash R(\varphi)$ then $\mathcal{M}' \vDash R(\varphi)$. Also, $\varphi \not\in E'_{Obj}$ then $\neg K\varphi \in E_{K_1}$ then $\neg R(\varphi) \in H(E'_{K_1})$ then $\mathcal{M}' \vDash \neg R(\varphi)$; a contradiction.

($\Leftarrow$). Assume $\mathcal{M} \not\preceq \mathcal{M}'$. Let $\varphi$ be in $L$ and satisfy $\mathcal{M} \vDash R(\varphi)$ and $\mathcal{M}' \not\vDash R(\varphi)$. Hence $\mathcal{M} \not\vDash \neg R(\varphi)$ and $\mathcal{M}' \not\vDash R(\varphi)$, therefore $\mathcal{M} \not\vDash H(\neg K\varphi)$ and $\mathcal{M}' \not\vDash H(K(\varphi))$, therefore $H(\neg K\varphi) \not\in H(E_{K_1})$ and $H(K(\varphi)) \not\in H(E'_{K_1})$, therefore (by stability of $E$ and $E'$, and by modal-freedom of $\varphi$) $K\varphi \not\in E_{K_1}$ and $K\varphi \not\in E'_{K_1}$, therefore $\varphi \in E$ and $\varphi \not\in E'$. Thus $E \not\subseteq E'$. $\square$

To accomplish the goal of this section I need the following technical lemmas.

**Lemma 7.6.** For every set $T$ of sentences of $L_K$,
\[ H(T_{K_1 \cap mPos}) = H(T_{K_1})_{Pos \cap QF}. \]

**Proof.** $H(T_{K_1 \cap mPos}) = H(T_{K_1}) \cap H(mPos) = (by \ Lemmas \ 7.4.2) H(T_{K_1}) \cap (Pos \cap QF) \supseteq Cn(H(T_{K_1})) \cap Pos \cap QF = H(T_{K_1})_{Pos \cap QF}$. Therefore, it suffices to prove that $H(T_{K_1})_{Pos \cap QF} \subseteq H(T_{K_1 \cap mPos})$. Let $\varphi \in H(T_{K_1})_{Pos \cap QF}$, i.e. $\varphi \in Cn(H(T_{K_1}))$ and $\varphi \in Pos \cap QF$, i.e. (by Lemma 7.4.1, 2, and 5) $H^{-1}(\varphi) \in Cn(T_{K_1})$ and $H^{-1}(\varphi) \in mPos$, i.e. $H^{-1}(\varphi) \in T_{K_1 \cap mPos}$, i.e., $\varphi \in H(T_{K_1 \cap mPos})$. $\square$

**Lemma 7.7.** For every set $T$ of $K_1$-sentences and every Herbrand structure $\mathcal{M}$ for $L^H$,
\[ \mathcal{M} \models H(T_{K_1 \cap mPos}) \iff \text{there is } \mathcal{N} \leq \mathcal{M} \text{ with } \mathcal{N} \models H(T_{K_1}). \]

**Proof.** $H(T_{K_1 \cap mPos}) = H(T_{K_1}) \cap H(mPos) = (by \ Lemmas \ 7.4.2) H(T_{K_1}) \cap Pos \cap QF \supseteq Cn(H(T_{K_1})) \cap Pos \cap QF = H(T_{K_1})_{Pos \cap QF}$. Therefore, it suffices to prove that $H(T_{K_1 \cap mPos}) \subseteq H(T_{K_1})_{Pos \cap QF}$. Let $\varphi \in H(T_{K_1 \cap mPos})$, i.e. $\varphi \in Cn(H(T_{K_1}))$ and $\varphi \in Pos \cap QF$, i.e. (by Lemma 7.4.1, 2, and 5) $H^{-1}(\varphi) \in Cn(T_{K_1})$ and $H^{-1}(\varphi) \in mPos$, i.e. $H^{-1}(\varphi) \in T_{K_1 \cap mPos}$, i.e., $\varphi \in H(T_{K_1 \cap mPos})$. $\square$
Proof. \( H(T_{K_1 \cap \text{mPos}}) = (\text{by Lemma 7.6}) H(T_{K_1})_{\text{Post} \cap \text{QE}}. \) Application of Theorem 7.2 completes the proof.

