Fixpoint 3-valued semantics for autoepistemic logic

Marc Denecker
Department of Computer Science
K.U.Leuven
Celestijnenlaan 200A, B-3001 Heverlee, Belgium
marcd@cs.kuleuven.ac.be

Victor Marek and Mirosław Truszczyński
Computer Science Department
University of Kentucky
Lexington, KY 40506-0046
marek|mirek@cs.engr.uky.edu

Abstract

The paper presents a constructive 3-valued semantics for autoepistemic logic (AEL). We introduce a derivation operator and define the semantics as its least fixpoint. The semantics is 3-valued in the sense that, for some formulas, the least fixpoint does not specify whether they are believed or not. We show that complete fixpoints of the derivation operator correspond to Moore's stable expansions. In the case of modal representations of logic programs our least fixpoint semantics expresses well-founded semantics or 3-valued Fitting-Kunen semantics (depending on the embedding used). We show that, computationally, our semantics is simpler than the semantics proposed by Moore (assuming that the polynomial hierarchy does not collapse).

Introduction

We describe a 3-valued semantics for modal theories that approximates skeptical mode of reasoning in the autoepistemic logic introduced in (Moore 1984; 1985). We present results demonstrating that our approach is, indeed, appropriate for modeling autoepistemic reasoning. We discuss computational properties of our semantics and connections to logic programming.

Autoepistemic logic is among the most extensively studied nonmonotonic formal systems. It is closely related to default logic (Reiter 1980). It can handle default reasonings under a simple and modular translation in the case of prerequisite-free defaults (Marek & Truszczyński 1993). In the case of arbitrary default theories, a somewhat more complex non-modular translation provides a one-to-one correspondence between default extensions and stable (autoepistemic) expansions (Gottlob 1995). Further, under the so called Gelfond translation, autoepistemic logic captures the semantics of stable models for logic programs (Gelfond 1987). Under the Konolige encoding of logic programs as modal theories, stable expansions generalize the concept of the supported model semantics

(Marek & Truszczyński 1993). Autoepistemic logic is also known to be equivalent to several other modal nonmonotonic reasoning systems.

The semantics for autoepistemic logic (Moore 1985) assigns to a modal theory T a collection of its *stable expansions*. This collection may be empty, may consist of exactly one expansion or may consist of several different expansions. Intuitively, expansions are designed to model belief states of agents with *perfect* introspection powers: for every formula F, either the formula KF (expressing a belief in F) or the formula $\neg KF$ (expressing that F is not believed) belongs to an expansion. We will say that expansions contain no *meta ignorance*.

In many applications, the phenomenon of multiple expansions is desirable. There are situations where we are not interested in answers to queries concerning a single atom or formula, but in a *collection* of atoms or formulas that satisfy some constraints. Planning and diagnosis in artificial intelligence, and a range of combinatorial optimization problems, such as computing hamilton cycles or k-colorings in graphs, are of this type. These problems may be solved by means of autoepistemic logic precisely due to the fact that multiple expansions are possible. The idea is to represent a problem as an autoepistemic theory so that solutions to the problem are in one-to-one correspondence with stable expansions. While conceptually elegant, this approach has its problems. Determining whether expansions exist is a Σ_2^P-complete problem (Niemelä 1992; Gottlob 1992), and all known algorithms for computing expansions are highly inefficient.

In a more standard setting of knowledge representation, the goal is to model the knowledge about a domain as a theory in some formal system and, then, to use some inference mechanism to resolve queries against the theory or, in other words, establish whether particular formulas are entailed by this theory. Autoepistemic logic (as well as other nonmonotonic systems) can be used in this mode, too. Namely, under

the so called *skeptical* model, a formula is entailed by a modal theory, if it belongs to all stable expansions of this theory. The problem is, again, with the computational complexity of determining whether a formula belongs to all expansions.

We propose an alternative semantics for autoepistemic reasoning that allows us to *approximate* the skeptical approach described above. Our semantics has the property that if it assigns to a formula the truth value **t**, then this formula belongs to all stable expansions and, dually, if it assigns to a formula the truth value **f**, then this formula does not belong to any expansion. Our semantics is 3-valued and some formulas are assigned the truth value **u** (unknown). While only approximating the skeptical mode of reasoning, it has one important advantage. Its computational complexity is lower (assuming that the polynomial hierarchy does not collapse on some low level). Namely, the problem to determine the truth value of a formula is in the class Δ_2^P.

