Fixpoint 3-valued semantics for autoepistemic logic

Marc Denecker[1], V. Wiktor Marek[2], and Mirosław Truszczyński[2]

[1] Department of Computer Science, K.U.Leeuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium
[2] Department of Computer Science, University of Kentucky, Lexington, KY 40506-0046, USA

Dedicated to Ray Reiter on his 60th birthday

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 both well-founded semantics and 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).

1 Introduction

We describe a 3-valued semantics for modal theories that approximates skeptical mode of reasoning in the autoepistemic logic introduced in [12,13]. 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 introduced by Reiter in [17]. It can handle default reasonings under a simple and modular translation in the case of prerequisite-free defaults [10]. 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 [5]. Further, under the so called Gelfond translation, autoepistemic logic captures the semantics of stable models for logic programs [3]. Under the Konolige encoding [6] of logic programs as modal theories, stable expansions generalize the concept of the supported model semantics [10].

Autoepistemic logic is also known to be equivalent to several other modal nonmonotonic reasoning systems including the *only-knowing* logic of Levesque [8] and the reflexive autoepistemic logic of Schwarz [18].

The semantics for autoepistemic logic [13] 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, consistent stable 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 [4,14], 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; this decision problem is Π_2^P-complete [4].

We propose an alternative semantics for autoepistemic reasoning that, in particular, allows us to *approximate* the skeptical approach described above (as well as the dual, *brave* mode of reasoning). 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 under our semantics is in the class Δ_2^P.

As mentioned above, the semantics we propose can be applied to approximate the skeptical mode of autoepistemic reasoning. However, it has also another important application. It can be used as a pruning mechanism in algorithms that compute expansions. While searching for expansions, one can compute our 3-valued semantics for a modal theory under consideration (as mentioned, it is a simpler task computationally than the task of computing an expansion). Formulas true under this semantics are guaranteed to belong to all expansions and those that are false belong to none. This information can be used to simplify the current theory and limit the search space. As a consequence, significant speedups may be achieved.

There are parallels between our semantics and the well-founded semantics in logic programming. The well-founded semantics approximates the stable model semantics (atoms true under the well-founded semantics are in all stable models and atoms that are false under the well-founded semantics belong to none). Moreover, computing well-founded semantics is polynomial while deciding whether an atom belongs to all stable models is a co-NP-complete problem. As a result, the well-founded semantics is used as a search space pruning mechanism by some algorithms to compute stable model semantics [15]. We will show in the paper that there is, indeed, a close formal connection between our 3-valued semantics of modal theories and the well-founded semantics of logic programs.

The 3-valued semantics for autoepistemic logic introduced in this paper is based on the notion of a *belief pair*, that is, a pair (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 [12]. Moore characterized expansions in terms of *possible-world structures*, that is, sets of 2-valued interpretations. A belief pair (P, S) 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. It turns out that while expansions (or the corresponding possible-world structures) do not contain meta-ignorance, belief pairs, in general, do.

There is a natural ordering of belief pairs. We say that (P_1, S_1) "better approximates" than (P, S) the agent's beliefs entailed by the agent's initial assumptions if $S \subseteq S_1 \subseteq P_1 \subseteq P$. We will denote the corresponding ordering relation in the set $\mathcal{B}$ of all belief pairs by $\leq_p$. Our semantics of modal theories is defined in terms of an operator on the set of belief pairs. This operator, $\mathcal{D}_T$, is determined by a modal theory T (the set of initial assumptions of the agent). Intuitively, it attempts to simulate a constructive process a rational agent might use to produce an "elementary" improvement on this agent's current set of beliefs and disbeliefs: given a belief pair $B = (P, S)$,

$\mathcal{D}_T(B)$ is a belief pair that provides another, under some assumptions better, approximation to the agent's beliefs.

An important property is that $\mathcal{D}_T$ is monotone with respect to $\leq_p$. Hence, it has the least fixpoint. This least fixpoint can be constructed by starting with the least informative belief pair (approximating every possible-world structure) and then iterating the operator $\mathcal{D}_T$, in each step improving on the previous belief pair until no further improvement is possible. We propose this fixpoint as a *constructive* approximation to the semantics of stable expansions.

A fundamental property that makes the above approach meaningful is that *complete* belief pairs (those with P equal to S) that are fixpoints of $\mathcal{D}_T$ are (under an obvious one-to-one correspondence) precisely Moore's autoepistemic 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 and brave reasoning based on expansions. Moreover, as mentioned above, the problem of computing the least fixpoint of the operator $\mathcal{D}_T$ 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 [20] and the 3-valued Fitting-Kunen semantics [2,7].

Our paper is structured as follows. The next section reviews the basics of autoepistemic logic including both syntactic and semantic definitions of expansions. We then investigate the properties of the partial ordering of belief pairs and study the operator $\mathcal{D}_T$. Subsequently, we show how the purely semantic approach can be described in proof-theoretic terms and use this proof-theoretic approach to study algorithmic issues of the least fixpoint of the operator $\mathcal{D}_T$. Next, we discuss connections between fixpoints of $\mathcal{D}_T$ and several semantics of logic programs with negation. Section 6 contains conclusions and a discussion of future work. The appendix that concludes the paper gives a proof of Theorem 6.

2 Autoepistemic logic — preliminaries

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

The notion of a 2-*valued interpretation* of the language $\mathcal{L}$ is defined as usual: it is a mapping from At to $\{\mathbf{t}, \mathbf{f}\}$. Throughout the paper $\mathcal{A}_{At}$ (or $\mathcal{A}$, if At is clear from the context) will always denote the set of all interpretations of the set At of atoms of $\mathcal{L}$.

Autoepistemic logic was first introduced by Moore in [12] and later studied in [13]. In [13], the semantics of an autoepistemic theory T is defined in terms

of *stable expansions*. For every two sets T and E of modal formulas, E is said to be a *stable expansion* of T if it satisfies the equation:

$$E = \{\varphi: T \cup \{\neg K\psi: \psi \notin E\} \cup \{K\psi: \psi \in E\} \models_{FOL} \varphi\}$$

(the symbol $\models_{FOL}$ stands for classical entailment, where all formulas $K\varphi$ are interpreted as propositional literals).

A possible-world treatment of autoepistemic logic was described by Moore [12]. A possible-world structure W (over At) is a set of 2-valued interpretations of At. Alternatively, it can be seen as a Kripke structure with a total accessibility relation. Given a pair (W, I), where W is a possible-world structure and I is an interpretation (not necessarily from W), one defines a truth assignment function $\mathcal{H}_{W,I}$ inductively as follows:

i. For an atom A, we define $\mathcal{H}_{W,I}(A) = I(A)$
ii. The boolean connectives are handled in the usual way
iii. For every 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 $(W, I) \models_{ael} F$ to denote that $\mathcal{H}_{W,I}(F) = \mathbf{t}$. Further, for a modal theory T, we write $(W, I) \models_{ael} T$ if $\mathcal{H}_{W,I}(F) = \mathbf{t}$ for every $F \in T$. Finally, for a possible world structure W we define the *theory* of W, $Th(W)$, by: $Th(W) = \{F: (W, I) \models_{ael} F, \text{ for all } I \in W\}$.

It is well known that for every formula F, either $KF \in Th(W)$ or $\neg KF \in Th(W)$ (since $\mathcal{H}_{W,I}(KF)$ is the same for all interpretations $I \in \mathcal{A}$). 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 [12,10].

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

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

The following theorem, relating stable expansions of [13] and autoepistemic models, was proved in [8] and was discussed in detail in [19].

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

[1] Observe that empty models are allowed. This assumption allows us to treat consistent and inconsistent expansions in a uniform manner.

3 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 $P, S \subseteq At$ such that $S \subseteq P$. When $B = (P, S)$, $S(B)$ denotes S and $P(B)$ denotes P. The belief pair $(\mathcal{A}, \emptyset)$ is denoted $\bot$. The set $\{(P, S)\colon P, S \subseteq \mathcal{A} \text{ and } P \supseteq S\}$ of all belief pairs is denoted by $\mathcal{B}$. The belief pair $(\emptyset, \emptyset)$ is called *inconsistent* and is denoted by $\top$.

A belief pair B can be seen as an approximation of a possible-world structure W such that $S(B) \subseteq W \subseteq P(B)$. The interpretations in $S(B)$ can be viewed as states of the world which are known to be possible (belong to W). The set of these interpretations forms a lower approximation to W. The set $P(B)$ of interpretations can be viewed as an upper approximation to W: interpretations not in $P(B)$ are known not to be in W.

