Uniform semantic treatment of default and autoepistemic logics*

Marc Denecker
Département d'Informatique
Université Libre de Bruxelles
Boulevard du Triomphe CP 212
1050 Brussels, Belgium

Victor W. Marek
Department of Computer Science
University of Kentucky
Lexington, KY 40506-0046
USA

Mirosław Truszczyński
Department of Computer Science
University of Kentucky
Lexington, KY 40506-0046
USA

Abstract

We revisit the issue of epistemological and semantic foundations for autoepistemic and default logics, two leading formalisms in nonmonotonic reasoning. We develop a general semantic approach to autoepistemic and default logics that is based on the notion of a *belief pair* and that exploits the lattice structure of the collection of all belief pairs. For each logic, we introduce a monotone operator on the lattice of belief pairs. We then show that a whole *family* of semantics can be defined in a systematic and principled way in terms of fixpoints of this operator (or as fixpoints of certain closely related operators). Our approach elucidates fundamental constructive principles in which agents form their belief sets, and leads to approximation semantics for autoepistemic and default logics. It also allows us to establish a precise one-to-one correspondence between the family of semantics for default logic and the family of semantics for autoepistemic logic. The correspondence exploits the modal interpretation of a default proposed by Konolige. Our results establish conclusively that default logic can be viewed as a fragment of autoepistemic logic, a result that has been long anticipated. At the same time, they explain the source of the difficulty to formally relate the semantics of default extensions by Reiter and autoepistemic expansions by Moore. These two semantics occupy *different* locations in the corresponding families of semantics for default and autoepistemic logics.

1 Introduction

In this paper, we conduct a systematic study of epistemological and semantic foundations of default and autoepistemic logics. We identify some basic *common* principles behind these two formalisms and state them in terms of lattices of possible-world structures and belief pairs, operators on these lattices and fixpoints of these operators. We use these principles to develop a comprehensive unified semantic treatment of default and autoepistemic logics. The approach

*This is a full version of the extended abstract published in the Proceedings of KR'2000 [DMT00a].

we propose helps explain how agents form their belief sets. It allows us to study constructive ways in which belief sets can be approximated and leads to a comprehensive and precise account of the relationship between the two logics in question.

The default logic introduced by Reiter [Rei80] and the autoepistemic logic introduced by Moore [Moo84, Moo85] are among the most widely studied nonmonotonic knowledge representation systems. Research monographs [Bes89, MT93, Ant97] provide extensive presentations of these two logics, and of their properties. The default and the autoepistemic logics were designed to model commonsense forms of reasoning, in particular, reasoning patterns of the form "*in the absence of any information to the contrary infer ...*". Such patterns were seen as a basic reasoning mechanism in the context of partial knowledge. Reiter referred to such patterns as *defaults*.

In the default logic of Reiter, a default is represented as a non-standard inference rule

$$\frac{\alpha : M\beta_1, \ldots, M\beta_k}{\gamma},$$

where α, β_i, $1 \leq i \leq k$, and γ are propositional formulas (we limit our attention to the propositional case only). Speaking informally, the intended meaning of a default is: *if α can be derived and if for every i, $1 \leq i \leq k$, β_i is consistent, then derive γ*. This intuition points to the key idea behind a default. It has premises of two different types. The premise α is called the *prerequisite* of a default and is treated just like premises of standard (monotone) inference rules. Premises β_i are called *justifications*. The symbol M that prefixes justifications is commonly used in modal logic to denote the modality of "being consistent". Reiter used it to emphasize the way in which justifications are interpreted. In order to apply the rule, they just need to be consistent (rather than derived). To formally define the semantics of default logic, Reiter provided a precise mathematical interpretation for the phrases "α can be derived and "β_i is consistent" and defined the notion of an *extension* as a formal representation of a belief set that an agent might adopt when reasoning on the basis of a default theory.

Though Reiter used the modal notation $M\beta$ for justifications to emphasize the intended meaning of justifications, he did not use the syntax of modal logic nor the semantic techniques developed there. In fact, in current literature dealing with default logic, the letter M is dropped from the notation of justifications. We will also do so throughout the paper. The idea to use a modal language and modal logic techniques in the area of nonmonotonic logics is due to McDermott and Doyle [MD80, McD82]. For the primary modality in the language they chose the modality of consistency which, as we mentioned earlier, is commonly denoted by M. McDermott and Doyle [MD80] and, later, McDermott [McD82] introduced several nonmonotonic semantics for the operator M. They suggested that an inference rule of the form "*infer γ if in the absence of any information contradicting β*" should be represented by the modal formula $M\beta \Rightarrow \gamma$.

Moore [Moo85] developed autoepistemic logic departing from the nonmonotonic logics of McDermott and Doyle. Moore pointed out some technical problems arising in the context of the original modal nonmonotonic logic described in [MD80], observing that the notion of consistency proposed there was too weak. He also discarded a more refined approach from [McD82], where modal nonmonotonic logics based on monotone modal systems T, S4 and S5 were studied. Moore suggested to use the modality of belief as the primary modality in the language. This modality is usually denoted by K and is related to M through the identity $M = \neg K \neg$. Further, and more importantly, Moore suggested that the semantics of a modal

nonmonotonic logic be designed so that to model reasoning of a rational agent reflecting on her own beliefs. Moore postulated that such a rational agent should have perfect introspection capabilities. That is, if φ is in the agent's belief set then $K\varphi$ should be a belief too (the agent believes in her own beliefs) and if φ is not in the agent's belief set then $\neg K\varphi$ should be in the belief set (the agent disbelieves her non-beliefs). Moore proposed a way to complete collections of base facts about the world (possibly referring to agent's beliefs or disbeliefs) to belief sets, called *expansions*, that the agent might hold given the base theory.

This brief overview points to similarities in motivations and intuitions behind default and autoepistemic logics. These similarities drew attention to the issue of a formal account of the relationship between the two logics, which quickly became a subject of active research. Konolige [Kon88] proposed to translate a default

$$\frac{\alpha : \beta_1, \ldots, \beta_k}{\gamma},$$

into the autoepistemic formula:

$$K\alpha \wedge \neg K\neg\beta_1 \wedge \ldots \wedge \neg K\neg\beta_k \rightarrow \gamma.$$

The translation was an attempt to reflect the intuition that in order to apply a default, its prerequisite must be derived and its justifications must be consistent. To model the statement "α is derived", Konolige used the formula $K\alpha$. To model the statement that "β_i are consistent", Konolige used the formula $\neg K\neg\beta_i$ (which is equivalent to $M\beta_i$).

There was, however, a problem. It turned out that, while seemingly well motivated, the translation does not relate extensions and expansions. Specifically, modal counterparts (under Konolige's translation) of default theories could have expansions not corresponding to extensions. That discovery suggested a possibility that the autoepistemic logic may not be the right modal counterpart to the default logic or that the modal reading of a default proposed by Konolige is not appropriate. Thus, researchers began to look for other modal logics and for other translations. Konolige related default logic to a *version* of autoepistemic logic based on the notion of a *strongly grounded expansion* [Kon88]. Marek and Truszczyński [MT89b] proposed an alternative translation and represented extensions as expansions of a modal nonmonotonic logic constructed by the method of McDermott from the weakest modal logic N. Truszczyński [Tru91] found that the Gödel translation of intuitionistic logic to modal logic S4 could be used to translate the default logic into the nonmonotonic modal logic S4F.

Gottlob [Got95] returned to the original problem of relating default and autoepistemic logic. He described a mapping translating default theories into modal ones so that extensions correspond precisely to expansions. The problem is that his translation is not modular. The autoepistemic representation of a default theory depends on the whole theory and cannot be obtained as the union of independent translations of individual defaults. Thus, the approach of Gottlob does not provide an autoepistemic reading of an individual default. In fact, in the same paper Gottlob proved that a modular translation from default logic with the semantics of extensions to autoepistemic logic with the semantics of expansions does not exist. In conclusion, there is *no* modal interpretation of a default under which *extensions* would correspond to *expansions*.