Clearly, the semantics we propose can be applied whenever the situation requires that autoepistemic logic be used in the skeptical mode. However, it has also another important application. It can be used as a pruning mechanism in algorithms that compute expansions. The idea is to first compute our 3-valued interpretation for an autoepistemic theory (which is computationally simpler than the task of computing an expansion) and, in this way, find some formulas which are in all expansions and some that are in none. This restricts the search space for expansions and may yield significant speedups.

Conceptually, our semantics plays the role similar to that played by the well-founded semantics in logic programming. Deciding whether an atom is in all stable models is a co-NP-complete problem. However, the well-founded semantics, which approximates the stable model semantics can be computed in polynomial time. Furthermore, well-founded semantics is used both as the basis for top-down query answering implementations of logic programming (Chen, Swift, & Warren 1994), and as a search space pruning mechanism by some implementations to compute stable model semantics (Niemelä & Simons 1995).

Our 3-valued semantics for autoepistemic logic is based on the notion of a *belief pair*. These are pairs (P, S), where P and S are sets of 2-valued interpretations of the underlying first-order language, and $S \subseteq P$. The motivation to consider belief pairs comes from Moore's possible-world characterization of stable expansions (Moore 1984). Moore characterized consistent expansions in terms of *possible-world structures*, that is, non-empty sets of 2-valued interpretations. A

belief pair (S, P) can be viewed as an approximation to a possible-world structure W such that $S \subseteq W \subseteq P$: interpretations not in P are known not to be in W, and those in S are known to be in W. Observe that while expansions (or the corresponding possible-world structures) do not contain meta ignorance, belief pairs, in general, do.

Our semantics is defined in terms of fixpoints of a monotone operator defined on the set of belief pairs. This operator, $\mathcal{D}_T$, is determined by an initial theory T. Given a belief pair $B = (P, S)$, $\mathcal{D}_T$ establishes that some interpretations that are in P must, in fact, belong to S. In addition, some other interpretations in P are eliminated altogether, as inappropriate for describing a possible state of the world (given the agent's initial knowledge). The operator attempts to simulate a constructive process rational agents might use to produce an "elementary" improvement on their current set of beliefs and disbeliefs.

We say that (P_1, S_1) "better approximates" the agent's beliefs and disbeliefs entailed by the agent's initial assumptions than (P, S) if $S \subseteq S_1 \subseteq P_1 \subseteq P$. The operator $\mathcal{D}_T$ is monotone with respect to this ordering and, thus, it has the least fixpoint. We propose this fixpoint as a *constructive* approximation to the semantics of stable expansions.

A fundamental property that makes the above construction meaningful is that *complete* belief pairs (those with P equal to S) that are fixpoints of $\mathcal{D}$ are precisely Moore's S5-models characterizing expansions. Thus, by the general properties of fixpoints of monotone operators over partially ordered sets, the least fixpoint described above indeed approximates the skeptical reasoning based on expansions. Moreover, as mentioned above, the problem of computing the least fixpoint of the operator $\mathcal{D}$ requires only polynomially many calls to the satisfiability testing engine, that is, it is in Δ_2^P. Another property substantiating our approach is that under some natural encodings of logic programs as modal theories, our semantics yields both well-founded semantics (Van Gelder, Ross, & Schlipf 1991) and the 3-valued Fitting-Kunen semantics (Fitting 1985; Kunen 1987).

Autoepistemic logic — preliminaries

The language of autoepistemic logic is the standard language of propositional modal logic over a set of atoms Σ and with a single modal operator K. We will refer to this language as $\mathcal{L}_K$. The first-order fragment of $\mathcal{L}_K$ will be denoted by $\mathcal{L}$.

The notion of a 2-*valued interpretation* of the alphabet Σ is defined as usual: it is a mapping from Σ to $\{\mathbf{t}, \mathbf{f}\}$. The set of all interpretations of Σ is denoted $\mathcal{A}_\Sigma$ (or $\mathcal{A}$, if Σ is clear from the context).