We will now extend the concept of an interpretation to the case of belief pairs and consider the question of meta-ignorance and meta-knowledge of belief pairs. We will see that, being only approximations to possible-world structures, belief pairs may contain meta-ignorance. We will use three logical values, **f**, **u** and **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 (min and max are evaluated with respect to the ordering $\leq_{tr}$):

(a) $\mathcal{H}_{B,I}(A) = I(A)$, if A is an atom
(b) $\mathcal{H}_{B,I}(\neg F) = \mathcal{H}_{B,I}(F)^{-1}$
(c) $\mathcal{H}_{B,I}(F_1 \vee F_2) = \max\{\mathcal{H}_{B,I}(F_1), \mathcal{H}_{B,I}(F_2)\}$
(d) $\mathcal{H}_{B,I}(F_1 \wedge F_2) = \min\{\mathcal{H}_{B,I}(F_1), \mathcal{H}_{B,I}(F_2)\}$
(e) $\mathcal{H}_{B,I}(F_2 \supset F_1) = \max\{\mathcal{H}_{B,I}(F_1), \mathcal{H}_{B,I}(F_2)^{-1}\}$

The formula KF is evaluated as follows:

$$\mathcal{H}_{B,I}(KF) = \begin{cases} \mathbf{t} & \text{if for every } J \in P,\ \mathcal{H}_{B,J}(F) = \mathbf{t} \\ \mathbf{f} & \text{if there is } J \in S \text{ such that } \mathcal{H}_{B,J}(F) = \mathbf{f} \\ \mathbf{u} & \text{otherwise} \end{cases}$$

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 $F \in \mathcal{L}_K$, $\mathcal{H}_{B,I}(F) = \mathcal{H}_{W,I}(F)$.*

We will now define two satisfaction relations: *weak*, denoted by $\models_w$, and *strong*, denoted by $\models$. Namely, for a belief pair B, an interpretation I and a modal formula F we define:

i. $(B, I) \models_w F$ if $\mathcal{H}_{B,I}(F) \neq \mathbf{f}$ (that is, if $\mathcal{H}_{B,I}(F) \geq_{tr} \mathbf{u}$), and
ii. $(B, I) \models F$ if $\mathcal{H}_{B,I}(F) = \mathbf{t}$

Let T be a modal theory and let B be a belief pair. We define:

$$\mathcal{D}_T(B) = (\{I : (B, I) \models_w T\}, \{I : (B, I) \models T\}). \tag{1}$$

Clearly, if $(B, I) \models F$ then $(B, I) \models_w F$. Hence, $\mathcal{D}_T(B)$ is a belief pair or, in other words, $\mathcal{D}_T$ is an operator on $(\mathcal{B}, \leq_p)$. In addition, $P(\mathcal{D}_T(B))$ consists of the interpretations which weakly satisfy T according to B, while $S(\mathcal{D}_T(B))$ consists of those interpretations which strongly satisfy T according to B. The subscript T in $\mathcal{D}_T$ is often omitted when T is clear from the context.

Example 1. Consider $T = \{Kp \supset q\}$. Then $\mathcal{D}(\bot) = (\mathcal{A}, \{pq, \overline{p}q\})$ (here, by pq we mean an interpretation that assigns $\mathbf{t}$ to both p and q while $\overline{p}q$ denotes an interpretation assigning $\mathbf{f}$ to p and $\mathbf{t}$ to q). Indeed, $\mathcal{H}_\bot(Kp) = \mathbf{u}$. Consequently, for every I, $\mathcal{H}_{\bot,I}(Kp \supset q) \neq \mathbf{f}$, that is, $(\bot, I) \models_w Kp \supset q$. For the same reason, $\mathcal{H}_{\bot,I}(Kp \supset q)) = \mathbf{t}$ if and only if $I(q) = \mathbf{t}$.

To compute $\mathcal{D}^2(\bot)$, observe that $\mathcal{H}_{\mathcal{D}(\bot)}(Kp) = \mathbf{f}$. Consequently, for every I, $\mathcal{H}_{\mathcal{D}(\bot),I}(Kp \supset q) = \mathbf{t}$. It follows that $\mathcal{D}^2(\bot) = (\mathcal{A}, \mathcal{A})$. It is also easy to see now that $(\mathcal{A}, \mathcal{A})$ is the fixpoint of $\mathcal{D}$, that is, $\mathcal{D}(\mathcal{A}, \mathcal{A}) = (\mathcal{A}, \mathcal{A})$.

The next result relates complete fixpoints of $\mathcal{D}$ to Moore's semantics of autoepistemic logic.

Theorem 2. *Let $T \subseteq \mathcal{L}_K$. Then:*

(a) *For every $W \subseteq \mathcal{A}$, (W) is a fixpoint of $\mathcal{D}_T$ if and only if $W = \{I : (W, I) \models_{ael} T\}$*
(b) *A possible-world structure W is an autoepistemic model of T if and only if (W) is a fixpoint of $\mathcal{D}_T$*

(c) *A modal theory E is a stable expansion of T if and only if $E = Th(S)$ for some complete fixpoint (S) of $\mathcal{D}_T$*

Proof: (a) Observe that for every $W \subseteq \mathcal{A}$, for every $I \in \mathcal{A}$ and for every $F \in T$, $\mathcal{H}_{(W),I}(F) = \mathcal{H}_{W,I}(F)$. Hence, $((W), I) \models F$ if and only if $(W, I) \models_{ael} F$. In addition, by Proposition 1, $(W, I) \models F$ if and only if $((W), I) \models F$. Thus, (W) is a fixpoint of $\mathcal{D}_T$ if and only if $W = \{I\colon (W, I) \models_{ael} T\}$. The assertion (b) follows directly from (a). The assertion (c) follows from (b) by Theorem 1. □

Theorem 2 demonstrates that complete fixpoints of the operator $\mathcal{D}_T$ describe stable expansions of T. However, in general, the operator $\mathcal{D}_T$ may also have fixpoints that are not complete. Such fixpoints provide 3-valued interpretations to modal formulas and can serve as approximations to complete fixpoints of $\mathcal{D}_T$.

The approach to autoepistemic reasoning that we present in this paper exploits the concept of a *least* fixpoint of $\mathcal{D}_T$. Namely, we show the existence of this least fixpoint and demonstrate that it can be constructed by iterating the operator $\mathcal{D}_T$ starting with the belief pair $\bot$. Intuitively, this iterative construction models the agent who, given an initial theory T, starts with the belief pair $\bot$ (with the smallest meta-knowledge content) and, then, iteratively constructs a sequence of belief pairs with increasing meta-knowledge (decreasing meta-ignorance) until no further improvement is possible.

Next, we demonstrate that the semantics implied by the least fixpoint of $\mathcal{D}_T$ approximates the semantics of Moore and that it coincides with the semantics of Moore on stratified modal theories. We show that the task to compute the least fixpoint of the operator $\mathcal{D}_T$ is simpler than computing autoepistemic expansions (unless the polynomial hierarchy collapses). Finally, we study connections of our semantics to several semantics used for logic programs with negation.

Our approach relies on an observation that there is a natural partial ordering of the set $\mathcal{B}$ of belief pairs. Recall that for two belief pairs B_1 and B_2, we defined

$$B_1 \leq_p B_2 \quad \text{if} \quad P(B_1) \supseteq P(B_2) \quad \text{and} \quad S(B_1) \subseteq S(B_2). \tag{2}$$

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). Clearly, the relation $\leq_p$ is reflexive, antisymmetric and transitive. Hence, $(\mathcal{B}, \leq_p)$ is a poset. The following two results gather some basic properties of the poset $(\mathcal{B}, \leq_p)$, truth assignment function $\mathcal{H}_{B,I}$ and the operator $\mathcal{D}$. The first one shows that the ordering $\leq_p$ is consistent with the concept of the *knowledge ordering* (also referred to as *information* ordering in the literature) of the truth values: $\mathbf{u} \leq_{kn} \mathbf{f}, \mathbf{u} \leq_{kn} \mathbf{t}$, $\mathbf{f} \not\leq_{kn} \mathbf{t}$ and $\mathbf{t} \not\leq_{kn} \mathbf{f}$. It also relates the ordering $\leq_p$ to the weak and strong entailment relations $\models_w$ and $\models$. The second result states that $\mathcal{D}$ is monotone with respect to $\leq_p$.

Proposition 2. *Let B_1 and B_2 be belief pairs such that $B_1 \leq_p B_2$. For every interpretation $I \in \mathcal{A}$ and every formula $F \in \mathcal{L}_K$:*

(a) $\mathcal{H}_{B_1,I}(F) \leq_{kn} \mathcal{H}_{B_2,I}(F)$.
(b) *If* $(B_2, I) \models_w F$ *then* $(B_1, I) \models_w F$
(c) *If* $(B_1, I) \models F$, *then* $(B_2, I) \models F$.

Proof: (a) We proceed by induction on the length of F. Thus, let us consider a modal formula F and let us assume that the assertion of the proposition holds for every modal formula G of length smaller than the length of F. There are three cases to consider.

First, assume that F is an atom. Then for every $I \in \mathcal{A}$, $\mathcal{H}_{B_1,I}(F) = I(F) = \mathcal{H}_{B_2,I}(F)$. In particular, $\mathcal{H}_{B_1,I}(F) \leq_{kn} \mathcal{H}_{B_2,I}(F)$ (this argument establishes the basis for the induction).

Next, assume that F is of the form $G \wedge G'$, $G \vee G'$, $G \supset G'$ or $\neg G$. In this case, the assertion follows immediately from the induction hypothesis and from the following observation: if a, b, a' and b' are truth values such that $a \leq_{kn} a'$ and $b \leq_{kn} b'$ then:

i. $(a \wedge b) \leq_{kn} (a' \wedge b')$
ii. $(a \vee b) \leq_{kn} (a' \vee b')$
iii. $(a \supset b) \leq_{kn} (a' \supset b')$
iv. $(\neg a) \leq_{kn} \neg(a')$.