Results of Gottlob provided strong evidence that extensions and expansions are, in some sense, essentially different. A careful examination of intuitions as well as of formal definitions

of extensions and expansions provides some further evidence to this effect. Moore's logic sometimes sanctions *unsupported* beliefs. For example, the theory $\{Kp \Rightarrow p\}$ has two expansions. One of them is generated by the set of all tautologies, the other one is generated by p. Presence of p in this latter expansion is justified only by the belief in p (by the formula Kp). In other words, the belief in p is self-supporting. In contrast, belief sets containing self-supporting beliefs are not sanctioned by the default logic. Even if a default

$$\frac{p:}{p},$$

providing self-supporting evidence for p, is included in the theory, the semantics of default logic does not make any use of it. The default theory $(\{\frac{p:}{p}\}, \emptyset)$ has only one extension and it consists of tautologies only. Thus, the autoepistemic logic of Moore could be viewed as a nonmonotonic logic of belief and the default logic of Reiter could be viewed as a nonmonotonic logic of *justified* belief.

As for the translation proposed by Konolige, it is clear that it does not relate extensions and expansions. It does, however, provide some link between default logic and autoepistemic logic. Marek and Truszczyński [MT89a] proposed the concept of a *weak extension* of a default theory and proved that under the translation of Konolige, weak extensions and expansions coincide. In other words, they proposed an alternative semantics for default theories that could be viewed as the semantics of belief, yielding a default version of the logic of belief. Thus, the Konolige's translation might be the right one, once proper semantics on each side are identified and correctly aligned.

The picture that emerges is that of a substantial research effort and several significant results but with no definite systematic account of semantics for default and autoepistemic theories and with no clear understanding of constructive principles behind the process of arriving at a belief set. There has been no satisfactory solution to the matter of the relationship between the two logics and there have been questions concerning the adequacy of the modal reading of defaults that was proposed by Konolige. In this paper we resolve all these issues.

We propose a unifying semantic treatment of default and autoepistemic logics in terms of *possible-world structures*. A possible-world structure is a collection of two-valued propositional interpretations each representing a state of the world that is *possible* according to the agent. Possible-world structures can be seen as special *Kripke structures* [HC84]. They are of fundamental importance in semantic studies of the modalities of knowledge and belief.

Possible-world structures were used in the study of autoepistemic logic. Moore described a possible-world characterization of expansions in [Moo84]. They also figured prominently in Levesque's studies of autoepistemic logic as the logic of "only knowing" [Lev90]. Possible-world structures appeared in the study of default logic but only marginally. Guerreiro and Casanova [GC90] found a characterization of extensions in terms of possible-world structures. Their work was later further expanded by Lifschitz [Lif90]. Possible-world structures in default logic were also studied by Besnard and Schaub [BS94].

In this paper, we propose a comprehensive semantic framework based on the concept of a possible-world structure and on the intuition that belief sets can be obtained in a constructive process of building their increasingly more precise approximations. To formally represent an approximation to a possible-world structure we use the concept of a *belief pair*, that is, a pair of possible-world structures one of which provides a conservative and the other one a liberal

estimate. We introduced this notion [DMT99] and used it there to study ways to approximate the semantics of expansions.

Belief pairs form a complete lattice. We model the process of revising one approximation (belief pair) to obtain another approximation (that is, another belief pair) in default and autoepistemic logics as monotone operators on the lattice of belief pairs. By selecting different fixpoints of these operators we obtain structured *families* of semantics for default and autoepistemic logics. Some of these semantics are appropriate to model the notion of belief. Others are well suited to model the concept of justified belief. Still others have a strong constructive flavor — the corresponding fixpoints can be obtained by iterating the operators over the least precise approximation. Our main contributions can be summarized as follows:

- With every modal theory T we associate an operator $\mathcal{D}_T$ defined on the lattice of belief pairs. Applying purely algebraic means to the operator $\mathcal{D}_T$, we obtain a family of semantics for T. These semantics capture different modes of reasoning in autoepistemic logic. One of them corresponds to the semantics of expansions as introduced by Moore. Another one is the semantics of justified belief that eliminates expansions containing self-supporting beliefs.

- The same approach works for the default logic! With every default theory Δ we associate an operator $\mathcal{E}_\Delta$ defined on the lattice of belief pairs. We apply to $\mathcal{E}_\Delta$ the same algebraic techniques we used in the study of the operator $\mathcal{D}_T$ and obtain a family of semantics for Δ. Among these semantics there are the semantics of weak extensions and the semantics of extensions capturing within default logic the concepts of belief and justified belief, respectively.

- The semantic operator of a default theory Δ and of the autoepistemic theory obtained by applying to Δ the translation of Konolige are identical. This fact has far reaching consequences. Konolige's translation establishes an isomorphism between the two families of semantics of default and autoepistemic logics. In particular, the meaning of a default theory under a particular semantics of default logic is identical to the meaning of its translation in the corresponding semantics of the autoepistemic logic.

  In this way, we resolve the issue of the relationship between default and autoepistemic logics and the question why Konolige's translation did not work. Default logic under the semantics of extensions and autoepistemic logic under the semantics of expansions model different modes of autoepistemic reasoning and occupy different locations in their respective families of semantics. This fact is responsible for Gottlob's result that defaults cannot be translated to autoepistemic formulas one by one. However, once we properly align different semantics of default and autoepistemic logics, we find that Konolige's translation is correct! Viewed in the context of a family of semantics, rather than in the context of a single one, default logic turns out to be just a fragment of autoepistemic logic. The original default logic with extensions can be seen as a fragment of the autoepistemic logic of justified belief. The default logic with weak extensions (expansions) is a fragment of the autoepistemic logic of belief (the original autoepistemic logic with expansions).

- We identify two different constructive semantics describing how to approximate the knowledge in an autoepistemic or default theory. They are obtained by iterating certain monotone operators on the lattice of belief pairs (operators $\mathcal{D}_T$, $\mathcal{E}_\Delta$ and two other

operators that are derived from them). This constructive process provides insights on how agents can gain information about an unknown belief set by starting with the weakest approximation possible (the bottom of the lattice of belief pairs) and by using their base theory to iteratively improve upon this approximation until further improvements are no longer possible. We show how these semantics can be used to obtain sufficient conditions for the existence of a single belief set. We also study the complexity of computing these approximation semantics and show that it is lower than the complexity of computing individual belief sets. This result may have implications for building automated reasoning systems for default and autoepistemic logics.

- Our investigations are based on algebraic considerations concerning fixpoints of operators on lattices. In that we follow the approach developed by Fitting [Fit02] and further extended in [DMT00b] to study semantics for logic programs with negation. Connections between logic programming with negation and autoepistemic and default logics were established a long time ago [Gel87, MT89b, BF91]. It turns out that the structure of most important semantics for logic programs revealed by Fitting's work is isomorphic to the structure of the semantics for autoepistemic and default logics that we derive in this paper.

The paper is organized as follows. First, in Section 2, we introduce basic logic terminology and review the semantic approach to autoepistemic logic proposed by Moore [Moo84]. In Section 3, we introduce and study the operator $\mathcal{D}_T$, defined on the lattice of belief pairs. Fixpoints of the operator $\mathcal{D}_T$ give rise to the semantics of partial expansions and expansions for autoepistemic logic. The least fixpoint of $\mathcal{D}_T$ yields the Kripke-Kleene semantics. In Section 4, we introduce *stable* operators associated with the operator $\mathcal{D}_T$. Fixpoints of these operators define several new semantics for the autoepistemic logic. In Section 5, we study semantical foundations of Reiter's default logic. We define an operator $\mathcal{E}_\Delta$ on the lattice of belief pairs, which is a default-logic counterpart of the operator $\mathcal{D}_T$. As in the case of autoepistemic logic, we derive from $\mathcal{E}_\Delta$ several other operators and show how their fixpoints describe major semantics of default theories. Exploiting the relationship between the operators $\mathcal{D}_T$ and $\mathcal{E}_\Delta$ (under the Konolige's modal interpretation of defaults), we establish in Section 6 a precise correspondence between default and autoepistemic logics and explain earlier problems with relating the two logics. In Section 7, we study the complexity of computation of suitably chosen representations of Kripke-Kleene and well-founded fixpoints and show that the corresponding decision problems are in the class Δ_2^P. The last section contains additional discussion of the results and conclusions.