**Lemma 7.8.** For every consistent stable theory \( E \) in \( L_K \) there exists a unique Herbrand structure \( \mathfrak{M} \) for \( L^H \) with \( \mathfrak{M} \models H(E_{K_1}). \)

**Proof.** For every \( \varphi \in L, K\varphi \in E_{K_1} \) or \( \neg K\varphi \in E_{K_1} \), i.e., \( R(\varphi) \in H(E_{K_1}) \) or \( \neg R(\varphi) \in H(E_{K_1}) \).

This determines interpretation of \( R \) in \( \mathfrak{M} \), which gives the uniqueness. Because \( E \) is consistent, the above condition guarantees also the existence.

**Lemma 7.9.** For every Herbrand structure \( \mathfrak{M} \) for \( L^H \) and any set \( T \) of sentences of \( L_K \), if \( \mathfrak{M} \models H(T_{K_1}) \) then there is a stable theory \( E \) in \( L_K \) with \( \mathfrak{M} \models H(E_{K_1}). \)

**Proof.** Let \( \Phi = \{ \varphi \mid \mathfrak{M} \models R(\varphi) \} \). Because \( T_{K_1} \) is closed under the propositional consequence, a routine induction verifies that \( \Phi \) is also closed under the propositional consequence. By [MaT91], Prop 2.5 p. 593, there is a stable theory \( E \) with \( E_{\text{Obj}} = \Phi \). A direct inspection shows that \( \mathfrak{M} \models H(E_{K_1}). \)

Now we are ready to translate Theorem 7.2 back to \( L_K \).

**Preservation Lemma 7.10.** For every stable theory \( E \) of \( L_K \),

\[
T_{\text{mPos}} \subseteq E \text{ iff there is stable } E' \subseteq E \text{ with } T \subseteq E'.
\]

**Proof.** \((\Rightarrow)\). \( T_{\text{mPos}} \subseteq E \) implies \( T_{K_1 \cap \text{mPos}} \subseteq E_{K_1} \) implies \( H(T_{K_1 \cap \text{mPos}}) \subseteq H(E_{K_1}). \) By Lemma 7.8, let \( \mathfrak{M} \) be the unique Herbrand model of \( H(E_{K_1}) \). Of course, \( \mathfrak{M} \models H(T_{K_1 \cap \text{mPos}}). \)

By Lemma 7.7, we get \( \mathfrak{N} \leq \mathfrak{M} \) with \( \mathfrak{N} \models H(T_{K_1}). \) By Lemma 7.9, there is a stable theory \( E' \) with \( \mathfrak{N} \models H(E'_{K_1}). \) Of course, \( H(T_{K_1}) \subseteq H(E'_{K_1}) \), hence \( T_{K_1} \subseteq E'_{K_1} \). Therefore \( T_{K_1} \subseteq E' \), and by Theorem 6.6, \( T \subseteq E' \). Lemma 7.5 gives \( E' \subseteq E \).

\((\Leftarrow)\). Let \( T \subseteq E' \subseteq E \), \( \mathfrak{N} \models H(E'_{K_1}), \) and \( \mathfrak{M} \models H(E_{K_1}). \) By Lemma 7.5 we have that \( \mathfrak{N} \leq \mathfrak{M}. \)

Moreover, \( \mathfrak{M} \models H(T_{K_1}) \) (because \( T_{K_1} \subseteq E'_{K_1} \)). By Lemma 7.7 we get \( \mathfrak{N} \models H(T_{K_1 \cap \text{mPos}}), \) which means that \( H(T_{K_1 \cap \text{mPos}}) \subseteq H(E_{K_1}) \), or \( T_{K_1 \cap \text{mPos}} \subseteq E_{K_1} \), i.e., \( T_{\text{mPos}} \cap K_1 \subseteq E \). Application of Theorem 6.6 yields \( T_{\text{mPos}} \subseteq E \).

I used the term “Preservation” in the name of Lemma 7.10 because of its following consequence.

I call a set \( T \) of sentences of \( L_K \) upward preserved under \( \sqsubseteq \) iff for every two stable theories \( E \) and \( E' \) with \( E \sqsubseteq E' \), \( T \subseteq E \) implies \( T \subseteq E' \).