A possible-world semantics for autoepistemic logic was introduced by Moore (Moore 1984) and proven equivalent with the semantics of stable expansions. A possible-world structure W (over Σ) is a set of 2-valued interpretations of Σ. Alternatively, it can be seen as a Kripke structure with a total accessibility relation. Given a pair (W, I), where W is a set of interpretations and I is an interpretation, one defines a truth assignment function $\mathcal{H}_{W,I}$ inductively as follows:
(1) For an atom A, we define $\mathcal{H}_{W,I}(A) = I(A)$; (2) The boolean connectives are handled in the usual way; (3) For any formula F, we define $\mathcal{H}_{W,I}(KF) = \mathbf{t}$ if for every $J \in W, \mathcal{H}_{W,J}(F) = \mathbf{t}$, and $\mathcal{H}_{W,I}(KF) = \mathbf{f}$, otherwise.

We write $(W, I) \models F$ to denote that $\mathcal{H}_{W,I}(F) = \mathbf{t}$. Further, for a modal theory T, we will write $(W, I) \models T$ if $\mathcal{H}_{W,I}(F) = \mathbf{t}$ for any $F \in T$. Finally, for a possible world structure W we define the *theory* of W, $Th(W)$, by:

$$Th(W) = \{F : (W, I) \models F, \text{ for all } I \in W\}.$$

It is well known that for every formula F, either $KF \in Th(W)$ or $\neg KF \in Th(W)$ ($\mathcal{H}_{W,I}(KF)$ is the same for all interpretations $I \in W$). Thus, possible-world structures have no meta ignorance and, as such, are suitable for modeling belief sets of agents with *perfect* introspection capabilities. It is precisely this property that made possible-world structures fundamental objects in the study of modal nonmonotonic logics (Moore 1984; Marek & Truszczyński 1993).

Definition 1 *An* autoepistemic model *of a modal theory T is a possible-world (**S5**) structure W which satisfies the following fixpoint equation:*

$$W = \{I : (W, I) \models T\}.$$

The following theorem, relating stable expansions of (Moore 1985) and autoepistemic models, was proved in (Levesque 1990) and was discussed in detail in (Schwarz 1992).

Theorem 1 *For any two modal theories T and E, E is a consistent stable expansion of T if and only if $E = Th(W)$ for some nonempty autoepistemic model W of T.*

A fixpoint 3-valued semantics for autoepistemic logic

Our semantics for autoepistemic logic is defined in terms of possible-world structures and fixpoint conditions. The key difference with the semantics proposed by Moore is that we consider *approximations* of possible-world structures by *pairs* of possible-world structures. Recall from the previous section, that $\mathcal{A}$

denotes the set of all interpretations of a fixed propositional language $\mathcal{L}$.

Definition 2 *A* belief pair *is a pair (P, S) of sets of interpretations P and S such that $P \supseteq S$. When $B = (P, S)$, $S(B)$ denotes S and $P(B)$ denotes P. The belief pair $(\mathcal{A}, \emptyset)$ is denoted $\bot$. The set $\{(P, S) : P, S \in \mathcal{A} \text{ and } P \supseteq S\}$ of all belief pairs is denoted by $\mathcal{B}$.*

The interpretations in $S(B)$ can be viewed as states of the world which are known to be possible (belong to W). They form a lower approximation to W. The interpretations in $P(B)$ can be viewed as an upper approximation to W. In other words, interpretations not in $P(B)$ are known not to be in W.

We will extend now the concept of a interpretation to the case of belief pairs and consider the question of meta ignorance and meta knowledge of belief pairs. We will show that, being only approximations to possible-world structures, belief pairs may contain meta ignorance. We will use three logical values, $\mathbf{f}$, $\mathbf{u}$ and $\mathbf{t}$. In the definition, we will use the *truth* ordering: $\mathbf{f} \leq_{tr} \mathbf{u} \leq_{tr} \mathbf{t}$ and define $\mathbf{f}^{-1} = \mathbf{t}, \mathbf{t}^{-1} = \mathbf{f}, \mathbf{u}^{-1} = \mathbf{u}$.