2 Autoepistemic logic — preliminaries

In this section, we introduce basic logic terminology that we will use in the paper. We also recall the semantic treatment of autoepistemic logic proposed by Moore [Moo84] and studied by Levesque [Lev90].

In the paper, we consider the language of propositional logic determined by a set of propositional atoms At. We denote this language by $\mathcal{L}$. We also consider the language of modal propositional logic obtained by extending $\mathcal{L}$ with a modal operator K. We denote this language by $\mathcal{L}_K$. We call formulas in $\mathcal{L}_K$ that do not contain any occurrences of K *modal-free*

or *propositional* formulas. Of particular interest in the paper are modal formulas of the form $K\varphi$. We call them *modal atoms*. We refer to collections of modal formulas (that is, subsets of $\mathcal{L}_K$) as *modal theories*.

A *two-valued interpretation* assigns to each atom from At a truth value **t** or **f**. These two truth values, together with the ordering $\mathbf{f} \leq \mathbf{t}$, form the standard Boolean lattice of truth values. The set of all two-valued interpretations of At will be denoted by $\mathcal{A}$.

Any set $Q \subseteq \mathcal{A}$ is called a *possible-world structure* and can be viewed as a universal Kripke model with a total accessibility relation [Che80, HC84][1]. Possible-world structures constitute a basic tool in semantic studies of modal logics. As we stated earlier, they represent the agent's knowledge about the world. Possible-world structures were used by Moore [Moo84] and later by Levesque [Lev90] in the investigations of autoepistemic logic.

We denote the collection of all possible-world structures (with respect to the set of atoms At) by $\mathcal{W}$. This set can be ordered by the *reverse set inclusion* $\sqsubseteq$: for $Q_1, Q_2 \in \mathcal{W}$, $Q_1 \sqsubseteq Q_2$ if $Q_2 \subseteq Q_1$. The reason for the choice of this ordering is that if $Q_1 \sqsubseteq Q_2$, and $T_i, i = 1, 2$, are the sets of sentences of $\mathcal{L}$ true in Q_i, then $T_1 \subseteq T_2$. The ordering $\sqsubseteq$ can be thought of as a knowledge ordering. As we move up in the lattice, more and more interpretations are excluded from possible-world structures. Thus, our knowledge of the interpretation describing the actual world improves. Clearly, $\langle \mathcal{W}, \sqsubseteq \rangle$ is a complete lattice.

Example 2.1 As a running example, to illustrate concepts in the paper, we will consider the language over the set of atoms $At_p = \{p\}$. We will denote this language as $\mathcal{L}_p$. The set of interpretations $\mathcal{A}$ (denoted $\mathcal{A}_p$, in this special case) consists of two interpretations, say I_p and J_p, where $I_p(p) = \mathbf{t}$ and $J_p(p) = \mathbf{f}$. The set $\mathcal{W}$ (denoted by $\mathcal{W}_p$) has four elements: $\emptyset$, $\mathcal{I}_p = \{I_p\}$, $\mathcal{J}_p = \{J_p\}$ and $\mathcal{A}_p$. The lattice $(\mathcal{W}_p, \sqsubseteq)$ is shown in Figure 1. □

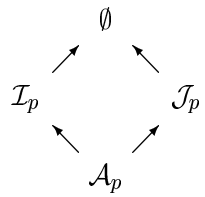

Figure 1: Lattice $(\mathcal{W}_p, \sqsubseteq)$

To study formalisms based on the modal language, we study operators on the lattice $\mathcal{W}$ and their fixpoints. We start by defining the *truth function* $\mathcal{H}_{Q,I}$ ($Q \subseteq \mathcal{A}$ is a possible-world structure, $I \in \mathcal{A}$ is an interpretation). The definition is inductive and proceeds as follows:

1. $\mathcal{H}_{Q,I}(p) = I(p)$, if p is an atom.
2. $\mathcal{H}_{Q,I}(\varphi_1 \wedge \varphi_2) = \mathbf{t}$ if $\mathcal{H}_{Q,I}(\varphi_1) = \mathbf{t}$ and $\mathcal{H}_{Q,I}(\varphi_2) = \mathbf{t}$. Otherwise, $\mathcal{H}_{Q,I}(\varphi_1 \wedge \varphi_2) = \mathbf{f}$.

[1] Usually, the universe of a Kripke model is required to be nonempty. The empty set of worlds satisfies all formulas, and thus corresponds to the inconsistent theory. At times, autoepistemic expansions are inconsistent, and allowing for an empty set of worlds, as we do in this paper, captures this case.

3. $\mathcal{H}_{Q,I}(\varphi_1 \vee \varphi_2) = \mathbf{t}$ if $\mathcal{H}_{Q,I}(\varphi_1) = \mathbf{t}$ or $\mathcal{H}_{Q,I}(\varphi_2) = \mathbf{t}$. Otherwise, $\mathcal{H}_{Q,I}(\varphi_1 \vee \varphi_2) = \mathbf{f}$.

4. $\mathcal{H}_{Q,I}(\neg\varphi) = \mathbf{t}$ if $\mathcal{H}_{Q,I}(\varphi) = \mathbf{f}$. Otherwise, $\mathcal{H}_{Q,I}(\varphi) = \mathbf{f}$.

5. $\mathcal{H}_{Q,I}(K\varphi) = \mathbf{t}$, if for every interpretation $J \in Q$, $\mathcal{H}_{Q,J}(\varphi) = \mathbf{t}$. Otherwise, $\mathcal{H}_{Q,I}(K\varphi) = \mathbf{f}$.

Let us note that the value of a modal atom $K\varphi$ given by $\mathcal{H}_{Q,I}$ does not depend on I. Thus, it is entirely determined by the possible-world structure Q. We will denote this value by $\mathcal{H}_Q(K\varphi)$. We define the *theory* of a possible-world structure Q as the set

$$Th(Q) = \{\varphi : \mathcal{H}_Q(K\varphi) = \mathbf{t}\}.$$

It is clear that every modal atom $K\varphi$ is either true or false with respect to Q. In other words, for every formula $\varphi \in \mathcal{L}_K$, its epistemic status is fully determined: it is either known in Q or it is not known in Q.

For every modal theory T, Moore [Moo84] defined an operator D_T on $\mathcal{W}$ by:

$$D_T(Q) = \{I : \mathcal{H}_{Q,I}(\varphi) = \mathbf{t}, \text{ for every } \varphi \in T\}.$$

The intuition behind this definition is as follows. The possible-world structure $D_T(Q)$ is a revision of a possible-world structure Q. This revision consists of the worlds that are acceptable given the constraints on agent's beliefs captured by T. That is, the revision consists precisely of these worlds that make all formulas in T true (in the context of Q — the current belief state). Fixpoints of the operator D_T represent "stable" belief sets — they cannot be revised any further. Moore called the theory of a fixpoint of D_T a *stable expansion* of T and proposed it as a basis of autoepistemic logic: a formal description of a belief set of a rational agent with full introspection powers reasoning from a base theory T. In this paper, we will use the term "expansion" instead of the original term "stable expansion". Somewhat abusing the notation, we will use the term "expansion" also to refer to fixpoints of the operator D_T (and not only to their theories).

Example 2.1 (cont'd). Let us consider a theory $T_p = \{Kp \Rightarrow p\}$ in the language $\mathcal{L}_p$ (we regard $\alpha \Rightarrow \beta$ as an abbreviation of $\neg\alpha \vee \beta$). We will determine the operator D_{T_p}. To this end, we will first determine the truth function $\mathcal{H}_{Q,I}$ for all possible-world structures $Q \in \mathcal{W}_p$ and all interpretations $I \in \mathcal{A}_p$. Because of the form of the theory T_p, it is enough to establish the values of $\mathcal{H}_{Q,I}$ for p, Kp and $Kp \Rightarrow p$, only. Table 1 lists for each pair Q, I, those formulas among p, Kp, $Kp \Rightarrow p$ and their negations that are true under $\mathcal{H}_{Q,I}$.