**Corollary 7.11.** A set \( T \) of sentences of \( L_K \) is upward preserved under \( \sqsubseteq \) if \( T \) is S5-equivalent to a set of modally positive formulas of \( L_K \).

**Proof.** \((\Leftarrow)\) Implication to the left follows from a routine induction on the length of a modally positive formula.

\((\Rightarrow)\) Let \( T \) be upward preserved under \( \sqsubseteq \). By Theorem 3.1, it suffices to show that for every stable theory \( E \), conditions \( T \subseteq E \) and \( T_{\text{mPos}} \subseteq E \) are equivalent. \( T \subseteq E \) obviously implies \( T_{\text{mPos}} \subseteq E \), therefore what remains to show is the opposite implication. Let \( T_{\text{mPos}} \subseteq E \). Indeed, by Lemma 7.10 there is stable theory \( E' \subseteq E \) with \( T \subseteq E' \), therefore the upward preservedness of \( T \) under \( \sqsubseteq \) yields \( T \subseteq E \).

I conclude this section with the translation of Lemma 7.1.

**Minimal Expandability Lemma 7.12.** For every set \( T \) of sentences of \( L_K \) and every stable theory \( E \supseteq T \) there is a minimal stable expansion \( E' \) of \( T \) with \( E' \subseteq E \).

**Proof.** Let \( E \) be stable with \( T \subseteq E \). We have \( H(T_{K_1}) \subseteq H(E_{K_1}). \) Let \( \mathfrak{M} \models H(E_{K_1}). \) Of course, \( \mathfrak{M} \models H(T_{K_1}). \) By Lemma 7.1 there is a minimal \( \mathfrak{N} \leq \mathfrak{M} \) with \( \mathfrak{N} \models H(T_{K_1}). \) By Lemma 7.9 there
exists stable $E'$ satisfying $\exists \mathcal{N} \models H(E'_{K_1})$. By Lemma 7.5, $E'$ is a minimal stable theory which satisfies $E' \subseteq E$ and $T_{K_1} \subseteq E'$. Hence by Theorem 6.6, $T \subseteq E'$.

Lemmas 7.10 and 7.12 are interesting in their own right. They will be instrumental in the proof of the main result of this paper. In particular, Lemma 7.12 assures that every consistent set of sentences of $L_K$ has a minimal stable expansion (Moore’s expansions do not possess this nice property).

8. THE COMPLETENESS OF THE MINIMAL-KNOWLEDGE ASSUMPTION

The minimal-knowledge assumption $MKA$, that I have been investigating with varying intensity for over a decade now (with the completeness of $MKA$ presented at [Suc05]), is defined for every set $T$ of sentences and every sentence $\varphi$ of $L_K$ as follows.

$$\varphi \in MKA(T) \text{ iff } T_{mPos} = (T \cup \{\varphi\})_{mPos}.$$ \hspace{1cm} (12)

The intuitive meaning of the above definition is that:

$\varphi$ follows from $T$ under $MKA$ iff $\varphi$ does not add new modally positive $S5$-consequences to $T$, that is, iff every modally positive sentence $S5$-provable from $T \cup \{\varphi\}$ is already $S5$-provable from $T$.

Because the modally positive sentences represent the actual knowledge contained in a knowledge base (as opposed to the modally negative ones, which represent the ignorance), $MKA(T)$ articulates the induction-like requirement that “nothing more than $T$ is known”. This indicates the circumscriptive nature of $MKA$.

Here are my main results which state that $MKA$, $Cn_{max}$, and $Cn_{AE}$ coincide.