Definition 3 *Let $B = (P, S)$ be a belief pair and let I be an interpretation. The truth function $\mathcal{H}_{B,I}$ is defined inductively:*

$$
\begin{aligned}
\mathcal{H}_{B,I}(A) &= I(A) \quad (A \text{ is an atom}) \\
\mathcal{H}_{B,I}(\neg F) &= \mathcal{H}_{B,I}(F)^{-1} \\
\mathcal{H}_{B,I}(F_1 \vee F_2) &= max\{\mathcal{H}_{B,I}(F_1), \mathcal{H}_{B,I}(F_2)\} \\
\mathcal{H}_{B,I}(F_1 \wedge F_2) &= min\{\mathcal{H}_{B,I}(F_1), \mathcal{H}_{B,I}(F_2)\} \\
\mathcal{H}_{B,I}(F_2 \supset F_1) &= max\{\mathcal{H}_{B,I}(F_1), \mathcal{H}_{B,I}(F_2)^{-1}\}
\end{aligned}
$$

The formula $K(F)$ is evaluated as follows:

$$
\mathcal{H}_{B,I}(K(F)) = \begin{cases} \mathbf{t} & \text{if } \forall_{J \in P} \mathcal{H}_{B,J}(F) = \mathbf{t} \\ \mathbf{f} & \text{if } \exists_{J \in S} \mathcal{H}_{B,J}(F) = \mathbf{f} \\ \mathbf{u} & \text{otherwise} \end{cases}
$$

The truth value of a modal atom KF, $\mathcal{H}_{B,I}(KF)$, does not depend on the choice of I. Consequently, for a modal atom KF we will write $\mathcal{H}_B(KF)$ to denote this, common to all interpretations from $\mathcal{A}$, truth value of KF.

Let us define the *meta knowledge* of a belief pair B as the set of formulas $F \in \mathcal{L}_K$ such that $\mathcal{H}_B(KF) = \mathbf{t}$ or $\mathcal{H}_B(KF) = \mathbf{f}$. The *meta ignorance* is formed by all other formulas, that is, those formulas $F \in \mathcal{L}_K$ for which $\mathcal{H}_B(KF) = \mathbf{u}$.

Clearly, a belief pair $B = (W, W)$ naturally corresponds to a possible-world structure W. Such a belief pair is called *complete*. We will denote it by (W). The following straightforward result indicates that $\mathcal{H}_{B,I}$ is a generalization of $\mathcal{H}_{W,I}$ to the case of belief pairs. It also states that a complete belief pair contains no meta ignorance.

Proposition 1 *If B is a complete belief pair (W), then $\mathcal{H}_{B,I}$ is 2-valued. Moreover, for every formula F, $\mathcal{H}_{B,I}(F) = \mathcal{H}_{W,I}(F)$.*

In our approach to autoepistemic reasoning we will model the agent who, given an initial theory T, starts with the belief pair $\perp$ (with the smallest meta knowledge content) and, then, iteratively constructs a sequence belief pairs with increasing meta knowledge (decreasing meta ignorance) until no more improvement is possible. To this end, we will introduce now a partial ordering on the set $\mathcal{B}$ of belief pairs. Given two belief pairs B_1 and B_2, we define $B_1 \leq_p B_2$ if $P(B_1) \supseteq P(B_2)$ and $S(B_1) \subseteq S(B_2)$. This ordering is consistent with the ordering defined by the "amount" of meta knowledge contained in a belief pair: the "higher" a belief pair in the ordering $\leq_p$, the more meta-knowledge it contains (and the less meta-ignorance). It is also consistent with the concept of the *information ordering* of the truth values: $\mathbf{u} \leq_{kn} \mathbf{f}, \mathbf{u} \leq_{kn} \mathbf{t}$.

Proposition 2 *Let B_1 and B_2 be belief pairs. If $B_1 \leq_p B_2$ then for every $F \in \mathcal{L}_K$ and for every interpretation I, $\mathcal{H}_{B_1,I}(F) \leq_{kn} \mathcal{H}_{B_2,I}(F)$.*

The ordered set $(\mathcal{B}, \leq_p)$ is not a lattice. In fact, for every $W \subseteq \mathcal{A}$, (W) is a maximal element in $(\mathcal{B}, \leq_p)$. If $W_1 \neq W_2$, then (W_1) and (W_2) have no least upper bound (l.u.b.) in $(\mathcal{B}, \leq_p)$. The pair $\perp = (\mathcal{A}, \emptyset)$ is the least element of $(\mathcal{B}, \leq_p)$.

The ordered set $(\mathcal{B}, \leq_p)$ is *chain-complete*. That is, every set of pairwise comparable elements has the l.u.b. It is a well-known (Markowsky 1976) that a monotone operator defined on a chain-complete ordered set with a least element has a least fixpoint. This fixpoint is the limit of the iterations of the operator starting at the least element $\perp$.