	$\emptyset$	$\mathcal{I}_p$	$\mathcal{J}_p$	$\mathcal{A}_p$
I_p	$p, Kp, Kp \Rightarrow p$	$p, Kp, Kp \Rightarrow p$	$p, \neg Kp, Kp \Rightarrow p$	$p, \neg Kp, Kp \Rightarrow p$
J_p	$\neg p, Kp, Kp \Rightarrow p$	$\neg p, Kp, \neg(Kp \Rightarrow p)$	$\neg p, \neg Kp, Kp \Rightarrow p$	$\neg p, \neg Kp, Kp \Rightarrow p$

Table 1: Truth assignment $\mathcal{H}_{Q,I}(\varphi)$.

The operation of D_{T_p} can be readily obtained from this table. In particular, for each possible-world structure Q, $D_{T_p}(Q)$ is the set of interpretations I for which the table entry (I, Q) contains $Kp \Rightarrow p$. Table 2 lists the values of the operator D_{T_p}. It follows that the theory T_p has two expansions: $\mathcal{I}_p$ and $\mathcal{A}_p$. Let us note that $\mathcal{A}_p$ is the least expansion in the knowledge ordering. □

X	$\emptyset$	$\mathcal{I}_p$	$\mathcal{J}_p$	$\mathcal{A}_p$
$D_{T_p}(X)$	$\mathcal{I}_p$	$\mathcal{I}_p$	$\mathcal{A}_p$	$\mathcal{A}_p$

Table 2: The operator D_{T_p}.

3 Autoepistemic logic — a multivalued generalization

In [DMT99], we generalized Moore's approach to the three-valued case, in which we allow for the possibility that the truth value of some modal atoms is neither **t** nor **f** but, instead, it is captured by a new truth value, *unknown* or **u**. In this section, we recall essential elements of our approach from [DMT99] and extend it to the four-valued case.

Let us consider a modal theory T. We are interested in ways to form a belief set corresponding to T or its representation in terms of some possible-world structure, say Q. However, instead of searching for direct ways to find Q, we exploit the idea of an approximation. An underestimate (or a conservative view) of Q is given by any superset P of Q. Indeed, any such superset bounds Q from below in the lattice $\mathcal{W}$ (with respect to the knowledge ordering $\sqsubseteq$). Similarly, an overestimate (or a liberal view) of Q is provided by any subset S of Q, as subsets of Q are upper bounds for Q in the lattice $\mathcal{W}$. Interpretations in P can be thought of as those that are still regarded by the agent as possible (we do not have reasons to eliminate any of them yet). Interpretations in S are those that are surely in Q — we already have established that they need to be included in the possible-world structure describing the agent's belief set. Together, P and S form an *approximation* to Q.

To study approximations we introduce the concept of a *belief pair*. A belief pair is any pair (P, S) of possible-world structures. The structure P is intuitively regarded as an underestimation, S is regarded as an overestimation. Consequently, we say that a belief pair (P, S) approximates a possible-world structure Q if $S \subseteq Q \subseteq P$ (or, in terms of the knowledge ordering, if $P \sqsubseteq Q \sqsubseteq S$). Clearly, the set of possible-world structures approximated by a belief pair (P, S) is not empty if and only if $S \subseteq P$ (equivalently, $P \sqsubseteq S$). We call such belief pairs *consistent* (intuitively, their conservative perspective is consistent with the liberal one). All other belief pairs are called *inconsistent* — they do not approximate any possible-world structures.

In [DMT99], we considered consistent belief pairs only. As a result we obtained a three-valued concept of belief set (beliefs could be true, false or undefined). In this paper we allow inconsistent belief pairs. First, the ways in which the agent establishes estimates P and S may be independent of each other and, at least at an abstract level, inconsistent belief pairs may arise. Second, admitting inconsistent belief pairs simplifies mathematical arguments and yields more elegant algebraic structures. Working in this extended setting, we propose four-valued semantics for autoepistemic logics and show that they generalize two- and three-valued semantics that were known before.

Belief pairs can be ordered by a *precision* ordering $\preceq_{pr}$. Namely, given two belief pairs (P, S) and (P', S') we define:

$$(P, S) \preceq_{pr} (P', S') \text{ if } P' \subseteq P, \text{ and } S \subseteq S',$$

or, equivalently,

$$(P, S) \preceq_{pr} (P', S') \text{ if } P \sqsubseteq P', \text{ and } S' \sqsubseteq S.$$

Let us denote by $[P, S]$ the set $\{Q \in \mathcal{W}: P \sqsubseteq Q \sqsubseteq S\}$. Clearly, if $(P, S) \preceq_{pr} (P', S')$ then

$$[P', S'] \subseteq [P, S].$$

In other words, larger (in the ordering $\preceq_{pr}$) belief pairs provide more *precise* approximations — the sets of approximated possible-world structures get smaller. This property motivates our choice of terminology.

We denote the set of all belief pairs by $\mathcal{B}$. Together with the precision ordering, the set $\mathcal{B}$ forms a complete lattice. Thus, by the theorem of Tarski and Knaster, monotone operators on $\mathcal{B}$ are guaranteed to have a least fixpoint. The lattice $(\mathcal{B}_p, \preceq_{pr})$ of belief pairs over the language $\mathcal{L}_P$ (Example 2.1) is shown in Figure 2.

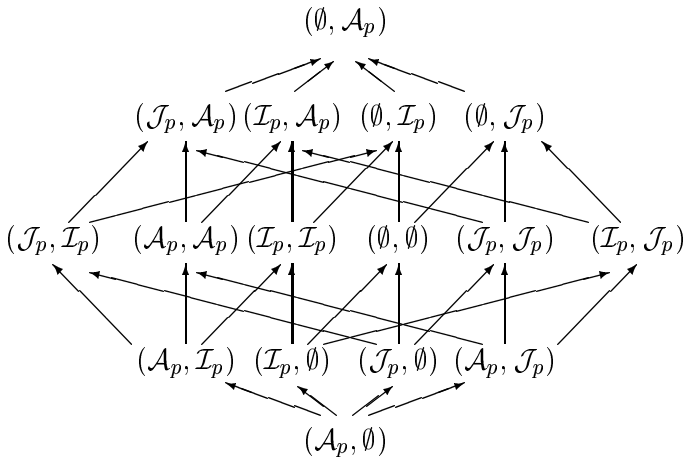

Figure 2: Lattice $(\mathcal{B}_p, \preceq_{pr})$

We will mention in passing that the set of all belief pairs can also be ordered by another ordering, $\sqsubseteq$, defined as follows:

$$(P, S) \sqsubseteq (P', S') \text{ if } P \sqsubseteq P' \text{ and } S \sqsubseteq S'.$$

We refer to this ordering as the *knowledge ordering* of $\mathcal{B}$ as it is a component-wise extension of the knowledge ordering $\sqsubseteq$ of possible-world structures (moreover, we use the same symbol to denote it). This ordering plays only a limited role in our considerations (it defines the concept of minimality used in Theorems 4.4 and 5.9) thus, we do not discuss its properties in any significant detail.

With a belief pair (P, S) and a two-valued interpretation I we associate a two-valued truth function $\mathcal{H}^2_{(P,S),I}$ defined on the set of all formulas of the modal language. Our intention is to define $\mathcal{H}^2_{(P,S),I}(\varphi)$ so that it provides a conservative estimate to the truth value of φ with respect to a belief pair (P, S). The definition follows a standard recursive pattern and the only more subtle point concerns the definition of $\mathcal{H}^2_{(P,S),I}$ for modal atoms $K\varphi$.

Before we give a formal definition, let us first consider a modal atom $K\varphi$, where φ does not contain any other occurrences of K. In the belief pair (P, S), P represents the conservative point of view. Thus, the conservative estimate for the truth value of $K\varphi$ will be obtained if

the set P is used in the evaluation: $K\varphi$ should be true according to the conservative point of view if φ is true in all valuations in P. The situation changes in the case of the formula $\neg K\varphi$ where, as before, φ is modal-free. To compute the conservative estimate for the truth value of this formula, we need to negate the *liberal* estimate for the truth value of $K\varphi$. This liberal estimate can be computed with the help of S: $K\varphi$ is true if φ is true in all interpretations from S.

This discussion can be generalized to arbitrary formulas and suggests the following approach. To obtain the conservative estimate for the truth value of a formula, modal atoms that appear *positively* in the formula must be evaluated with respect to the conservative point of view (that is with respect to the set of interpretations P). On the other hand, modal atoms that appear *negatively* must be evaluated according to the liberal perspective (that is with respect to the interpretations in S). Formally, we have the following inductive definition of $\mathcal{H}^2_{(P,S),I}$.