Finally, let us assume $F = KG$ for some modal formula G. Take any $I \in \mathcal{A}$. Assume that $\mathcal{H}_{B_1,I}(KG) = \mathbf{t}$. It follows that for every $J \in P(B_1)$, $\mathcal{H}_{B_1,J}(G) = \mathbf{t}$. Since $B_1 \leq_{kn} B_2$, $P(B_2) \subseteq P(B_1)$. Hence, by the induction hypothesis, for every $J \in P(B_2)$, $\mathcal{H}_{B_2,J}(G) = \mathbf{t}$. Consequently, $\mathcal{H}_{B_2,I}(KG) = \mathbf{t}$ and $\mathcal{H}_{B_1,I}(F) \leq_{kn} \mathcal{H}_{B_2,I}(F)$.

The argument in the case when $\mathcal{H}_{B_1,I}(KG) = \mathbf{f}$ is similar. Since $\mathbf{u} \leq_{kn} \mathbf{t}$ and $\mathbf{u} \leq_{kn} \mathbf{f}$, the assertion follows in the case when $\mathcal{H}_{B_1,I}(KG) = \mathbf{u}$, too.
(b) Assume that $(B_1, I) \not\models_w T$. Then, there is a formula $F \in T$ such that $\mathcal{H}_{B_1,I}(F) = \mathbf{f}$. By the assertion (a), $\mathcal{H}_{B_2,I}(F) = \mathbf{f}$. Consequently, $(B_2, I) \not\models_w T$.
(c)Assume that $(B_1, I) \models T$. Then $\mathcal{H}_{B_1,I}(F) = \mathbf{t}$ for every $F \in T$. By the assertion (a), $\mathcal{H}_{B_2,I}(F) = \mathbf{t}$ for every $F \in T$. Hence, $(B_2, I) \models T$. □

Proposition 3. *Let B_1 and B_2 be belief pairs such that $B_1 \leq_p B_2$. For every theory $T \subseteq \mathcal{L}_K$, $\mathcal{D}_T(B_1) \leq_p \mathcal{D}_T(B_2)$, that is, the operator $\mathcal{D}_T$ is monotone on $(\mathcal{B}, \leq_p)$.*

Proof: Assume that $B_1 \leq_p B_2$. By Proposition 2(b), $\{I\colon (B_2, I) \models_w T\} \subseteq \{I\colon (B_1, I) \models_w T\}$. Hence, $P(\mathcal{D}(B_2)) \subseteq P(\mathcal{D}(B_1))$. Similarly (by Proposition 2(c)), $S(\mathcal{D}(B_1)) \subseteq S(\mathcal{D}(B_2))$. Thus, $\mathcal{D}(B_1) \leq_p \mathcal{D}(B_2)$. □

Proposition 3 is especially important. The monotonicity of the operator $\mathcal{D}$ will allow us to assert the existence of a least fixpoint of $\mathcal{D}$. However, let us note that the poset $(\mathcal{B}, \leq_p)$ is not a lattice (and, hence, not a complete

lattice). Indeed, 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 $(\mathcal{B}, \leq_p)$. Thus, we will not be able to use the theorem Tarski-Knaster in its classic form. Instead, we will use its generalization (see [11]) developed for the case of posets that are *chain complete*. Let us recall that a poset is chain complete if its every *chain* (that is, a totally ordered subposet) has a l.u.b. [1,11]. Note also that every chain complete partially ordered set has a least element. It follows from the observation that the empty set is a chain.

Theorem 3 ([11]). *Let $(P, \leq)$ be a chain-complete poset. Let D be a monotone operator on $(P, \leq)$. Then, D has a least fixpoint. This fixpoint is the limit of the sequence of iterations of D starting with the least element of $(P, \leq)$.*

To use Theorem 3, we will now show that the poset $(\mathcal{B}, \leq_p)$ is chain complete.

Proposition 4. *The poset $(\mathcal{B}, \leq_p)$ is chain complete.*

Proof: For a nonempty set C of belief pairs define $P_C = \bigcap\{P(B) : B \in C\}$ and $S_C = \bigcup\{S(B) : B \in C\}$. Consider now a chain $C \subseteq \mathcal{B}$ of belief pairs. Assume that $I \in S_C$. There exists a belief pair $(P, S) \in C$ such that $I \in S$. Since (P, S) is a belief pair, $I \in P$. Let $(P', S') \in C$. Then we have $(P', S') \leq_p (P, S)$ or $(P, S) \leq_p (P', S')$. In the first case, $P \subseteq P'$. Hence, $I \in P'$. In the second case we have $S \subseteq S' \subseteq P'$ and, again, $I \in P'$. It follows that $I \in P_C$ and, consequently, that $S_C \subseteq P_C$.

We have just proved that (P_C, S_C) is a belief pair. It is easy to see that for every $(P, S) \in C$, $(P, S) \leq_p (P_C, S_C)$. Moreover, any other upper bound B of C satisfies $(P_C, S_C) \leq_p B$. Hence, (P_C, S_C) is the l.u.b. of C.

Finally, it is evident that the belief pair $\bot = (\mathcal{A}, \emptyset)$ is a least element of the poset $(\mathcal{B}, \leq_p)$. Thus the empty chain also has its least upper bound (the least element of $(P, \leq)$). □

As an immediate consequence of Theorem 3 and Proposition 4 we obtain the following crucial corollary.

Corollary 1. *For every theory $T \subseteq \mathcal{L}_K$, the operator $\mathcal{D}_T$ has a least fixpoint.*

The least fixpoint of the operator $\mathcal{D}_T$ will be denoted by $\mathcal{D}_T\uparrow$. We propose this fixpoint as the semantics of modal theory T. This semantics reflects the reasoning process of an agent who gradually constructs belief pairs with increasing knowledge (information) content.

The following three results provide justification for our least fixpoint semantics. The first of these results shows that the least fixpoint semantics provides a lower approximation to the skeptical semantics based on expansions and an upper approximation to the brave reasoning based on expansions.

Theorem 4. *Let T be a modal theory. If $\mathcal{H}_{\mathcal{D}_T\uparrow}(KF) = \mathbf{t}$ then F belongs to all expansions of T. If $\mathcal{H}_{\mathcal{D}_T\uparrow}(KF) = \mathbf{f}$ then F does not belong to any expansion of T.*

Proof: Assume that $\mathcal{H}_{\mathcal{D}_T\uparrow}(KF) = \mathbf{t}$. Let E be a stable expansion of T. Then $E = Th(W)$ for some autoepistemic model W of T (Theorem 1). Clearly, (W) is then a fixpoint of $\mathcal{D}_T$ (Theorem 2). Since $\mathcal{D}_T\uparrow \leq_p (W)$, by Proposition 2 it follows that $\mathcal{H}_{(W)}(KF) = \mathbf{t}$. Hence, for every $I \in W$, $\mathcal{H}_{(W),I}(F) = \mathbf{t}$ or, equivalently (Proposition 1), $\mathcal{H}_{W,I}(F) = \mathbf{t}$. Consequently, $F \in Th(W) = E$. A similar argument can be used to prove the second part of the assertion. □

The second result shows that if the least fixpoint is complete (that is, no further improvement in meta-knowledge is possible) than the least fixpoint semantics coincides with the semantics of Moore.

Theorem 5. *If $\mathcal{D}_T\uparrow$ is complete then $\mathcal{D}_T\uparrow$ is the unique autoepistemic model of T.*

Proof: For every complete fixpoint (W) of $\mathcal{D}_T$, $\mathcal{D}_T\uparrow \leq_p (W)$. Moreover, complete elements of $(\mathcal{B}, \leq_p)$ are maximal. Hence, if $\mathcal{D}_T\uparrow$ is complete, it is a unique complete fixpoint of $\mathcal{D}_T$. Thus, by Theorem 2(b), $\mathcal{D}_T\uparrow$ is the unique autoepistemic model of T. □

In the last result of this section we will show that the least fixpoint semantics is complete for the class of stratified theories, introduced by Gelfond [3] and further generalized in [9]. This property, in combination with Theorem 5, implies that for stratified theories the least fixpoint semantics coincides with the skeptical (and brave) autoepistemic semantics of Moore. This is an important property since the semantics of Moore is commonly accepted for the class of stratified theories and the agreement with this semantics is regarded as a test of "correctness" of a semantics for a modal nonmonotonic logic. Let us note that a similar test of agreement with the perfect model semantics on stratified programs is used in logic programming to justify semantics for logic programs with negation. In particular, the well-founded and stable model semantics both coincide with the perfect model semantics on stratified logic programs. This property is not quite coincidental as connections between autoepistemic logic and logic programming are well known [10] and are also discussed below in Section 5.

Theorem 6. *If T is a stratified autoepistemic theory then:*

(a) *$\mathcal{D}_T\uparrow$ is complete*
(b) *T has a unique stable expansion*
(c) *$\mathcal{D}_T\uparrow$ is consistent if and only if the lowest stratum T_0 is consistent.*

The proof of this theorem (as well as a precise definition of a stratified modal theory) can be found in the appendix.