We will now define a monotone operator on the ordered set $(\mathcal{B}, \leq_p)$ and will use it to define a step-wise process of constructing belief pairs with increasing meta knowledge. To this end, we will define two satisfaction relations: *weak* (denoted by $\models_w$) and *strong* (denoted by $\models$):

$(B, I) \models_w F$ if $\mathcal{H}_{B,I}(F) \neq \mathbf{f}$ (i.e. $\mathcal{H}_{B,I}(F) \geq_{tr} \mathbf{u}$)

$(B, I) \models F$ if $\mathcal{H}_{B,I}(F) = \mathbf{t}$

Definition 4 *Given $B \in \mathcal{B}$, the value of the derivation operator $\mathcal{D}_T$ is defined as follows:*

$\mathcal{D}_T(B) = (\{I \mid B, I \models_w T\}, \{I \mid B, I \models T\}).$

Thus, $P(\mathcal{D}_T(B))$ consists of the states which weakly satisfy T, according to B, while $S(\mathcal{D}_T(B))$ are the states which strongly satisfy T according to B. The subscript T is omitted when T is clear from the context.

Example 1 Consider $T = \{K(p) \supset q\}$. Then $\mathcal{D}(\perp) = (\mathcal{A}, \{pq, \overline{p}q\})$. Indeed, $\mathcal{H}_{\perp}(Kp) = \mathbf{u}$. Consequently, for every I, $\mathcal{H}_{\perp,I}(Kp \to q) \neq \mathbf{f}$, that is, $(\perp, I) \models_w Kp \to q$. For the same reason, $\mathcal{H}_{\perp,I}(Kp \to q)) = \mathbf{t}$ if and only if $I(q) = \mathbf{t}$. To compute $\mathcal{D}^2(\perp)$, observe that $\mathcal{H}_{\mathcal{D}(\perp)}(Kp) = \mathbf{f}$. Consequently, for every I, $\mathcal{H}_{\mathcal{D}(\perp),I}(Kp \to q) = \mathbf{t}$. It follows that $\mathcal{D}^2(\perp) = (\mathcal{A}, \mathcal{A})$. It is also easy to see now that $(\mathcal{A}, \mathcal{A})$ is the fixpoint of $\mathcal{D}$. Notice that the belief pair $(\mathcal{A})$, obtained by iterating $\mathcal{D}$, corresponds to the possible-world structure $\mathcal{A}$ that defines the unique stable expansion of T (according to the definition by Moore).

Basic properties of the operator $\mathcal{D}$ are gathered in the following proposition.

Proposition 3 *Let T be a propositionally consistent modal theory. Then, for every belief pair B:*
(1) $\mathcal{D}_T(B)$ is a belief pair.
(2) $\mathcal{D}_T$ is monotone on $\mathcal{B}$.
(3) If B is complete, then $\mathcal{D}_T(B)$ is complete.

Since $(\mathcal{B}, \leq_p)$ is a chain-complete ordered set with least element $\perp$ and $\mathcal{D}$ is monotonic, $\mathcal{D}$ has a least fixpoint. We will denote it by $\mathcal{D} \uparrow$. We propose this fixpoint as the semantics of modal theories reflecting the reasoning process of the agent of gradually constructing belief pairs with increasing information content. In the remainder of the paper, we will study properties of this semantics and, more generally, of fixpoints of the operator $\mathcal{D}$. The next three results relate fixpoints of $\mathcal{D}$ to Moore's semantics.

Theorem 2 *Let $\subseteq \mathcal{L}_K$. A possible-world structure W is an autoepistemic model of T if and only if (W) is a fixpoint of $\mathcal{D}_T$. If $\mathcal{D}_T \uparrow$ is complete then it is the unique autoepistemic model of T.*

Using Propositions 1 and 2, we can extract a relationship between stable expansions and fixpoints of the derivation operator $\mathcal{D}_T$.

Corollary 1 *For any pair T, E of modal theories, E is a consistent stable expansion of T if and only if $E = Th(S)$ for some complete fixpoint (S) of $\mathcal{D}_T$.*
If $\mathcal{H}_{\mathcal{D}_T \uparrow}(F) = \mathbf{t}$ then F belongs to all expansions of T. If $\mathcal{H}_{\mathcal{D}_T \uparrow}(F) = \mathbf{f}$ then F does not belong to any expansion of T.