1. $\mathcal{H}^2_{(P,S),I}(p) = I(p)$, for every atom p.
2. $\mathcal{H}^2_{(P,S),I}(\varphi_1 \wedge \varphi_2) = \mathbf{t}$ if $\mathcal{H}^2_{(P,S),I}(\varphi_1) = \mathbf{t}$ and $\mathcal{H}^2_{(P,S),I}(\varphi_2) = \mathbf{t}$. Otherwise, $\mathcal{H}^2_{(P,S),I}(\varphi_1 \wedge \varphi_2) = \mathbf{f}$.
3. $\mathcal{H}^2_{(P,S),I}(\varphi_1 \vee \varphi_2) = \mathbf{t}$ if $\mathcal{H}^2_{(P,S),I}(\varphi_1) = \mathbf{t}$ or $\mathcal{H}^2_{(P,S),I}(\varphi_2) = \mathbf{t}$. Otherwise, $\mathcal{H}^2_{(P,S),I}(\varphi_1 \vee \varphi_2) = \mathbf{f}$.
4. $\mathcal{H}^2_{(P,S),I}(\neg\varphi) = \neg\mathcal{H}^2_{(S,P),I}(\varphi)$.
5. $\mathcal{H}^2_{(P,S),I}(K\varphi) = \mathbf{t}$ if $\mathcal{H}^2_{(P,S),J}(\varphi) = \mathbf{t}$ for all $J \in P$. Otherwise, $\mathcal{H}^2_{(P,S),I}(K\varphi) = \mathbf{f}$.

Step (4) is the key. It ensures that when evaluating the negation of a formula the roles of P and S are switched. Consequently, modal atoms appearing positively in a formula are evaluated with respect to the belief pair (P, S) and modal atoms that appear negatively are evaluated with respect to the belief pair (S, P).

Clearly, to construct a liberal estimate for the truth value of φ with respect to a belief pair (P, S) we can proceed similarly and use S (respectively, P) to evaluate modal literals appearing positively (negatively) in φ. It is easy to see, however, that the resulting truth function can be expressed as $\mathcal{H}^2_{(S,P),I}$ (we reverse the roles of P and S).

Conservative and liberal estimates of truth values of formulas can be combined into a single estimate from a four-valued Belnap's lattice of truth values. The elements of the Belnap's lattice are: $\mathbf{t}_4 = (\mathbf{t}, \mathbf{t})$ (true), $\mathbf{f}_4 = (\mathbf{f}, \mathbf{f})$ (false), $\mathbf{u} = (\mathbf{f}, \mathbf{t})$ (unknown) and $\mathbf{i} = (\mathbf{t}, \mathbf{f})$. These values are related by the following lattice order $\preceq_{pr}$ (the *precision* ordering in Belnap's lattice):

$$(u, v) \preceq_{pr} (u', v') \textit{ if } u \leq u' \textit{ and } v \geq v'$$

(let us recall that $\leq$ is the ordering of truth values $\mathbf{f}$ and $\mathbf{t}$, and that $\mathbf{f} \leq \mathbf{t}$). The Belnap's lattice is shown in Figure 3.

An element of the Belnap's lattice can be viewed as an approximation to an unknown two-valued truth value. Clearly, the higher we are in the Belnap's lattice, the more precise is the approximation. At the bottom both $\mathbf{t}$ and $\mathbf{f}$ are possible (thus, the term "unknown" for the truth value $\mathbf{u}$). Each of the approximations at the second level represent exactly one truth

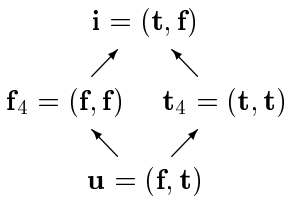

Figure 3: The Belnap's lattice

value. Lastly, no two-valued truth value is represented by the top element. It is due to these "precision of approximation" intuitions (similar to intuitions underlying the precision ordering of belief pairs), that we refer to the ordering $\preceq_{pr}$ as the *precision* ordering[2].

On the elements of the Belnap lattice one can define the negation operator:

$$\neg \mathbf{f}_4 = \mathbf{t}_4, \quad \neg \mathbf{t}_4 = \mathbf{f}_4, \quad \neg \mathbf{u} = \mathbf{u}, \quad \neg \mathbf{i} = \mathbf{i}.$$

Under this definition, for every element (x, y) of the Belnap's lattice we have:

$$\neg(x, y) = (\neg y, \neg x),$$

where the negation operator on the right-hand side is the standard two-valued negation operator.

Using the Belnap lattice of truth values, we now define the four-valued truth function combining lower and upper estimates $\mathcal{H}^2_{(P,S),I}(\varphi)$ and $\mathcal{H}^2_{(S,P),I}(\varphi)$ as follows:

$$\mathcal{H}^4_{(P,S),I}(\varphi) = (\mathcal{H}^2_{(P,S),I}(\varphi), \mathcal{H}^2_{(S,P),I}(\varphi)).$$

Directly from this definition, it follows that

$$\mathcal{H}^4_{(P,S),I}(\neg\varphi) = \neg\mathcal{H}^4_{(P,S),I}(\varphi),$$

where the negation operator on the right-hand side is the negation operator in the Belnap's lattice.

It is clear that for a modal atom $K\varphi$, the logical values $\mathcal{H}^2_{(P,S),I}(K\varphi)$ and $\mathcal{H}^4_{(P,S),I}(K\varphi)$ do not depend on the interpretation I. Therefore, we will denote them by $\mathcal{H}^2_{(P,S)}(K\varphi)$ and $\mathcal{H}^4_{(P,S)}(K\varphi)$, respectively.

Example 2.1 (cont'd). To illustrate the concepts we just introduced we will now evaluate $\mathcal{H}^4_{(P,S),I_p}(Kp \Rightarrow p)$ and $\mathcal{H}^4_{(P,S),J_p}(Kp \Rightarrow p)$ for all belief pairs $(P, S) \in \mathcal{B}_p$. First, let us notice that for every interpretation I,

$$\begin{aligned}\mathcal{H}^4_{(P,S),I}(Kp \Rightarrow p) &= (\mathcal{H}^2_{(P,S),I}(\neg Kp \vee p),\ \mathcal{H}^2_{(S,P),I}(\neg Kp \vee p)) = \\ &\quad (\mathcal{H}^2_{(P,S),I}(\neg Kp) \vee I(p),\ \mathcal{H}^2_{(S,P),I}(\neg Kp) \vee I(p)) = \\ &\quad (\neg\mathcal{H}^2_{(S,P)}(Kp) \vee I(p),\ \neg\mathcal{H}^2_{(P,S)}(Kp) \vee I(p)).\end{aligned}$$

[2]As in the case of belief pairs, there is an alternative ordering of the elements in the Belnap's lattice. It is a component-wise extension of the standard ordering of the truth values **f** and **t**.

Since $I_p(p) = \mathbf{t}$, it follows that for every belief pair (P, S),

$$\mathcal{H}^4_{(P,S),I_p}(Kp \Rightarrow p) = (\mathbf{t}, \mathbf{t}) = \mathbf{t}_4.$$

Similarly, since $J_p(p) = \mathbf{f}$, we have

$$\mathcal{H}^4_{(P,S),J_p}(Kp \Rightarrow p) = (\neg\mathcal{H}^2_{(S,P)}(Kp), \neg\mathcal{H}^2_{(P,S)}(Kp)) = \neg\mathcal{H}^4_{(P,S)}(Kp).$$

Let us consider the belief pair $(\mathcal{A}_p, \mathcal{I}_p)$. Since $\mathcal{A}_p$ contains worlds in which p is false, a conservative estimate (an underestimate) for the truth value of Kp is given by $\mathbf{f}$. On the other hand, since p is true in all worlds in $\mathcal{I}_p$, the liberal estimate (an overestimate) for this truth value is $\mathbf{t}$. Thus, $\mathcal{H}^4_{(\mathcal{A}_p,\mathcal{I}_p)}(Kp) = (\mathbf{f}, \mathbf{t}) = \mathbf{u}$. Consequently, $\mathcal{H}^4_{(\mathcal{A}_p,\mathcal{I}_p),J_p}(Kp \Rightarrow p) = \neg\mathcal{H}^4_{(\mathcal{A}_p,\mathcal{I}_p)}(Kp) = \mathbf{u}$.