To conclude this section let us observe that the semantics defined by the least fixpoint of the operator $\mathcal{D}$ has several attractive features. It is defined for every modal theory T. It coincides with the semantics of autoepistemic logic on stratified theories. In the general case, it provides a lower approximation to the intersection of all stable expansions (skeptical autoepistemic reasoning) and upper approximation to the union of all stable expansions (brave autoepistemic reasoning).

4 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.

To this end, in addition to the propositional language $\mathcal{L}$ (generated, recall, by the set of atoms At), we will also consider the extension of $\mathcal{L}$ by three new constants **t**, **f** and **u**. We will call this language *3-FOL*. Formulas and theories in this language will be called *3-FOL formulas* and *3-FOL theories*, respectively. Our strategy is now as follows. First, we will show that a wide class of belief pairs can be represented by 3-FOL theories. Next, using this representation, we will describe a method to compute fixpoints of the operator $\mathcal{D}$ that is algorithmically more feasible than the direct approach implied by the definition of $\mathcal{D}$.

We start by discussing a class of 3-valued truth assignments on the language 3-FOL that are generated by 2-valued interpretations from $\mathcal{A}$ under the assumption that the new constants **t**, **f** and **u** are always interpreted by the logical values they represent. Formally, given an interpretation $I \in \mathcal{A}$, we define a valuation I^e on the language 3-FOL inductively as follows (minima and maxima are computed with respect to the truth ordering of **t**, **f** and **u**):

i. $I^e(A) = I(A)$, if $A \in At$
ii. $I^e(a) = a$, for $a \in \{\mathbf{t}, \mathbf{f}, \mathbf{u}\}$
iii. $I^e(\neg F) = (I^e(F))^{-1}$
iv. $I^e(F_1 \vee F_2) = \max\{I^e(F_1), I^e(F_2)\}$
v. $I^e(F_1 \wedge F_2) = \min\{I^e(F_1), I^e(F_2)\}$
vi. $I^e(F_2 \supset F_1) = \max\{I^e(F_1), (I^e(F_2))^{-1}\}$.

Let F be a 3-FOL formula. By F^{wk} we denote the formula obtained by substituting **t** for all positive occurrences of **u** and **f** for all negative occurrences of **u**. Similarly, by F^{str} we denote the formula obtained by substituting **t** for all negative occurrences of **u** and **f** for all positive occurrences of **u**. Given a 3-FOL theory Y, we define Y^{str} and Y^{wk} by the standard setwise extension. Before we proceed let us note the following useful identities (the proof is straightforward and is omitted):

$$(\neg F)^{str} = \neg(F^{wk}) \quad \text{and} \quad (\neg F)^{wk} = \neg(F^{str}). \tag{3}$$

Clearly, F^{str} and F^{wk} do not contain **u**. Consequently, they can be regarded as formulas in the propositional language generated by the atoms in

At and the two constants $\mathbf{t}$ and $\mathbf{f}$. We will call this language *2-FOL*. Formulas F^{wk} and F^{str} can be viewed as lower and upper approximations to the formula F.

It is clear that for every 2-FOL formula F, and for every interpretation $I \in \mathcal{A}$, $I^e(F) \in \{\mathbf{t}, \mathbf{f}\}$. We say that an interpretation $I \in \mathcal{A}$ is a *model* of a 2-FOL theory T if $I^e(F) = \mathbf{t}$. We will write $I \models F$ in such case. An interpretation $I \in \mathcal{A}$ is a *model* of a 2-FOL theory T ($I \models T$) if I is a model of every formula from T. The set of interpretations from $\mathcal{A}$ that are models of a 2-FOL formula F will be denoted by $\mathcal{M}od(T)$. The entailment relation in the language 2-FOL is now defined in the standard way: for two 2-FOL theories T_1 and T_2, $T_1 \models T_2$ if $\mathcal{M}od(T_2) \subseteq \mathcal{M}od(T_1)$. We have the following technical lemma.

Lemma 1. *For every interpretation $I \in \mathcal{A}$ and for every 3-FOL formula F:*

(a) $I^e(F) = \mathbf{t}$ *if and only if* $I \models F^{str}$
(b) $I^e(F) = \mathbf{f}$ *if and only if* $I \not\models F^{wk}$.

Proof: We will prove both (a) and (b) simultaneously by induction. Clearly, both (a) and (b) are true for every atom At and for the constants $\mathbf{t}$, $\mathbf{f}$ and $\mathbf{u}$.

Consider a 3-FOL formula G and assume both (a) and (b) hold for all 3-FOL formulas with length smaller than the length of G. Assume first that $G = F_1 \vee F_2$. Clearly, $I^e(F_1 \vee F_2) = \mathbf{t}$ if and only if $I^e(F_1) = \mathbf{t}$ or $I^e(F_2) = \mathbf{t}$. Similarly, $I \models (F_1 \vee F_2)^{str}$ if and only if $I \models {F_1}^{str}$ or $I \models {F_2}^{str}$. By the induction hypothesis, $I^e(F_i) = \mathbf{t}$ if and only if $I \models {F_i}^{str}$, $i = 1, 2$. Hence, the assertion (a) holds for $G = F_1 \vee F_2$.

Analogous arguments can be used to show that the assertion (b) holds for $G = F_1 \vee F_2$ and that both assertions (a) and (b) hold for $G = F_1 \wedge F_2$. Thus, to complete the proof, consider the case when $G = \neg F$. Then, $I^e(\neg F) = \mathbf{t}$ if and only if $I^e(F) = \mathbf{f}$. By the induction hypothesis, $I^e(F) = \mathbf{f}$ if and only if $I \not\models F^{wk}$. Moreover, by (3), $I \not\models F^{wk}$ if and only if $I \models (\neg F)^{str}$. By the induction hypothesis, $I^e(F) = \mathbf{f}$ if and only if $I \not\models F^{str}$. Hence, (a) holds for $G = \neg F$. The proof of (b) for $G = \neg F$ is similar. □

Lemma 1 has an important consequence. It implies that each 3-FOL theory generates a belief pair.

Corollary 2. *Let Y be a 3-FOL theory. Then, $\mathcal{M}od(Y^{str}) \subseteq \mathcal{M}od(Y^{wk})$. That is, equivalently, $(Mod(Y^{wk}), Mod(Y^{str}))$ is a belief pair.*

Proof: Let $I \in Mod(Y^{str})$ and let $F \in Y$. Then, $I \models F^{str}$. By Lemma 1(a), $I^e(F) = \mathbf{t}$. Hence, $I^e(F) \neq \mathbf{f}$ and, by Lemma 1(b), $I \models F^{wk}$. Consequently, $I \in Mod(Y^{wk})$ and $\mathcal{M}od(Y^{str}) \subseteq \mathcal{M}od(Y^{wk})$ follows. □

Let Y be a 3-FOL theory. The belief pair $(Mod(Y^{wk}), Mod(Y^{str}))$ will be denoted by $Bel(Y)$. We say that a belief pair B is *represented by a 3-FOL theory* Y if $B = Bel(Y)$. Clearly, the belief pair $\perp = (\mathcal{A}, \emptyset)$ is represented by the 3-FOL theory $\{\mathbf{u}\}$. We will now show that every belief pair in the range of the operator $\mathcal{D}_T$ is representable by a 3-FOL theory.

Let B be a belief pair and let F be a modal formula. By F_B we will denote a 3-FOL formula that is obtained from F by replacing each top level modal atom KG in F by the constant corresponding to the logical value $\mathcal{H}_B(KG)$. For a modal theory T, we define $T_B = \{F_B : F \in T\}$. We have the following result.

Theorem 7. *For every modal theory $T \subseteq \mathcal{L}_K$ and every belief pair B we have $\mathcal{D}_T(B) = Bel(T_B)$.*

Proof: First, observe that directly from the definitions of the truth assignment I^e and a 3-FOL formula F_B it follows that

$$I^e(F_B) = \mathcal{H}_{B,I}(F). \tag{4}$$

Now, we have

$$\mathcal{D}_T(B) = (\{I : \mathcal{H}_{B,I}(F) \geq_{tr} \mathbf{u}, \text{ for all } F \in T\}, \{I : \mathcal{H}_{B,I}(F) = \mathbf{t}, \text{ for all } F \in T\}).$$

Hence, by (4),

$$\mathcal{D}_T(B) = (\{I : I^e(F_B) \neq \mathbf{f}, \text{ for all } F \in T\}, \{I : I^e(F_B) = \mathbf{t}, \text{ for all } F \in T\}).$$

Finally, by Lemma 1,

$$\mathcal{D}_T(B) = (Mod(T_B^{wk}), Mod(T_B^{str})).$$

That is, $\mathcal{D}_T(B) = Bel(T_B)$. □

We will now show that, similarly to belief pairs, 3-FOL theories can be used to assign truth values to *modal* atoms (and, hence, to all modal formulas). We will then exhibit (Theorem 8) the relationship between this truth assignment and the truth assignment $\mathcal{H}_{B,I}$ introduced in Section 3. In order for the inductive argument in the proof of Theorem 8 to work, we need to extend the modal language $\mathcal{L}_K$ by the constants **t**, **f** and **u**. We call the resulting language *3-AEL*. We call formulas and theories in this language *3-AEL* formulas and *3-AEL* theories, respectively. Observe that the definition of the truth assignment $\mathcal{H}_{B,I}$ from Section 3 naturally extends to 3-AEL formulas.