Consistent stratified autoepistemic theories (Gelfond 1987) have a unique autoepistemic model (stable expansion). Our semantics coincides with the Moore's semantics on stratified theories.

Theorem 3 *If T is a consistent stratified autoepistemic theory, then $\mathcal{D} \uparrow$ is complete. Hence, it is the unique autoepistemic model of T.*

Thus, the semantics defined by the least fixpoint of the operator $\mathcal{D}$ has several attractive features. It is defined for every consistent modal theory T. It coincides with the semantics of autoepistemic logic on stratified theories and, in the general case, provides an approximation to all stable expansions (or, in other words, to skeptical autoepistemic reasoning).

An effective implementation of $\mathcal{D}$

The approach proposed and discussed in the previous section does not directly yield itself to fast implementations. The definition of the operator $\mathcal{D}$ refers to all interpretations of the language $\mathcal{L}$. Thus, computing $\mathcal{D}(B)$ by following the definition is exponential, even for modal theories of a very simple syntactic form. Moreover, representing belief pairs is costly. Each of the sets $P(B)$ and $S(B)$ may contain exponentially many elements.

In this section, we describe a characterization of the operator $\mathcal{D}$ that is much more suitable for investigations of algorithmic issues associated with our semantics. The strategy is to represent a belief set B as a theory $Rep(B)$. Since the theory $Rep(B)$ needs to represent *two* sets of valuations, $Rep(B)$ will be a theory in the propositional language extended by three constants $\mathbf{t}$, $\mathbf{f}$ and $\mathbf{u}$. These constants will always be interpreted by the logical values $\mathbf{t}$, $\mathbf{f}$ and $\mathbf{u}$, respectively. We will call such theories 3-FOL theories.

Let F be a 3-FOL formula. By $(F)^{wk}$ we denote the formula obtained by substituting $\mathbf{t}$ for all positive occurrences of $\mathbf{u}$ and $\mathbf{f}$ for all negative occurrences of $\mathbf{u}$. Similarly, by $(F)^{str}$ we denote the formula obtained by substituting $\mathbf{t}$ for all negative occurrences of $\mathbf{u}$ and $\mathbf{f}$ for all positive occurrences of $\mathbf{u}$. Given a 3-FOL theory Y, we define $(Y)^{str}$ and $(Y)^{wk}$ by standard setwise extension.

Clearly, $(F)^{str}$ and $(F)^{wk}$ do not contain $\mathbf{u}$. Consequently, they can be regarded as formulas in the propositional language extended by two constants $\mathbf{t}$ and $\mathbf{f}$ with standard interpretations as truth and falsity, respectively. We will call this language 2-FOL. We will write $\vdash$ and $\models$ to denote provability and entailment relations in 2-FOL. An important observation here is that if an interpretation satisfies the formula $(F)^{str}$ then it also satisfies $(F)^{wk}$. That is $(F)^{str} \supset (F)^{wk}$ is a tautology of 2-FOL. For a 2-FOL theory U, we define: $Mod(U) = \{I : \text{for all } F \in U, \ I \models F\}$. It follows that for a 3-FOL theory Y, $Mod((Y)^{str}) \subseteq Mod((Y)^{wk})$. Thus, $(Mod((Y)^{wk}), Mod((Y)^{str}))$ is a belief pair and Y can be viewed as its representation.

We show now how, similarly to belief pairs, 3-FOL theories can be used to assign truth values to *modal* atoms (and, hence, to all modal formulas). Let Y be a 3-FOL theory, and let F be a modal formula. Define

$\mathcal{H}_Y(K(F))$ by induction of depth of formula F as follows:

(1) If F is objective, then define:

$$\mathcal{H}_Y(K(F)) = \begin{cases} \mathbf{t} & \text{if } (Y)^{wk} \vdash (F)^{str} \\ \mathbf{f} & \text{if } (Y)^{str} \not\vdash (F)^{wk} \\ \mathbf{u} & \text{otherwise.} \end{cases}$$

(2) If F is not objective, then replace all modal atoms $K(G)$ in F by $\mathcal{H}_Y(K(G))$. This yields an objective formula F'. Define $\mathcal{H}_Y(K(F)) = \mathcal{H}_Y(K(F'))$.