	$\emptyset$	$\mathcal{I}_p$	$\mathcal{J}_p$	$\mathcal{A}_p$
$\emptyset$	$\mathbf{f}_4$	$\mathbf{f}_4$	$\mathbf{i}$	$\mathbf{i}$
$\mathcal{I}_p$	$\mathbf{f}_4$	$\mathbf{f}_4$	$\mathbf{i}$	$\mathbf{i}$
$\mathcal{J}_p$	$\mathbf{u}$	$\mathbf{u}$	$\mathbf{t}_4$	$\mathbf{t}_4$
$\mathcal{A}_p$	$\mathbf{u}$	$\mathbf{u}$	$\mathbf{t}_4$	$\mathbf{t}_4$

Table 3: The truth values $\mathcal{H}^4_{(P,S),J_p}(Kp \Rightarrow p)$.

The values $\mathcal{H}^4_{(P,S),J_p}(Kp \Rightarrow p)$ for the remaining belief pairs can be computed in the same way by computing $\neg\mathcal{H}^4_{(P,S)}(Kp)$. They are all listed in Table 3. The value $\mathcal{H}^4_{(P,S),J_p}(Kp \Rightarrow p)$ is given at the intersection of the row labeled with P and the column labeled with S. □

Using the truth function $\mathcal{H}^4_{(P,S)}$ we can associate with each belief pair (P, S) its *epistemic content*. We define the *knowledge* of (P, S), denoted $kn(P, S)$, by

$$kn(P, S) = \{\varphi \in \mathcal{L}_K : \mathcal{H}^4_{(P,S)}(K\varphi) = \mathbf{t}_4\}.$$

Similarly, we define the *ignorance* of (P, S), denoted $ig(P, S)$, by

$$ig(P, S) = \{\varphi \in \mathcal{L}_K : \mathcal{H}^4_{(P,S)}(K\varphi) = \mathbf{f}_4\}.$$

The set $kn(P, S)$ can be viewed as the set of formulas that are known in the belief pair (P, S). The set $ig(P, S)$ can be regarded as the set of formulas that are unknown in the belief pair (P, S).

By the *meta-knowledge* of a belief pair (P, S) we mean the set of those formulas whose epistemic status is determined: the corresponding belief is either true or false (as opposed to unknown or inconsistent). These are precisely the formulas in $kn(P, S) \cup ig(P, S)$. We call the set of all other formulas, $\mathcal{L}_K \setminus (kn(P, S) \cup ig(P, S))$, the *meta-ignorance* of (P, S) as their epistemic status is not clear.

The concept of a belief pair generalizes that of a possible-world structure and the truth function $\mathcal{H}^4_{(P,S),I}$ is a four-valued generalization of the two-valued truth function $\mathcal{H}_{P,I}$. Moreover, the concept of knowledge of a belief pair generalizes the notion of the theory of a possible-world structure. We define a *complete* belief pair to be any belief pair of the form (P, P). We have the following result (the proof is self-evident and we omit it).

Proposition 3.1 *Let P be a possible-world structure. Then*

1. *For every formula $\varphi \in \mathcal{L}_K$ and every interpretation $I \in \mathcal{A}$, $\mathcal{H}^2_{(P,P),I}(\varphi) = \mathcal{H}_{P,I}(\varphi)$ and $\mathcal{H}^4_{(P,P),I}(\varphi) = (\mathcal{H}_{P,I}(\varphi), \mathcal{H}_{P,I}(\varphi))$.*
2. *$kn(P,P) = \{\varphi\colon \mathcal{H}_P(K\varphi) = \mathbf{t}\} = Th(P)$.*
3. *$kn(P,P) \cup ig(P,P) = \mathcal{L}_K$.*

The last assertion of Proposition 3.1 states that the meta-knowledge of a complete belief pair is complete — the epistemic status of each formula φ is fully determined: the logical value of knowing the formula (the logical value of the modal atom $K\varphi$) is either $\mathbf{t}_4$ or $\mathbf{f}_4$.

We will now study some basic properties of the truth functions $\mathcal{H}^2_{(P,S),I}$ and $\mathcal{H}^4_{(P,S),I}$ involving the orderings of truth values in the Boolean and Belnap lattices.

Proposition 3.2 *Let (P,S) and (P',S') be belief pairs from $\mathcal{B}$ such that $(P,S) \preceq_{pr} (P',S')$. Then, for every interpretation $I \in \mathcal{A}$ and for every modal formula φ we have:*

1. *$\mathcal{H}^2_{(P,S),I}(\varphi) \leq \mathcal{H}^2_{(P',S'),I}(\varphi)$.*
2. *$\mathcal{H}^4_{(P,S),I}(\varphi) \preceq_{pr} \mathcal{H}^4_{(P',S'),I}(\varphi)$.*

Proof: It is clear that statement (1) implies statement (2). So, we prove statement (1) only. We proceed by induction on the complexity of the formula φ. The induction base is obvious. The cases of $\varphi = \psi_1 \wedge \psi_2$ and $\varphi = \psi_1 \vee \psi_2$ follow immediately from the fact that the operators $\wedge$ and $\vee$ are monotone with respect to $\leq$.

Next, we will consider the case $\varphi = K\psi$. If $\mathcal{H}^2_{(P,S),I}(K\psi) = \mathbf{f}$, the inequality (1) follows. So, let us assume that $\mathcal{H}^2_{(P,S),I}(K\psi) = \mathbf{t}$. Let $J \in P'$. Since $(P,S) \preceq_{pr} (P',S')$, we have that $P' \subseteq P$. Thus, $J \in P$ and, consequently, $\mathcal{H}^2_{(P,S),J}(\psi) = \mathbf{t}$. By the induction hypothesis it follows that $\mathcal{H}^2_{(P,S),J}(\psi) \leq \mathcal{H}^2_{(P',S'),J}(\psi)$. Hence, $\mathcal{H}^2_{(P',S'),J}(\psi) = \mathbf{t}$. Since J is an arbitrary element of P', we obtain that $\mathcal{H}^2_{(P',S'),I}(K\psi) = \mathbf{t}$. Thus, the inequality (1) follows in the case $\mathcal{H}^2_{(P,S),I}(K\psi) = \mathbf{t}$, too.

Finally, we will consider the case when $\varphi = \neg\psi$. As before, it is enough to consider the case when $\mathcal{H}^2_{(P,S),I}(\varphi) = \mathbf{t}$. In this case, we have that $\mathcal{H}^2_{(S,P),I}(\psi) = \mathbf{f}$. Moreover, since $(P,S) \preceq_{pr} (P',S')$, we also have that $(S',P') \preceq_{pr} (S,P)$. Thus, by the induction hypothesis, $\mathcal{H}^2_{(S',P'),I}(\psi) \leq \mathcal{H}^2_{(S,P),I}(\psi) = \mathbf{f}$. It follows that $\mathcal{H}^2_{(S',P'),I}(\psi) = \mathbf{f}$ and, consequently, $\mathcal{H}^2_{(P',S'),I}(\varphi) = \mathbf{t}$. Hence, the inequality (1) holds for $\varphi = \neg\psi$. □

This result has several interesting corollaries. The first of them is that for a consistent belief pair (P,S), the truth function $\mathcal{H}^4_{(P,S),I}$ assigns only consistent truth values.

Corollary 3.3 *If (P,S) is a consistent belief pair then for every interpretation $I \in \mathcal{A}$ and every formula φ, $\mathcal{H}^4_{(P,S),I}(\varphi)$ is consistent (that is, $\mathcal{H}^4_{(P,S),I}(\varphi) \neq \mathbf{i}$).*

Proof: It is easy to see that if (P,S) is consistent, then:

$$(P,S) \preceq_{pr} (S,P).$$

By Proposition 3.2(1), $\mathcal{H}^2_{(P,S),I}(\varphi) \leq \mathcal{H}^2_{(S,P),I}(\varphi)$, and hence $\mathcal{H}^4_{(P,S),I}(\varphi)$ is consistent. □

The next corollary is concerned with the concept of the epistemic content of a belief pair. We show that, under the restriction to consistent belief pairs, the notion is monotone with respect to the ordering $\preceq_{pr}$.