Definition 4. Let Y be a 3-FOL theory, and let F be a 3-AEL formula. We define $\mathcal{H}_Y(KF)$ as follows. If F is a modal-free formula (that is a 3-FOL formula), then define:

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

If F is not modal free, then replace every modal atom KG in F, not under the scope of any other occurrence of the modal operator, by the constant corresponding to the value of $\mathcal{H}_Y(KG)$. Call the resulting formula F'. Define $\mathcal{H}_Y(KF) = \mathcal{H}_Y(KF')$ (notice that F' is a modal-free formula and the first part of the definition applies).

Let T be a modal theory and let Y be a 3-FOL theory. By the *Y-instance* of T, T_Y, we mean the 3-FOL theory obtained by substituting in each formula from T all modal atoms KF (not appearing under the scope of any other occurrence of the modal operator K) by the constant corresponding to $\mathcal{H}_Y(KF)$.

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

Theorem 8. *Let Y be a 3-FOL theory. Then, for every 3-AEL formula F,*

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

Proof: The proof is by induction on the length of the formula F. In what follows we denote $Bel(Y)$ by B. Thus, we also have $P(B) = \mathcal{M}od(Y^{wk})$ and $S(B) = \mathcal{M}od(Y^{str})$.

First consider the case when F is a 3-FOL formula (the argument in this case will establish the basis of the induction). By the definition, $\mathcal{H}_B(KF) = \mathbf{t}$ if and only if

$$\text{for every } I \in P(B),\ \mathcal{H}_{B,I}(F) = \mathbf{t}. \tag{5}$$

Since F is a 3-FOL formula, $\mathcal{H}_{B,I}(F) = I^e(F)$. Hence, by Lemma 1 and by the equality $P(B) = \mathcal{M}od(Y^{wk})$, the statement (5) is equivalent to:

$$\text{for every } I \in \mathcal{M}od(Y^{wk}),\ I \in \mathcal{M}od(F^{str}). \tag{6}$$

The statement (6), in turn, is equivalent to $Y^{wk} \models F^{str}$. Thus, $\mathcal{H}_B(KF) = \mathbf{t}$ if and only if $\mathcal{H}_Y(KF) = \mathbf{t}$. In a similar way one can prove that $\mathcal{H}_B(KF) = \mathbf{f}$ if and only if $\mathcal{H}_Y(KF) = \mathbf{f}$. Consequently, $\mathcal{H}_B(KF) = \mathcal{H}_Y(KF)$.

Second, consider the case when F is a modal 3-AEL formula. Let F' be a formula obtained from F by replacing each modal atom KG (not in the scope of any other occurrence of K in F) by the constant corresponding to the truth value $\mathcal{H}_B(KG)$. By the induction hypothesis $\mathcal{H}_B(KG) = \mathcal{H}_Y(KG)$. Hence, by the definition of $\mathcal{H}_Y(KF)$, $\mathcal{H}_Y(KF) = \mathcal{H}_Y(KF')$.

Since F' is modal-free, $\mathcal{H}_B(KF') = \mathcal{H}_Y(KF')$. In addition, it is easy to see that $\mathcal{H}_B(KF) = \mathcal{H}_B(KF')$. Thus, $\mathcal{H}_B(KF) = \mathcal{H}_Y(KF)$. □

Let T be a modal theory. We will now define an operator $\mathcal{SD}_T$ on 3-FOL theories. We will then show that this new operator is closely related 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. The number of such calls is bounded by the number of occurrences of the modal operator K in the theory T. In each call we verify whether some 2-FOL theory X_1 entails another 2-FOL theory X_2, where the sizes of X_1 and X_2 are bounded by the sizes of the theories T and Y.

Theorems 8 and 7 imply the main result of this section.

Theorem 9. *Let T be a modal theory and let Y be a 3-FOL theory. Then,*

(a) $T_Y = T_{\mathcal{B}el(Y)}$ *and* $\mathcal{SD}_T(Y) = T_{\mathcal{B}el(Y)}$
(b) $\mathcal{B}el(\mathcal{SD}_T(Y)) = \mathcal{D}_T(\mathcal{B}el(Y))$.
(c) *If a belief pair B is a fixpoint of $\mathcal{D}_T$, then T_B is a fixpoint of $\mathcal{SD}_T$.*
(d) *If Y is a fixpoint of $\mathcal{SD}_T$ then $\mathcal{B}el(Y)$ is a fixpoint of $\mathcal{D}_T$.*

Proof: (a) This statement follows directly from Theorem 8.
(b) By Theorem 7, $\mathcal{D}_T(\mathcal{B}el(Y)) = \mathcal{B}el(T_{\mathcal{B}el(Y)})$. By (a), $\mathcal{D}_T(\mathcal{B}el(Y)) = \mathcal{B}el(\mathcal{SD}_T(Y))$ follows.
(c) If a belief pair B is a fixpoint of $\mathcal{D}_T$, then $B = \mathcal{D}_T(B) = \mathcal{B}el(T_B)$ (the last equality follows by Theorem 7). Hence, $T_B = T_{\mathcal{B}el(T_B)} = \mathcal{SD}_T(T_B)$ (the last equality follows by (a)).
(d) This statement follows directly from (b). □

Let us denote $B_\alpha = \mathcal{D}_T^\alpha(\bot)$ and $Y_\alpha = \mathcal{SD}_T^\alpha(\mathbf{u})$. It follows directly from Theorem 9(b) (by an easy induction) that for every ordinal number α, $B_\alpha = Bel(Y_\alpha)$. Hence, if $Y_\alpha = Y_{\alpha+1}$ (that is, if Y_α is a fixpoint of the operator $\mathcal{SD}_T$) $B_\alpha = B_{\alpha+1}$ (that is, B_α is a fixpoint of the operator $\mathcal{D}_T$).

Next, by (a), for every ordinal α, $T_{Y_\alpha} = T_{B_\alpha}$. Hence, if $B_\alpha = B_{\alpha+1}$ (that is, if B_α is a fixpoint of the operator $\mathcal{D}_T$) then

$$Y_{\alpha+2} = \mathcal{SD}_T(Y_{\alpha+1}) = T_{Y_{\alpha+1}} = T_{B_{\alpha+1}} = T_{B_\alpha} = T_{Y_\alpha} = \mathcal{SD}_T(Y_\alpha) = Y_{\alpha+1}.$$

That is, $Y_{\alpha+1}$ is a fixpoint o the operator $\mathcal{SD}_T$.

It follows that

$$\mathcal{D}_T{\uparrow} = \mathcal{B}el(\mathcal{SD}_T{\uparrow}).$$

In the case when T is finite, the number of iterations needed to compute $\mathcal{SD}_T{\uparrow}$ is limited by the number of top level (unnested) modal literals in T. Originally, they may all be evaluated to **u**. However, at each step, at least one **u** changes to either **t** or **f** and this value is preserved in the subsequent evaluations.

Once $\mathcal{SD}_T{\uparrow}$ is computed, one can evaluate the truth value $\mathcal{H}_{\mathcal{SD}_T{\uparrow}}(KG)$ for any modal atom of the language $\mathcal{L}_K$. This task again requires polynomially many calls to a propositional provability procedure. A key point is that the logical value so computed is exactly the logical value of the modal atom KG with respect to the belief pair $\mathcal{D}_T{\uparrow}$. In other words, determining the logical value of a modal formula with respect to the semantics defined by the least fixpoint of the operator $\mathcal{D}_T$ takes a polynomial number of calls to a propositional provability procedure. Consequently, the problem to decide whether a logical value of a modal atom under this semantics is **t** is in the class Δ_P^2 (the same is true for two other decision problems of deciding whether the logical value of a modal atom is **u** and **f**, respectively). Since deciding whether a modal atom is in all (some of the) expansions of a modal theory is Π_P^2-complete (Σ_P^2-complete), our 3-valued semantics is computationally simpler (unless the polynomial hierarchy collapses at some low level). These considerations yield the following formal result.

Theorem 10. *The problems to decide whether* $\mathcal{H}_{\mathcal{D}_T\uparrow}(KF) = \mathbf{t}$, $\mathcal{H}_{\mathcal{D}_T\uparrow}(KF) = \mathbf{f}$ *and* $\mathcal{H}_{\mathcal{D}_T\uparrow}(KF) = \mathbf{u}$ *are in the class* Δ_P^2.

5 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, [10]). In this section, we discuss connections of the semantics defined by the least fixpoint of the operator $\mathcal{D}$ to some 3-valued semantics of logic programs.