**The Completeness Theorem 8.1.** For every set $T$ of sentences of $L_K$,

$$Cn_{AE}(T) = MKA(T).$$

**Proof.** By equation (12), we have to prove that for every sentence $\varphi$ of $L_K$,

$$\varphi \in Cn_{AE}(T) \text{ iff } (T \cup \{\varphi\})_{K_1 \cap mPos} = T_{K_1 \cap mPos}.$$ \hspace{1cm} (\Rightarrow). Let $\varphi \in Cn_{AE}(T)$, i.e. for every minimal expansion $E$ of $T$, $\varphi \in E$. Because $T_{K_1 \cap mPos} \subseteq (T \cup \{\varphi\})_{K_1 \cap mPos}$, it suffices to show that for every stable $E$, if $T_{K_1 \cap mPos} \subseteq E$ then $(T \cup \{\varphi\})_{K_1 \cap mPos} \subseteq E$. So, assume the former. By Lemma 7.12 we get a minimal stable $E'$ with $T_{K_1 \cap mPos} \subseteq E'$ and $E' \subseteq E$. By Lemma 7.10 there is a $E'' \subseteq E'$ with $T \subseteq E''$. In particular, $T_{K_1 \cap mPos} \subseteq E''$. But $E'$ is a minimal expansion for $T_{K_1 \cap mPos}$, therefore $E'' = E'$. Hence $T \cup \{\varphi\} \subseteq E'$. Now, because $E' \subseteq E$, using Lemma 7.10 again we conclude $(T \cup \{\varphi\})_{K_1 \cap mPos} \subseteq E$.

$$(\Leftarrow).$$ Assume $(T \cup \{\varphi\})_{K_1 \cap mPos} = T_{K_1 \cap mPos}$. By Lemma 7.12 there is a minimal stable expansion $E$ of $T$. We have $(T \cup \{\varphi\})_{K_1 \cap mPos} \subseteq E$. By Lemma 7.10 there is $E' \subseteq E$ with $T \cup \{\varphi\} \subseteq E'$. In particular, $T \subseteq E'$. Minimality of $E$ yields $E = E'$, i.e. $\varphi \in E$.

In conclusion we obtain

**Corollary 8.2.** For every set $T$ of sentences of $L_K$,

$$MKA(T) = Cn_{max}(T).$$

**Proof.** by application of Theorems 5.3 and 8.1.

Theorem 8.1 determines the limit of the proving power of the operation $Cn_{AE}$: no modally positive sentence $\varphi$ may be derived from $T$ using $Cn_{AE}$ unless $\varphi$ is $S5$-derivable from $T$. This restriction does not apply to iterative applications of $Cn_{AE}$ which, as I will indicate in Section 9, can produce certain new modally positive consequences of the knowledge base in question.

The following two results show close relationship between $Cn_{AE}$ and $Cn_{S5}$.

**Theorem 8.3.** For every set $T$ of sentences of $L_K$,

$$Cn_{AE}(T) = Cn_{AE}(Cn_{S5}(T)).$$

**Proof.** follows from Lemma 8.1 and fact that $MKA(T) = MKA(Cn_{S5}(T))$. 

$(Cn_{exp}$ does not have the above property; e.g. for $T = \{Kp_1\}$ the equality does not hold.)

The second, intuitively obvious, result allows one to reverse the order of definitions used in this paper and to define the consequence operation $Cn_{S5}$ of system $S5$ by means of $Cn_{AE}$. 

...
Theorem 8.4. $Cn_{S5}$ is the maximal monotonic consequence operation with

$$Cn_{S5}(T) \subseteq Cn_{AE}(T)$$

holding for every set of $T$ of sentences of $L_K$.

Proof. The inclusion follows from Theorem 3.1. Let $Cn_i$ be a monotonic consequence operation satisfying

$$Cn_{S5}(T) \subseteq Cn_i(T) \subseteq Cn_{AE}(T)$$

for every $T$, with $Cn_{S5}(T) \neq Cn_i(T)$ for some $T$. $Cn_{S5}(T)$ is characterized by all stable theories containing $T$, and $Cn_{AE}(T)$ is characterized by all minimal stable theories containing $T$. Monotonic consequence $Cn_i$ must eliminate certain stable $E$'s (otherwise it would coincide with $Cn_{S5}$). Let $E$ be an example of such eliminated stable theory. Since every stable theory is its own minimal expansion, we have $Cn_{AE}(E) = E$, but $Cn_i \nsubseteq E$. Hence $Cn_i(T) \nsubseteq Cn_{AE}(T)$.

Obviously, if $Cn_i(T) \subseteq Cn_{AE}(T)$ and $Cn_i'(T) \subseteq Cn_{AE}(T)$ then $Cn_i(T) \cup Cn_i'(T) \subseteq Cn_{AE}(T)$, thus there is only one maximal monotonic consequence operation $Cn_{S5}$ that satisfies (13).