Corollary 3.4 *Let B and B' be consistent belief pairs. If $B \preceq_{pr} B'$ then*

$$kn(B) \subseteq kn(B') \quad \textit{and} \quad ig(B) \subseteq ig(B').$$

Proof: Let us consider a formula $\varphi \in kn(B)$. Then, we have that $\mathcal{H}^4_B(K\varphi) = \mathbf{t}_4$. Since B' is consistent, $\mathcal{H}^4_{B'}(K\varphi) \neq \mathbf{i}$ (Corollary 3.3). Further, since $B \preceq_{pr} B'$, $\mathcal{H}^4_B(K\varphi) \preceq_{pr} \mathcal{H}^4_{B'}(K\varphi)$. Thus, $\mathcal{H}^4_{B'}(K\varphi) = \mathbf{t}_4$ and, consequently, $\varphi \in kn(B')$. The proof in the case when $\varphi \in ig(B)$ is similar. □

We will now study operators on the lattice $(\mathcal{B}, \preceq_{pr})$ and their properties. We will focus on operators that are monotone with respect to $\preceq_{pr}$ ($\preceq_{pr}$-monotone, for short). Each such an operator has a $\preceq_{pr}$-least fixpoint by the theorem of Tarski and Knaster. In addition to monotonicity, we will impose on operators one more condition, symmetry. An operator O on $\mathcal{B}$ is *symmetric* if for every belief pairs (P, S) and (P', S')

$$O(P, S) = (P', S') \text{ if and only if } O(S, P) = (S', P').$$

Proposition 3.5 *Let O be an operator on the lattice $\mathcal{B}$ that is $\preceq_{pr}$-monotone and symmetric.*

1. *For every consistent belief pair B, $O(B)$ is consistent.*
2. *The least fixpoint of O is consistent. Moreover, if it is complete, it is a unique fixpoint of O.*

Proof: (1) Let $B = (P, S)$ be a consistent belief pair. Then, $P \sqsubseteq S$ and, consequently, $(P, S) \preceq_{pr} (S, P)$. By the $\preceq_{pr}$-monotonicity of O, $O(P, S) \preceq_{pr} O(S, P)$. Let $O(P, S) = (P', S')$. By the symmetry of O, $O(S, P) = (S', P')$. Thus, $(P', S') \preceq_{pr} (S', P')$ and, consequently, $P' \sqsubseteq S'$. That is, $O(P, S)$ is consistent.
(2) Let us denote the least fixpoint of O by (P, S). Since O is symmetric, (S, P) is also a fixpoint of O. Thus, $(P, S) \preceq_{pr} (S, P)$. Consequently, $P \sqsubseteq S$ and (P, S) is consistent.

Let us now assume that (P, S) is complete. Let (P', S') be a fixpoint of O. By the symmetry of O, (S', P') is also a fixpoint of O. Since (P, S) is the least fixpoint of O and since $P = S$ (by the completeness of (P, S)), we obtain

$$(P, P) \preceq_{pr} (P', S') \text{ and } (P, P) \preceq_{pr} (S', P').$$

The first relation implies that $P' \subseteq P$ and $P \subseteq S'$. The second relation implies that $S' \subseteq P$ and $P \subseteq P'$. Thus, $P' = S' = P$ or, equivalently, $(P', S') = (P, S)$. □

The next result concerning fixpoints of $\preceq_{pr}$-monotone and symmetric operators on $\mathcal{B}$ shows that the knowledge and ignorance of the least fixpoint of an operator can be used to approximate the knowledge and ignorance of any other fixpoint.

Proposition 3.6 *Let B be the least fixpoint of a $\preceq_{pr}$-monotone and symmetric operator O defined on the lattice $(\mathcal{B}, \preceq_{pr})$. For every fixpoint B' of O we have:*

$$kn(B) \subseteq kn(B') \quad \textit{and} \quad ig(B) \subseteq ig(B').$$

Proof: Let us assume that $B = (P, S)$ and $B' = (P', S')$. Let us consider a formula $\varphi \in kn(B)$ and an interpretation $I \in P$. We have $\mathcal{H}^4_{(P,S),I}(\varphi) = \mathbf{t}_4$. Since (P, S) is the least fixpoint of O, $(P, S) \preceq_{pr} (P', S')$. By Proposition 3.2, it follows that $\mathcal{H}^4_{(P',S'),I}(\varphi) = \mathbf{t}_4$ or $\mathcal{H}^4_{(P',S'),I}(\varphi) = \mathbf{i}$. Let us assume that $\mathcal{H}^4_{(P',S'),I}(\varphi) = \mathbf{i}$. Then, $\mathcal{H}^4_{(S',P'),I}(\varphi) = \mathbf{u}$. However, by the symmetry of O, (S', P') is also a fixpoint of O. It follows that $(P, S) \preceq_{pr} (S', P')$ and

$$\mathbf{t}_4 = \mathcal{H}^4_{(P,S),I}(\varphi) \preceq_{pr} \mathcal{H}^4_{(S',P'),I}(\varphi) = \mathbf{u},$$

a contradiction. Consequently, $\mathcal{H}^4_{(P',S'),I}(\varphi) = \mathbf{t}_4$. Since I is an arbitrary element of P and $P' \subseteq P$, it follows that $\mathcal{H}^4_{(P',S')}(K\varphi) = \mathbf{t}_4$. Thus, $\varphi \in kn(B')$ and $kn(B) \subseteq kn(B')$, as claimed. The other inclusion can be proved in a similar fashion. □

This result is related to Corollary 3.4. We do not require here that belief pairs B and B' be consistent. Instead, we require that one of them is a least fixpoint (and so, it is consistent), and another one is an arbitrary fixpoint (possibly inconsistent) of a $\preceq_{pr}$-monotone and symmetric operator on the lattice $(\mathcal{B}, \preceq_{pr})$.

Let T be a modal theory. We will now associate with T an operator on the lattice $\mathcal{B}$. Let (P, S) be a belief pair. Extending the definition from [DMT99], we set

$$\mathcal{D}_T(P, S) = (\mathcal{D}^l_T(P, S), \mathcal{D}^u_T(P, S)),$$

where

$$\mathcal{D}^l_T(P, S) = \{I\colon \mathcal{H}^2_{(S,P),I}(T) = \mathbf{t}\} \text{ and } \mathcal{D}^u_T(P, S) = \{I\colon \mathcal{H}^2_{(P,S),I}(T) = \mathbf{t}\},$$

and where $\mathcal{H}^2_{B,I}(T)$ stands for the greatest lower bound of the set $\{\mathcal{H}^2_{B,I}(\varphi)\colon \varphi \in T\}$ (in other words, $\mathcal{H}^2_{B,I}(T) = \mathbf{t}$ if and only if $\mathcal{H}^2_{B,I}(\varphi) = \mathbf{t}$ for every $\varphi \in T$). We refer to fixpoints of the operator $\mathcal{D}_T$ as *partial expansions*.

Intuitively, the operator $\mathcal{D}_T$ describes how an agent might revise a belief pair (P, S). The objective is to obtain a new underestimate P' and a new overestimate S'. Given the current belief pair (P, S), the agent can exclude from P', as definitely impossible, all these interpretations in which at least one formula in T is false even according to the liberal estimate of truth values. All other must still be regarded as possible and included in P'. Thus, P' consists of all those interpretations for which all formulas from T are true according to the liberal estimates of truth values (given the current approximation (P, S)). To construct S' (an overestimate) the agent includes in S' only those interpretations that the agent is certain should be included, given the knowledge captured by the current belief pair (P, S). Thus, the agent includes in S' all those interpretations which make all formulas in T true even according to conservative estimates.