We will be interested in propositional logic programs over a set of atoms At. However, to prove the main results of the section and to state some auxiliary facts, we will also consider a wider class of programs. These programs, called 3-FOL programs, will play a similar role as 3-FOL theories in Section 4. Formally, a *3-FOL program clause* is an expression of the form

$$a \leftarrow b_1, \ldots, b_k, \mathbf{not}(c_1), \ldots, \mathbf{not}(c_m), l_1, \ldots, l_n,$$

where a, each b_i, $1 \leq i \leq k$, and each c_i, $1 \leq i \leq m$ are atoms from At, and each l_i, $1 \leq i \leq n$, is one of $\mathbf{t}$, $\mathbf{f}$, $\mathbf{u}$ or their negation. The literals l_i will be referred to as *truth-value literals*. A 3-FOL clause in which $m = 0$ (no literals of the form $\mathbf{not}(c)$ in the body) is called a *definite* 3-FOL clause. A collection of 3-FOL clauses (definite 3-FOL clauses, respectively) is a *3-FOL program* (*definite 3-FOL program*).

We will often interpret a definite 3-FOL logic program P as a 3-FOL theory (by regarding program clauses as implications). This allows us to use for definite 3-FOL programs concepts introduced in Section 4 for 3-FOL theories. In particular, with every definite 3-FOL program we will associate 2-FOL theories P^{str} and P^{wk}, as well as the belief pair $\mathcal{B}el(P) = (Mod(P^{wk}), Mod(P^{str}))$.

Consider a 3-FOL definite logic program P. We say that a 3-valued interpretation I strongly satisfies P if for each rule

$$a \leftarrow b_1, \ldots, b_k, l_1, \ldots, l_n$$

from P, $I^e(a) \geq_{tr} I^e(b_i)$, for some i, $1 \leq i \leq k$ or $I^e(a) \geq_{tr} I^e(l_i)$, for some i, $1 \leq i \leq n$ (I^e is obtained from I by extending I naturally to the constants $\mathbf{t}$, $\mathbf{f}$ and $\mathbf{u}$)[2].

It is easy to see that every definite 3-FOL program has a least 3-valued model with respect to the truth ordering (see [16]). We will denote this model by $LM_3(P)$.

Since P is a definite 3-FOL program, theories P^{wk} and P^{str} are both definite 3-FOL programs. Thus, each has a least model (with respect to the

[2] This means that I satisfies the rule $p \leftarrow B$ in the strong Kleene truth table.

truth ordering). Moreover, since $\mathbf{u}$ does not occur in P^{wk} and P^{str}, these least models are two valued. We will denote them by $LM(P^{wk})$ and $LM(P^{str})$, respectively. Let $I, J \in \mathcal{A}$. Define $I \leq_{tr} J$ if for all atoms A, $I(A) \leq_{tr} J(A)$. We have the following simple technical lemma connecting the three interpretations $LM_3(P)$, $LM(P^{wk})$ and $LM(P^{str})$. The proof is easy and is left to the reader.

Lemma 2. *Let P be a definite 3-FOL program. Then:*

(a) $LM(P^{wk}) \leq_{tr} LM(P^{str})$
(b) $LM_3(P) = \mathbf{t}$ *if and only if* $LM(P^{wk}) = \mathbf{t}$, *and* $LM_3(P) = \mathbf{f}$ *if and only if* $LM(P^{str}) = \mathbf{f}$.

Let B be a belief pair. Define the *projection*, $Proj(B)$, as the 3-valued interpretation I such that $I(p) = \mathcal{H}_B(Kp)$. We have the following theorem relating, for definite 3-FOL program P its belief pair $\mathcal{B}el(P)$ with its least model $LM_3(P)$.

Theorem 11. *For any 3-FOL definite program P, $Proj(\mathcal{B}el(P)) = LM_3(P)$.*

Proof: By the definition of $Proj(\mathcal{B}el(P))$ and by Theorem 8 we have

$$Proj(\mathcal{B}el(P))(p) = \mathcal{H}_{\mathcal{B}el(P)}(Kp) = \mathcal{H}_P(Kp) \tag{7}$$

(slightly abusing the notation, we use the same symbol P to denote both a 3-FOL program and the corresponding 3-FOL theory).

Hence, by (7) and by Definition 4, $Proj(\mathcal{B}el(P))(p) = \mathbf{t}$ if and only if $P^{wk} \models p$. The entailment $P^{wk} \models p$ is, in turn, equivalent to $LM(P^{wk})(p) = \mathbf{t}$. By Lemma 2(b), it follows then that $Proj(\mathcal{B}el(P))(p) = \mathbf{t}$ if and only if $LM_3(P)(p) = \mathbf{t}$.

Similarly, by (7) and by Definition 4, $Proj(\mathcal{B}el(P))(p) = \mathbf{f}$ if and only if $P^{str} \not\models p$ or, equivalently, if and only if $LM(P^{str})(p) = \mathbf{f}$. Hence, by Lemma 2(b), $Proj(\mathcal{B}el(P))(p) = \mathbf{f}$ if and only if $LM_3(P)(p) = \mathbf{f}$. □

We will now study the relationship between logic programming and autoepistemic logic. Given a logic programming clause (over the alphabet At)

$$r = a \leftarrow b_1, \ldots, b_k, \mathbf{not}(c_1), \ldots, \mathbf{not}(c_m)$$

define:

$$ael_1(r) = Kb_1 \wedge \ldots \wedge Kb_k \wedge \neg Kc_1 \wedge \ldots \wedge \neg Kc_m \supset a$$

and

$$ael_2(r) = b_1 \wedge \ldots \wedge b_k \wedge \neg Kc_1 \wedge \ldots \wedge \neg Kc_m \supset a$$

Embeddings $ael_1(\cdot)$ and $ael_2(\cdot)$ naturally extend to logic programs P.

In the remainder of this paper we show 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 $\mathcal{D}_{ael_1(P)}$ ($\mathcal{D}_{ael_2(P)}$) describe 2-valued supported (stable, respectively) models of P. Finally, the least fixpoint of $\mathcal{D}_{ael_1(P)}$ captures the Fitting-Kunen 3-valued semantics of a program P, and the least fixpoint of $\mathcal{D}_{ael_2(P)}$ captures the well-founded semantics of P.

We will focus first on the embedding $ael_1(\cdot)$. It establishes the relationship between stable expansions and supported models and between the least fixpoint of the operator $\mathcal{D}_{ael_1(P)}$ and the Fitting-Kunen 3-valued semantics of a program P.

Let P be a logic program. Let us recall a definition of the 3-valued stepwise inference operator $\mathcal{T}_P$ [2]:

$$\mathcal{T}_P(I) = I' \quad \text{where } I'(p) = \max(\{I^e(body) \colon p \leftarrow body \in P\})$$

(here we treat *body* as the conjunction of its literals). It is well known that fixpoints of the operator $\mathcal{T}_P$ are models of the program P. These models are called *3-valued supported models*. It is also known [2] that every logic program P has a least (with respect to the knowledge ordering) 3-valued supported model. This model determines a semantics of P known as Fitting-Kunen semantics.

For a given 3-valued interpretation I, and a logic program P, define P_I^{sp} as the *definite* 3-FOL program obtained from P by substituting, in the bodies of rules in P $I(p)$, for each atom p occurring positively and $\neg I(p)$ for each literal $\mathbf{not}(p)$. We have the following lemma (its proof is straightforward and is omitted).

Lemma 3. *Let P be a logic program over the set of atoms At.*

(a) *Let I be a 3-valued interpretation of At. $\mathcal{T}_P(I) = LM_3(P_I^{sp})$*
(b) *If B is a belief pair then $P^{sp}_{Proj(B)} = (ael_1(P))_B$.*

Equipped with Lemma 3 we are ready to prove the first of the two main results of this section.

Theorem 12. *Let P be a logic program over the set of atoms At.*

(a) *For every belief pair B, $\mathcal{T}_P(Proj(B)) = Proj(\mathcal{D}_{ael_1(P)}(B))$*
(b) *If a belief pair B is a fixpoint of $\mathcal{D}_{ael_1(P)}$ then $Proj(B)$ is a 3-valued supported model of P*
(c) *If I is a 3-valued supported model of P, then $B = \mathcal{B}el(P_I^{sp})$ is a fixpoint of $\mathcal{D}_{ael_1(P)}$ and $Proj(B) = I$*
(d) *$Proj(\mathcal{D}_{ael_1(P)}\uparrow)$ is the $\leq_{kn}$-least 3-valued supported model of P (the model defining the Fitting-Kunen semantics)*
(e) *If a belief pair B is a complete fixpoint of $\mathcal{D}_{ael_1(P)}$, then $Proj(B)$ is a 2-valued supported model of P. Moreover, each 2-valued supported model of P is of this form.*

Proof: (a) Clearly,

$$\mathcal{T}_P(Proj(B)) = LM_3(P^{sp}_{Proj(B)}) = LM_3(ael_1(P)_B) = Proj(\mathcal{B}el(ael_1(P)_B)) = Proj(\mathcal{D}_{ael_1(P)}(B))$$

(the first two equalities follow by Lemma 3, the third one follows by Theorem 11 and the last one by Theorem 7).
(b) If B is a fixpoint of $\mathcal{D}_{ael_1(P)}$ then by (a), $\mathcal{T}_P(Proj(B)) = Proj(\mathcal{D}_{ael_1(P)}(B)) = Proj(B)$.
(c) Since I is a fixpoint of $\mathcal{T}_P$, $I = \mathcal{T}_P(I) = LM_3(P^{sp}_I)$ (Lemma 3(a)). Let $B = \mathcal{B}el(P^{sp}_I)$. By Theorem 11, $Proj(B) = LM_3(P^{sp}_I) = I$. Hence, by Lemma 3(b), $(ael_1(P))_B = P^{sp}_I$. Consequently, by Theorem 7, $\mathcal{D}_{ael_1(P)}(B) = \mathcal{B}el(ael_1(P)_B) = \mathcal{B}el(P^{sp}_I) = B$.
(d) Let $B = \mathcal{D}_{ael_1(P)}\uparrow$. By (b), $Proj(B)$ is a supported model of P. Consider another supported model I of P. It follows that $B' = \mathcal{B}el(P^{sp}_I)$ is a fixpoint of $\mathcal{D}_{ael_1(P)}$ and that $Proj(\mathcal{B}el(P^{sp}_I)) = I$.