One can extend the non-modal methods used in [Suc90; SS90; Suc93] to investigate stronger and weaker versions of $MKA$, analogically as it is the case with various versions of the closed-world assumption. For instance, one can consider a Minker style (cf. [Min82]) weak $MKA$ which restricts its content to the sentences of the form $\neg K\varphi$, where $\varphi$ is modal-free. It also appears a routine matter to bring in the quantifiers to $L_K$.

9. STRONGER VARIANTS OF $MKA$

To all appearances, $MKA$ seems like a very weak consequence operation, which is not capable of deriving modally positive conclusions from a knowledge base unless these conclusions are $S5$-derivable from the base's contents. For instance, $p_2 \not\in MKA(\neg Kp_1 \supset p_2)$. This fact seems counterintuitive to some researchers. However, if one needs to infer $p_2$ from $\{\neg Kp_1 \supset p_2\}$ (e.g., in logic programming applications it may be a legitimate requirement), then the set of premises should be partitioned onto strata, according to intentional priorities of the occurrences of negation. In our case, one can split $\{\neg Kp_1 \supset p_2\}$ onto 0 (the set of its positive clauses) and $\{\neg Kp_1 \supset p_2\}$ itself (the remainder), and easily verify that

$$p_2 \in MKA(\{\neg Kp_1 \supset p_2\} \cup MKA \upharpoonright p_1(0)),$$

where “$|$ $p_1$” means “restricted to the language of $p_1$”.

More generally, one can use mapping $H$ of Section 7 to translate the completeness theorem of prioritized closed-world assumption with respect to hierarchically minimal model semantics (thm. 5.5 in [SS90] and thm. 4.5.5 in [Suc00a]) to obtain analogous result for stratified $MKA$ and a restricted form of maximal semantics. This, for instance, has been done in [Suc00c] (cf. [Rin94] for a similar strengthening of autoepistemic logic).

10. OTHER FORMALIZATIONS OF AUTOEPISTEMIC INFERENCE

In this section I briefly discuss other suggested formalizations of autoepistemic logic known from professional literature.

The inherent inability of modal system $S5$ to derive $\neg Kp_2$ from $\{p_1\}$ was by no means easy to fix. For example, McDermott and Doyle (cf. [MD80], p. 50) added to $S5$ the anti-necessitation rule of inference (3) that they somewhat simplistically interpreted as:

$$\varphi \notin E \quad \neg K\varphi \in E$$

rather than using the fixed-point equation (1) and interpretation (4). (In [MD80], (14) was formulated as:

$$\text{If } \varphi \notin Cn_{MDD}(T) \text{ then } \neg K\varphi \in Cn_{MDD}(T).$$

This made the resulting logic an inconsistent system. For instance, both $\neg Kp_1$ and $\neg Kp_2$ are in $Cn_{MDD}(\{Kp_1 \lor Kp_2\})$, therefore $(\neg Kp_1 \lor Kp_2) \in Cn_{MDD}(\{Kp_1 \lor Kp_2\})$. (Similar anomaly plagued early versions of the closed-world assumption, e.g., the one of [Rei78]). Although the McDermott-Doyle rule (14) is inconsistent with $S5$, every stable theory $E$ does obey it via the fixed-point interpretation of
the anti-necessitation: if $\varphi \notin E$ then $\neg K \varphi \in E$! This example visualizes the subtlety of autoepistemic inference.

The inconsistency of McDermott and Doyle’s rule was addressed by many researchers. Here I briefly comment on work of Moore [Moo85], Parikh [Par91], Levesque [Lev90], Schwarz [Sch92], and Kaminski [Kam91].

As I have indicated in Section 1, the approach proposed by Moore restricted stable theories considered as models for a knowledge base $T$ to its expansions $E$, which Moore defined by means of the following fixed-point equation that encapsulated both necessitation and anti-necessitation:

$$E = Cn(T \cup \{K \varphi \mid \varphi \in E\} \cup \{\neg K \varphi \mid \varphi \notin E\}),$$

(16)

where $Cn$ denotes the propositional consequence operation. The corresponding consequence operation $Cn_{Exp}$ was given by:

$$Cn_{Exp}(T) = \bigcap\{E \mid E \text{ is an expansion of } T\}$$

(17)

Operation $Cn_{Exp}$ is nonmonotonic and properly captures the easy case of modal-free $T$. As I pointed out in Example 1.1, in some other cases $Cn_{Exp}$ reveals paradoxical behavior, e.g. the innocent theory $\{Kp_1\}$ has no expansions at all. Moreover, the implicit form of the definition of $Cn_{Exp}$ given by the fixed-point equation (16) does not make it particularly easy to compute, because in order to decide if $\varphi \in Cn_{Exp}(T)$ one has to find all expansions of $T$ first. (Substantially faster algorithms are known; cf. [MaT91]; see also [AAA10] for a more recent methods of computing expansions.)