Let T be a modal theory and let Y be a 3-FOL theory. By the Y-instance of T, T_Y, we mean a 3-FOL theory obtained by substituting all modal literals $K(F)$ (not appearing under the scope of any other occurrence of K) by $\mathcal{H}_Y(K(F))$. Observe that for a finite modal theory T and a finite 3-FOL theory Y, T_Y can be computed by means of polynomially many calls to the propositional provability procedure.

Let T be a modal theory. We will now define a counterpart to the operator $\mathcal{D}_T$. Let Y be a 3-FOL theory. Define $\mathcal{SD}_T(Y) = T_Y$.

The key property of the operator $\mathcal{SD}_T$ is that, for a finite modal theory T and for a finite 3-FOL theory Y, $\mathcal{SD}_T(Y)$ can be computed by means of polynomially many calls to the propositional provability procedure.

We will show that $\mathcal{SD}_T$ can be used to compute $\mathcal{D}_T$. In particular, we will show that the least fixpoint of $\mathcal{D}_T$ can be computed by iterating the operator $\mathcal{SD}_T$. To this end, for every 3-FOL theory Y, define $\mathcal{B}el(Y) = (Mod((T)^{wk}), Mod((T)^{str}))$.

First, the following theorem shows that the truth values of modal atoms evaluated according to a 3-FOL theory T and according to the corresponding belief pair $Bel(T)$ coincide.

Theorem 4 *Let Y be a 3-FOL theory. Then, for every modal formula F,*

$$\mathcal{H}_{Bel(Y)}(K(F)) = \mathcal{H}_Y(K(F)).$$

Next, let us observe that the operator $\mathcal{D}$ can be described in terms of the operator Bel. Let T be a modal theory and let B be a belief pair. By the B-instance of T, T_B, we mean a 3-FOL theory obtained by substituting all modal literals $K(F)$ (not appearing under the scope of any other occurrence of K) by $\mathcal{H}_B(K(F))$.

Theorem 5 *Let T be a modal theory and let B be a belief pair. Then, $\mathcal{D}_T(B) = \mathcal{B}el(T_B)$*

This theorem indicates that, given a modal theory T, belief pairs that are in the range of the operator $\mathcal{D}_T$ can be represented by objects of size polynomial in the size of T. Namely, every belief pair of the form $\mathcal{D}_T(B)$ can be represented by a 3-FOL theory T_B.

Theorems 4 and 5 imply the main result of this section.

Theorem 6 *Let T be a modal theory and let Y be a 3-FOL theory.*
(1) $\mathcal{B}el(\mathcal{SD}_T(Y)) = \mathcal{D}_T(\mathcal{B}el(Y))$.
(2) If a belief pair B is a fixpoint of $\mathcal{D}_T$, then T_B is a fixpoint of $\mathcal{SD}_T$.
(3) If Y is a fixpoint of $\mathcal{SD}_T$ then $\mathcal{B}el(Y)$ is a fixpoint of $\mathcal{D}_T$.

Observe that $\mathcal{B}el(\{\mathbf{u}\}) = \perp$. It follows directly from Theorem 6 (by induction) that for every ordinal number α, $\mathcal{D}_T^\alpha(\perp) = \mathcal{B}el(\mathcal{SD}_T^\alpha(\{\mathbf{u}\}))$.

Clearly, if $\mathcal{SD}_T^\alpha(\perp) = \mathcal{SD}_T^{\alpha+1}(\perp)$ then $\mathcal{D}_T^\alpha(\perp) = \mathcal{D}_T^{\alpha+1}(\perp)$. Moreover, by Theorems 4, 5 and 6, and by induction, it is easy to show that if $\mathcal{D}_T^\alpha(\perp) = \mathcal{D}_T^{\alpha+1}(\perp)$ then $\mathcal{SD}_T^{\alpha+1}(\perp) = \mathcal{SD}_T^{\alpha+2}(\perp)$.

Consequently, the least fixpoint of $\mathcal{D}_T$ (its polynomial-size representation) can be computed by iterating the operator $\mathcal{SD}_T$. In the case when T is finite, the number of iterations is limited by the number of top level (unnested) modal literals in T. Originally, they may all be evaluated to $\mathbf{u}$. However, at each step, at least one $\mathbf{u}$ changes to either $\mathbf{t}$ or $\mathbf{f}$ and this value is preserved in the subsequent evaluations. Thus, the problem of computing a polynomial size representation of the least fixpoint of the operator $\mathcal{D}$, the corresponding 3-FOL theory, is in the class Δ_2^P.