Clearly, $B \leq_p B'$. Proposition 2 entails that for each atom p, $\mathcal{H}_B(Kp) \leq_{kn} \mathcal{H}_{B'}(Kp)$. By the definition of $Proj(\cdot)$, $Proj(B) \leq_{kn} Proj(B') = I$, that is, $Proj(B)$ is the $\leq_{kn}$-least 3-valued supported model of P.
(e) This assertion follows from the observation that if B is complete then $Proj(B)$ is 2-valued (Proposition 1). □

We will now discuss the second embedding, $ael_2(\cdot)$, of logic programs into autoepistemic logic.

Recall the definition of the 3-valued version $\mathcal{GLP}_P$ of the Gelfond and Lifschitz operator (see, for instance, [16]). Given a logic program P and a 3-valued interpretation I, P_I is the program where negative body literals $\mathbf{not}(p)$ are replaced by $\neg I(p)$[3]. Then, $\mathcal{GLP}_P(I)$ is defined as $LM_3(P_I)$. Fixpoints of the operator $\mathcal{GLP}_P$ are known to be 3-valued models of P. These 3-valued models are called *stable*. The *well-founded model* of P is the $\leq_{kn}$-least fixpoint of $\mathcal{GLP}_P$ [16].

We have now the following technical lemma and the second main result of this section on the relationship between fixpoints of the operator $\mathcal{D}_{ael_2(P)}$ and 3-valued stable models of P.

Lemma 4. *If $I = Proj(B)$, then $P_I = (ael_2(P))_B$.*

Theorem 13. *Let P be a logic program over the set of atoms At.*

(a) $\mathcal{GLP}_P(Proj(B)) = Proj(\mathcal{D}_{ael_2(P)}(B))$
(b) *If a belief pair B is a fixpoint of $\mathcal{D}_{ael_2(P)}$ then $Proj(B)$ is a 3-valued stable model of P*
(c) *If I is a 3-valued stable model of P, then $B = \mathcal{B}el(P_I)$ is a fixpoint of $\mathcal{D}_{ael_2(P)}$ and $Proj(B) = I$*
(d) *$Proj(\mathcal{D}_{ael_2(P)}\uparrow)$ is the well-founded model of P.*

[3] Normally, P_I is further simplified, by deleting rules with $\neg\mathbf{t}$ in the body and deleting literals $\neg\mathbf{f}$ in the body of rules.

(e) *If a belief pair B is a complete fixpoint of $\mathcal{D}_{ael_2(P)}$, then Proj(B) is a 2-valued stable model of P. Moreover, all 2-valued stable models of P are of this form.*

6 Conclusions and future work

In this paper we investigated the conctructive approximation scheme for Moore's autoepistemic logic. We introduced the notion of a belief pair — a Kripke-style 3-valued structure for the modal language. The set of belief pairs $\mathcal{B}$ is endowed with a natural ordering $\leq_p$. This ordering is chain complete, which guarantees that every monotone operator on $(\mathcal{B}, \leq_p)$ has a least fixpoint. With every modal theory T we associated a monotone *derivation* operator $\mathcal{D}_T$ on $(\mathcal{B}, \leq_p)$. We proposed the least fixpoint of the operator $\mathcal{D}_T$ as the intended constructive 3-valued semantics of modal theory T. We proved that the complete fixpoints of the operator $\mathcal{D}_T$ coincide with Moore's autoepistemic models of T. Thus, the semantics specified by the least fixpoint of $\mathcal{D}_T$ approximates Moore's semantics. Under appropriate embeddings of a logic program P as a modal theory T ($T = ael_1(P)$ or $T = ael_2(P)$), the least fixpoint of the operator $\mathcal{D}_T$ generalizes Kunen-Fitting semantics and Van Gelder-Ross-Schlipf well-founded semantics. These results provide further evidence of the correctness of our approach.

It is natural to ask how general is the technique proposed in our paper. In the forthcoming work we show that the scheme proposed in this paper can be generalized and that one can develop a theory of approximating operators. Specifically, we elucidate the abstract content of the well-founded semantics in terms of a suitably chosen approximation operator in a chain-complete poset.

Acknowledgments

This work was partially supported by the NSF grants IRI-9400568 and IRI-9619233

References

1. N. Bourbaki. *Elements of Mathematics Theory of Sets.* Hermann, 1968.
2. M. C. Fitting. A Kripke-Kleene semantics for logic programs. *Journal of Logic Programming*, 2(4):295–312, 1985.
3. M. Gelfond. On stratified autoepistemic theories. In *Proceedings of AAAI-87*, pages 207–211. Morgan Kaufmann, 1987.
4. G. Gottlob. Complexity results for nonmonotonic logics. *Journal of Logic and Computation*, 2(3):397–425, 1992.
5. G. Gottlob. Translating default logic into standard autoepistemic logic. *Journal of the ACM*, 42(4):711–740, 1995.

6. K. Konolige. On the relation between default and autoepistemic logic. *Artificial Intelligence*, 35(3):343–382, 1988.
7. K. Kunen. Negation in logic programming. *Journal of Logic Programming*, 4(4):289–308, 1987.
8. H. J. Levesque. All I know: a study in autoepistemic logic. *Artificial Intelligence*, 42(2-3):263–309, 1990.
9. W. Marek and M. Truszczyński. Autoepistemic logic. *Journal of the ACM*, 38(3):588–619, 1991.
10. W. Marek and M. Truszczyński. *Nonmonotonic logics; context-dependent reasoning*. Springer-Verlag, Berlin, 1993.
11. G. Markowsky. Chain-complete posets and directed sets with applications. *Algebra Universalis*, 6(1):53–68, 1976.
12. R.C. Moore. Possible-world semantics for autoepistemic logic. In *Proceedings of the Workshop on Non-Monotonic Reasoning*, pages 344–354, 1984. Reprinted in: M. Ginsberg, ed., *Readings on nonmonotonic reasoning*, pp. 137–142, Morgan Kaufmann, 1990.
13. R.C. Moore. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, 25(1):75–94, 1985.
14. I. Niemelä. On the decidability and complexity of autoepistemic reasoning. *Fundamenta Informaticae*, 17(1-2):117–155, 1992.
15. I. Niemelä and P. Simons. Evaluating an algorithm for default reasoning. In *Proceedings of the IJCAI-95 Workshop on Applications and Implementations of Nonmonotomic Reasoning Systems*, 1995.
16. T.C. Przymusinski. The well-founded semantics coincides with the three-valued stable semantics. *Fundamenta Informaticae*, 13(4):445–464, 1990.
17. R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1-2):81–132, 1980.
18. G.F. Schwarz. Autoepistemic logic of knowledge. In A. Nerode, W. Marek, and V.S. Subrahmanian, editors, *Logic programming and nonmonotonic reasoning (Washington, DC, 1991)*, pages 260–274, Cambridge, MA, 1991. MIT Press.
19. G.F. Schwarz. Minimal model semantics for nonmonotonic modal logics. In *Proceedings of LICS-92*, pages 34–43, 1992.
20. A. Van Gelder, K.A. Ross, and J.S. Schlipf. The well-founded semantics for general logic programs. *Journal of the ACM*, 38(3):620–650, 1991.

Appendix: Stratified autoepistemic theories

We will present here a proof of Theorem 6. We will start by recalling the concept of stratification. We will use the original definition by Gelfond [3]. However, our argument can easily be extended to a slightly wider class of theories considered in [9].

A modal formula is called a *modal clause* if it is of the form

$$l_1 \wedge \ldots \wedge l_k \wedge KF_1 \wedge \ldots \wedge KF_m \neg KG_1 \wedge \ldots \wedge \neg KG_r \supset p_1 \vee \ldots \vee p_s$$

where $l_1, \ldots, l_k$ are literals of $\mathcal{L}$, $p_1, \ldots, p_s$ are atoms of $\mathcal{L}$, and $F_1, \ldots, G_r$ are formulas of $\mathcal{L}$.