Parikh improved the McDermott and Doyle’s rule (15) by restricting it to cases when this rule was already applied to certain subformulas of $\varphi$. However, $\{Kp_1 \lor Kp_2\}$ still remains $S5$-inconsistent under this restriction.

Schwarz investigated $S$-expansions of $T$, that is, stable solutions of fixed-point equation (1) only under the anti-necessitation rule interpreted as (4) and closed under consequence operation $Cn_S$ of a subsystem $S$ of $S5$. (In [Sch92], the fixed-point equation had the following form:

$$E = Cn_S(T \cup \{\neg K \varphi \mid \varphi \notin E\}),$$

(18)

where $S$ is a subsystem of $S5$.) Corresponding autoepistemic consequence operation was given by

$$Cn_{S-exp}(T) = \bigcap\{E \mid E \text{ is an } S\text{-expansion of } T\}.$$  

(19)

This approach yielded a variety of nonmonotonic logic which, however, did not comprise nonmonotonic $S5$ since in the case of $S = S5$, (19) defines the monotonic consequence operation $Cn_{S5}$. It has been demonstrated that Moore’s formalization of autoepistemic logic is a special case of a system defined by (19), where $S = KD45$.

Levesque suggested a use of an extra operator $O$, with sentences $O \varphi$ having the intentional meaning “only $\varphi$ is known” (although expressed in terms of belief rather than knowledge). Using the axiom of the form $(\forall x)(\text{bird}(x) \supset (\neg \text{fly}(x) \lor K \text{fly}(x)))$ rather than rule (15), he formally derived from a formalization of the well known puzzle about Tweety a modally positive sentence $K \text{fly}(\text{Tweety})$, not $S5$-provable from that formalization. His proof used, in fact, only symmetric axioms which are true for a predicate $P$ if they are true for $P$’s negation, therefore the sentence $K \neg \text{fly}(\text{Tweety})$ has a similar proof in his system. This paradox suggests either an error in the above mentioned proof or inconsistency of axiomatization of operator $O$.

Kaminski’s approach seems somewhat complementary to ours. The logics he considered are defined by (19) and (4), the latter being restricted to modal-free premises $\varphi$. (In [Kam91], (18) was replaced by:

$$E = Cn_S(T \cup \{\neg K \varphi \mid \varphi \in L \setminus E\}),$$

(20)

where $L$ was the set of modal-free sentences of $L_K$.)

For $S = S5$, the solutions of (20) are the minimal expansions of $T$, and, consequently, (19) paired with (20) characterize $Cn_{AE}$. Because of implicit character of (20), characterization (19) seems more difficult to compute (one has to find all the solutions of the equation (20) first) than $MKA$. Because both (12) and (19) may be computationally simplified, comparison of complexity of these two characterizations may require further studies. However, general undecidability of $Cn_{AE}$ (cf. [Suc03; Suc06]) makes any computational simplifications rather limited. (Cf. also [TV10] for a comprehensive review of complexity issues pertinent to autoepistemic logic.)
11. FORMALIZATIONS OF OTHER SCHEMES OF NONMONOTONIC DEDUCTION

Default logic, introduced by Reiter around 1980, is (according to characterization in [Suc00b]) a generalization of the Moore’s syntactic scheme (3) discussed in Section 1. In addition to standard propositional axioms and modus ponens, it allows for nonmonotonic rules of inference called defaults. They have the form of

\[ \varphi : M \psi_1, \ldots, M \psi_n \quad \xrightarrow{\chi} \quad (21) \]

where \( M \) is the modal operator of possibility that may be understood as an abbreviation of \( \neg K \neg \).