Relationship to Logic Programming

Autoepistemic logic is closely related to several semantics for logic programs with negation. It is well-known that both stable and supported models of logic programs can be described as expansions of appropriate translations of programs into modal theories (see, for instance, (Marek & Truszczyński 1993)). In this section, we briefly discuss connections of the semantics defined by the least point of the operator $\mathcal{D}$ to some 3-valued semantics of logic programs. The details will be provided in a forthcoming work.

Given a clause $r = a \leftarrow b_1, \ldots, b_k, c_1, \ldots, c_m$, define:

$$ael_1(r) = b_1 \wedge \ldots \wedge b_k \wedge \neg K c_1 \wedge \ldots \wedge \neg K c_m \supset a$$

and

$$ael_2(r) = K b_1 \wedge \ldots \wedge K b_k \wedge \neg K c_1 \wedge \ldots \wedge \neg K c_m \supset a$$

Embeddings $ael_1()$ and $ael_2()$ can be extended to logic programs P.

Let B be a belief pair. Define the *projection Proj(B)* as the 3-valued interpretation I such that $I(p) = \mathcal{H}_B(K(p))$.

It turns out that fixpoints of the operator $\mathcal{D}_{ael_1(P)}$ ($\mathcal{D}_{ael_2(P)}$, respectively) precisely correspond to 3-valued supported (stable, respectively) models of

P (the projection function $Proj(\cdot)$ establishes the correspondence). Moreover, complete fixpoints of $\mathcal{D}_{ael_1(P)}$ describe 2-valued supported (stable, respectively) models of P. Finally, the least fixpoint of $\mathcal{D}_{ael_2(P)}$ captures the Fitting-Kunen 3-valued semantics of a program P, and the least fixpoint of $\mathcal{D}_{ael_1(P)}$ captures the well-founded semantics of P.

References

Chen, W.; Swift, T.; and Warren, D. 1994. Efficient top-down computation of queries under the well-founded semantics. *Journal of Logic Programming.*

Fitting, M. C. 1985. Kripke-Kleene semantics for logic programs. *Journal of Logic Programming* 2:295–312.

Gelfond, M. 1987. On stratified autoepistemic theories. In *Proceedings of AAAI-87*, 207–211. Los Altos, CA: American Association for Artificial Intelligence.

Gottlob, G. 1992. Complexity results for nonmonotonic logics. *Journal of Logic and Computation* 2:397–425.

Gottlob, G. 1995. Translating default logic into standard autoepistemic logic. *Journal of the ACM* 42:711–740.

Kunen, K. 1987. Negation in Logic Programming. *Journal of Logic Programming* 4(3):289–308.

Levesque, H. J. 1990. All I know: a study in autoepistemic logic. *Artificial Intelligence* 42:263–309.

Marek, W., and Truszczyński, M. 1993. *Nonmonotonic logics; context-dependent reasoning*. Berlin: Springer-Verlag.

Markowsky, G. 1976. Chain-complete posets and directed sets with applications. *Algebra Universalis* 6:53–68.

Moore, R. 1984. Possible-world semantics for autoepistemic logic. In Reiter, R., ed., *Proceedings of the workshop on non-monotonic reasoning*, 344–354. (Reprinted in: M.Ginsberg, editor, *Readings on nonmonotonic reasoning*. pages 137–142, 1990, Morgan Kaufmann.).

Moore, R. 1985. Semantical considerations on nonmonotonic logic. *Artificial Intelligence* 25:75–94.

Niemelä, I., and Simons, P. 1995. Evaluating an algorithm for default reasoning. In *Proceedings of the IJCAI-95 Workshop on Applications and Implementations of Non-monotomic Reasonigs Systems.*

Niemelä, I. 1992. On the decidability and complexity of autoepistemic reasoning. *Fundamenta Informaticae* 17:117–155.

Reiter, R. 1980. A logic for default reasoning. *Artificial Intelligence* 13:81–132.

Schwarz, G. 1992. Minimal model semantics for nonmonotonic modal logics. In *Proceedings of LICS-92.*

Van Gelder, A.; Ross, K.; and Schlipf, J. 1991. Unfounded sets and well-founded semantics for general logic programs. *Journal of the ACM* 38:620 – 650.