A theory consisting of modal clauses is called *stratified* if there are pairwise disjoint theories $T_0, \ldots, T_n$ such that

i. $\bigcup_{i=0}^{n} T_i = T$
ii. T_0 is modal-free
iii. For every m, $0 < m \leq n$, all clauses in T_m have nonempty conclusions (that is, $s > 0$)
iv. Whenever p appears in a conclusion of a clause in T_j, $j > 0$, then p does not appear in T_i, $i < j$ and p does not appear within the scope of the modal operator K in T_i, $i \leq j$.

We call the list $\langle T_0, \ldots, T_n \rangle$ a *stratification* of T. In the remainder of this section, we write $T = T_0 \cup \ldots \cup T_n$ to indicate that $\langle T_0, \ldots, T_n \rangle$ is a stratification of T.

A stratification $T = T_0 \cup \ldots \cup T_n$ generates an increasing family of subsets of the set of atoms At. Namely, At_0 is the set of those atoms in At that do not occur in the conclusions of modal clauses from T_i, where $i > 0$, and

$$At_i = At_{i-1} \cup \{p\colon p \text{ occurs in the conclusion of a clause in } T_i\},$$

for $i = 1, \ldots, n$.

For an interpretation I and a set $Z \subseteq At$, by $I|Z$ we denote the restriction of I to Z. This concept is naturally extended to sets of interpretations and to belief pairs. For a set R of interpretations, we define $R|Z = \{I|Z : I \in R\}$ and, for a belief pair B, we define $B|Z = (P(B)|Z, S(B)|Z)$.

We say that a formula F is *based* on set of atoms Z if all atoms occurring in F belong to Z. The following simple lemma (we leave it without proof) gathers several facts on restrictions.

Lemma 5. *Let $Z \subseteq At$ and let F be a formula based on Z. Then for every belief pair B and interpretation I:*

(a) $(B, I) \models F$ *if and only if* $(B|Z, I|Z) \models F$
(b) $(B, I) \models_w F$ *if and only if* $(B|Z, I|Z) \models_w F$.

Consider a stratified theory $T = T_0 \cup \ldots \cup T_n$. We will now construct a sequence of belief pairs $B_0, \ldots, B_{n+1}$. Namely, we set $B_0 = \bot$ and for every i, $0 \leq i \leq n$, $B_{i+1} = (P_{i+1}, S_{i+1})$ where:

$$P_{i+1} = \{I \in P_i\colon (B_i, I) \models T_i\}$$

and

$$S_{i+1} = \{I \in P_i\colon (B_i, I) \models T_i \text{ and for every } p \in At \setminus At_i, I(p) = \mathbf{t}\}.$$

Lemma 6. *For every i, $1 \leq i \leq n+1$, $B_i|At_{i-1}$ is complete. Furthermore, for every interpretation $I \in \mathcal{A}$, $(B_i, I) \models T_i$ if and only if $(B_i, I) \models_w T_i$.*

Proof: Clearly, $S_i \subseteq P_i$. In particular, it follows that $S_i|At_{i-1} \subseteq P_i|At_{i-1}$. Consider now a valuation $I' \in P_i|At_{i-1}$. Then, there is a valuation $I \in P_i$ such that $I|At_{i-1} = I'$. Denote by J a valuation obtained from I by setting:

$$J(p) = \begin{cases} I(p) & \text{if } p \in At_{i-1} \\ \mathbf{t} & \text{if } p \in At \setminus At_{i-1}. \end{cases}$$

Since $I \in P_i$, $(B_{i-1}, I) \models T_{i-1}$. By Lemma 5, $(B_{i-1}, J) \models T_{i-1}$. Thus, by the definition of J, $J \in S_i$ and, consequently, $I' = I|At_{i-1} = J|At_{i-1} \in S_i|At_{i-1}$. Hence, for every i, $1 \leq i \leq n+1$, $P_i|At_{i-1} = S_i|At_{i-1}$. In other words, $B_i|At_{i-1}$ is complete.

By the definition of stratification, every modal atom KF occurring in a modal clause from T_i is based on the set of atoms At_{i-1}. Thus, the second part of the assertion follows from the completeness of the belief pair $B_i|At_{i-1}$ and from Lemma 5. □

The following lemma plays the key role in the proof of Theorem 6.

Lemma 7. *Let $T = T_0 \cup \ldots \cup T_n$ be a stratified theory. Then for every i, $0 < i \leq n+1$, $B_i \leq_p \mathcal{D}^i(\bot)$.*

Proof: We will proceed by induction on i. Let $i = 1$. Clearly, $P(B_1) = \{I : (\bot, I) \models T_0\}$ and $P(\mathcal{D}_T(\bot)) = \{I : (\bot, I) \models_w T\}$. Since T_0 is modal-free,

$$(\bot, I) \models T_0 \text{ if and only if } (\bot, I) \models_w T_0$$

Thus $P(\mathcal{D}_T(\bot)) \subseteq P(B_1)$ follows.

Consider now $I \in S(B_1)$. Then $(\bot, I) \models T_0$ and for every $p \in At \setminus At_0$, $I(p) = \mathbf{t}$. Since every clause in $T \setminus T_0$ has at least one positive atom in the conclusion, $(\bot, I) \models T$. Thus, $I \in S(\mathcal{D}_T(\bot))$. Consequently, $B_0 \leq_p \mathcal{D}^0(\bot)$. That is, the basis for the induction is established.

For the inductive step, we need to prove that $P(B_{i+1}) \supseteq P(\mathcal{D}_T^{i+1}(\bot))$ and $S(B_{i+1}) \subseteq S(\mathcal{D}_T^{i+1}(\bot))$. Consider an interpretation $I \notin P_{i+1}$. Then, either $I \notin P_i$ or $(B_i, I) \not\models T_i$. In the first case, since $B_i \leq_p \mathcal{D}_T^i(\bot) \leq_p \mathcal{D}_T^{i+1}(\bot)$, $I \notin P(\mathcal{D}_T^{i+1}(\bot))$. In the second case, by Lemma 6, $(B_i, I) \not\models_w T_i$. Consequently, by Proposition 2, $(\mathcal{D}_T^i(\bot), I) \not\models_w T_i$ and, hence also in this case, $I \notin P(\mathcal{D}_T^{i+1}(\bot))$. Thus, $P(B_{i+1}) \supseteq P(\mathcal{D}_T^{i+1}(\bot))$ follows.

Next, consider $I \in S_{i+1}$. By the definition, $I \in P_i$, $(B_i, I) \models T_i$ and for every $p \in At \setminus At_i$, $I(p) = \mathbf{t}$. We will show that $(\mathcal{D}_T^i(\bot), I) \models T$ (or, equivalently, that $I \in S(\mathcal{D}_T^{i+1}(\bot))$).

Consider stratum T_j with $j < i$. Then $P_i \subseteq P_{j+1}$. Since $I \in S_{i+1}$, $I \in P_i$ and, hence, $I \in P_{j+1}$. By the definition of P_{j+1}, $(B_j, I) \models T_j$. By the induction hypothesis, $B_j \leq_p \mathcal{D}_T^j(\bot)$. Thus, $B_j \leq_p \mathcal{D}_T^i(\bot)$. It now follows from Proposition 2 that $(\mathcal{D}_T^i(\bot), I) \models T_j$.

Next, consider stratum T_i. Since $I \in S_{i+1}$, $(B_i, I) \models T_i$. By the induction hypothesis, $B_i \leq_p \mathcal{D}_T^i(\bot)$. Hence, by Proposition 2, $(\mathcal{D}_T^i(\bot), I) \models T_i$.

Finally, consider stratum T_j with $j > i$. Since the conclusion of every modal clause in T_j contains a positive occurrence of an atom in $At \setminus At_i$ and since $I(p) = \mathbf{t}$ for every atom $p \in At \setminus At_i$, $(\mathcal{D}_T^i(\bot), I) \models T_j$.

To summarize, it follows that $(\mathcal{D}_T^i(\bot), I) \models T$. Consequently, $I \in S(\mathcal{D}_T^{i+1}(\bot))$.
□

We now prove Theorem 6 from Section 3.

Theorem 6 *If T is a stratified autoepistemic theory then:*

(a) *$\mathcal{D}_T\uparrow$ is complete*
(b) *T has a unique stable expansion*
(c) *$\mathcal{D}_T\uparrow$ is consistent if and only if the lowest stratum T_0 is consistent.*

Proof: (a) Clearly, Lemma 6 implies that B_{n+1} is a complete belief pair. By Lemma 7, $B_{n+1} \leq_p \mathcal{D}_T^{n+1}(\bot)$. Hence, it follows that $B_{n+1} = \mathcal{D}_T^{n+1}(\bot)$. Thus, $\mathcal{D}_T^{n+1}(\bot)$ is a fixpoint. Hence, it is a least fixpoint and, since it coincides with B_{n+1}, it is complete.
(b) The assertion follows directly from (a) by Theorem 1.
(c) Clearly, if T_0 is inconsistent, $B_1 = (\emptyset, \emptyset)$ and it is a least fixpoint of $\mathcal{D}_T$. On the other hand, if T_0 is consistent, it is easy to see that $S_1 \neq \emptyset$. Hence, $\mathcal{D}_T$ is consistent. □