(Some authors prefer to skip occurrences of \( M \) in front of \( \psi_i \)'s; others use symbol \( \Diamond \), instead; also \( \varphi \) and/or \( M \psi_i \)'s are/is usually skipped if tautologically true.)

Definition of semantics of default logic is somewhat similar to Moore’s autoepistemic logic. It is based on two sets of propositions:

- \( T \) - the set being the subject of closure under the default rules, and
- \( E \) - the set of assumptions, sometimes referred to as the context.

The meaning of the rule (21) is formally described by:

\[ \frac{\varphi \in T \mid \neg \psi_1 \notin E \mid \ldots \mid \neg \psi_n \notin E}{\chi \in \Phi(T,E)} \quad (22) \]

Given a set \( D \) of defaults, operator \( \Phi(T,E) \) of closure of \( T \) under rules (21) of \( D \) relative to \( E \) is defined as the closure of \( T \) both under propositional consequence and under the corresponding productions (22) of \( D \).

Set \( E \) is called and extension (a nonmonotonic deductive closure in the sense of Section 1, that is) of \( T \) iff

\[ \Phi(T,E) = E. \]

As a result, \( E \) is both supported and closed under all rules (21) of \( D \).

**Example 11.1.** The empty theory 0 with two default rules

\[ \frac{M p}{\neg q} \quad \text{and} \quad \frac{M q}{\neg p} \]

has two extensions:

\[ E_1 = Cn(\neg p) \quad \text{and} \quad E_1 = Cn(\neg q), \]

while the empty theory 0 with one default rule

\[ \frac{M p}{\neg q} \]

has no extensions. (The latter rather counterintuitive fact is a result of the implicit requirement of the supportedness that all extensions must satisfy.) \( \square \)

Several other schemes of nonmonotonic deduction addressed specifically the use of universally quantified formulas, the so-called clauses.

More analysis of default logic may be found in [MaT90; MaT93; MN94].

Predicate (resp.: domain) circumscription, introduced by McCarthy in late 70-ties (see [McC80]) aims at defining the concept of relation-minimal model (resp.: Herbrand model) of first-order theory within its language, which goal was not met (as shown in [EMR85]) because it requires second-order (resp.: infinitary) logic. It has been later accomplished by Lifschitz (see [Lif85]) in terms of the so-called second-order circumscription. Its first-order counterpart, substantially different from the original McCarthy’s attempt to express second-order properties with first-order formulas, is given by the minimal entailment \( \vdash_\prec \), or, equivalently, by the set \( \text{Circ}_2(T) \cap L \), where \( \text{Circ}_2 \) is the second-order circumscription consequence operator and \( L \) is the set of first-order sentences.

Complete and sound characterizations of \( \vdash_\prec \) for certain classes of formulas in terms of provability (something along the lines of the completeness theorem) for a number of partial orderings \( \prec \), including the ones mentioned above, are known from the literature (e.g., [Min82; YH85; Suc94; Suc97; Suc00a]).
Also, some relationships between the logic of minimal entailment and default logic are known. For instance, a minimal model of set $T$ which admits elimination of quantifiers can be defined in terms of an extension of $T$ and the set of defaults

$$D_{nas} = \{ \frac{M \dashv \varphi}{\varphi} \mid \varphi \text{ is an atomic sentence} \}.$$ 

On the other hand, minimal entailment seems to evade the schemes defined by the defaults because it is characterized by the inference rule (first introduced in [Suc88b] and extensively studied in [Suc88a; Suc88]) of the form:

“If adding $\varphi$ to the premise set $T$ does not enlarge the set of positive consequences of $T$ then infer $\varphi$”

which clearly doesn’t fall under the scheme (22) (obviously, not under $D_{nas}$).

It turns out that (some variants of) minimal model semantics provide logic programs, usually represented as finite sets of clauses, with standard meanings. Regular behavior of such semantics is a consequence of Bossu-Siegel’s classic result of [BS84] which ascertains that for every set $T$ of clauses and every non-minimal model $\mathfrak{M}$ of $T$ there is a minimal model of $T$ below $\mathfrak{M}$.

**REFERENCES**

Esben H. Lian A, Einar Broch Johnsen A, and Arild Waaler A. Confluent term rewriting for only-knowing logics.