Example 2.1 (cont'd). We will compute $\mathcal{D}^l_{T_p}(X, \mathcal{I}_p)$ for all possible-world structures $X \in \mathcal{W}_p$. Let us observe that

$$\mathcal{H}^2_{(\mathcal{I}_p,\mathcal{A}_p)}(\neg Kp) = \neg\mathcal{H}^2_{(\mathcal{A}_p,\mathcal{I}_p)}(Kp) = \mathbf{t}.$$

Thus,

$$\mathcal{D}^l_{T_p}(\mathcal{A}_p, \mathcal{I}_p) = \{I\colon \mathcal{H}^2_{(\mathcal{I}_p,\mathcal{A}_p),I}(\neg Kp \vee p) = \mathbf{t}\} = \mathcal{A}_p.$$

Similarly,

$$\mathcal{H}^2_{(\mathcal{I}_p,\emptyset)}(\neg Kp) = \neg\mathcal{H}^2_{(\emptyset,\mathcal{I}_p)}(Kp) = \mathbf{f}.$$

Thus,

$$\mathcal{D}^l_{T_p}(\emptyset, \mathcal{I}_p) = \{I : \mathcal{H}^2_{(\mathcal{I}_p,\emptyset),I}(\neg Kp \vee p) = \mathbf{t}\} = \mathcal{I}_p.$$

The remaining values $\mathcal{D}^l_{T_p}(X, \mathcal{I}_p)$ can be computed in the same fashion. They are all are shown in Table 4.

X	$\emptyset$	$\mathcal{I}_p$	$\mathcal{J}_p$	$\mathcal{A}_p$
$\mathcal{D}^l_{T_p}(X, \mathcal{I}_p)$	$\mathcal{I}_p$	$\mathcal{I}_p$	$\mathcal{A}_p$	$\mathcal{A}_p$

Table 4: The operator $\mathcal{D}^l_{T_p}(\cdot, \mathcal{I}_p)$.

Proceeding in a similar way, we can also compute values $\mathcal{D}^u_{T_p}(X, \mathcal{I}_p)$. First, it is easy to see that for every $X \in \mathcal{W}_p$,

$$\mathcal{H}^2_{(X,\mathcal{I}_p)}(\neg Kp) = \neg\mathcal{H}^2_{(\mathcal{I}_p,X)}(Kp) = \mathbf{f}.$$

Thus, for every $X \in \mathcal{W}_p$, $\mathcal{D}^u_{T_p}(X_p, \mathcal{I}_p) = \mathcal{I}_p$ (Table 5). □

X	$\emptyset$	$\mathcal{I}_p$	$\mathcal{J}_p$	$\mathcal{A}_p$
$\mathcal{D}^u_{T_p}(X, \mathcal{I}_p)$	$\mathcal{I}_p$	$\mathcal{I}_p$	$\mathcal{I}_p$	$\mathcal{I}_p$

Table 5: The operator $\mathcal{D}^u_{T_p}(\cdot, \mathcal{I}_p)$.

The operator $\mathcal{D}_T$ plays a fundamental role in our study of autoepistemic logic. It allows us to derive all major semantics of autoepistemic theories in two-valued, three-valued and four-valued settings. We will first briefly discuss how one can reconstruct from the operator $\mathcal{D}_T$ the two-valued approach of Moore and his semantics of expansions.

Proposition 3.7 *Let T be a modal theory. Then, for every possible-world structure P, we have $\mathcal{D}_T(P, P) = (D_T(P), D_T(P))$. Consequently, a complete belief pair (P, P) is a fixpoint of $\mathcal{D}_T$ if and only if P is a fixpoint of D_T.*

Proof: The equality $\mathcal{D}_T(P, P) = (D_T(P), D_T(P))$ follows directly from Proposition 3.1 and the definitions of the operators $\mathcal{D}_T$ and D_T.

Let us assume that $D_T(P) = P$. Then

$$\mathcal{D}_T(P, P) = (D_T(P), D_T(P)) = (P, P).$$

Conversely, if $\mathcal{D}_T(P, P) = (P, P)$, then $(D_T(P), D_T(P)) = \mathcal{D}_T(P, P) = (P, P)$. Thus, $D_T(P) = P$. □

Example 2.1 (cont'd). Since $\mathcal{I}_p$ and $\mathcal{A}_p$ are fixpoints of the operator D_{T_p}, by Proposition 3.7 the belief pairs $(\mathcal{I}_p, \mathcal{I}_p)$ and $(\mathcal{A}_p, \mathcal{A}_p)$ are fixpoints of the operator $\mathcal{D}_{T_p}$. From the results summarized in Tables 4 and 5, it follows that $\mathcal{D}_{T_p}$ has two more fixpoints: $(\mathcal{A}_p, \mathcal{I}_p)$ and $(\mathcal{I}_p, \mathcal{A}_p)$. These fixpoints are not complete. The first of them is consistent, the other one is not. □

Next, we observe that two belief pairs that define the same truth value for all modal atoms occurring in theory T, that is, are epistemically equivalent, are revised by the operator $\mathcal{D}_T$ into the same belief pair. Formally, we have the following result.

Proposition 3.8 *Let T be a modal theory and let (P, S) and (P', S') be belief pairs such that for every modal atom $K\psi$ of T, $\mathcal{H}^4_{(P,S)}(K\psi) = \mathcal{H}^4_{(P',S')}(K\psi)$. Then, $\mathcal{D}_T(P, S) = \mathcal{D}_T(P', S')$.*

Proof: Clearly,

$$\mathcal{H}^2_{(P,S)}(K\psi) = \mathcal{H}^2_{(P',S')}(K\psi) \text{ and } \mathcal{H}^2_{(S,P)}(K\psi) = \mathcal{H}^2_{(S',P')}(K\psi).$$

Thus, for every interpretation I and every formula $\varphi \in T$,

$$\mathcal{H}^2_{(P,S),I}(\varphi) = \mathcal{H}^2_{(P',S'),I}(\varphi) \text{ and } \mathcal{H}^2_{(S,P),I}(\varphi) = \mathcal{H}^2_{(S',P'),I}(\varphi).$$

Therefore,

$$\mathcal{D}^l_T(P, S) = \mathcal{D}^l_T(P', S') \text{ and } \mathcal{D}^u_T(P, S) = \mathcal{D}^u_T(P', S').$$

Consequently, $\mathcal{D}_T(P, S) = \mathcal{D}_T(P', S')$. □

The next result is of fundamental importance. It asserts that the operator $\mathcal{D}_T$ is $\preceq_{pr}$-monotone and symmetric. Thus, by the theorem of Tarski and Knaster, it has a unique least fixpoint. In addition, Propositions 3.5 and 3.6 apply to $\mathcal{D}_T$.

Proposition 3.9 *The operator $\mathcal{D}_T$ is symmetric and $\preceq_{pr}$-monotone.*

Proof: Directly from the definitions it follows that $\mathcal{D}^l_T(P, S) = \mathcal{D}^u_T(S, P)$, hence $\mathcal{D}_T$ is symmetric.

To prove the monotonicity part of the claim, let us consider two belief pairs (P, S) and (P', S') such that $(P, S) \preceq_{pr} (P', S')$. We need to prove that

$$\mathcal{D}^l_T(P', S') \subseteq \mathcal{D}^l_T(P, S) \text{ and } \mathcal{D}^u_T(P, S) \subseteq \mathcal{D}^u_T(P', S')$$

Let $I \in \mathcal{D}^l_T(P', S')$. Then, $\mathcal{H}^2_{(S',P'),I}(T) = \mathbf{t}$. Since $(P, S) \preceq_{pr} (P', S')$, $(S', P') \preceq_{pr} (S, P)$. Thus, by Proposition 3.2, $\mathcal{H}^2_{(S,P),I}(T) = \mathbf{t}$ and, consequently, $I \in \mathcal{D}^l_T(P, S)$. The second inclusion can be proved in the same manner. □

Corollary 3.3 and Propositions 3.5 and 3.9 provide a connection between the approach in this paper and our earlier work [DMT99]. Corollary 3.3 shows that for a *consistent* belief pair (P, S), the truth function $\mathcal{H}^4_{(P,S),I}$ is three-valued. In fact, one can check that if (P, S) is a consistent belief pair, then the truth function $\mathcal{H}^4_{(P,S),I}$ coincides with the three-valued truth function considered in [DMT99]. Proposition 3.5 implies that the operator $\mathcal{D}_T$ maps consistent belief pairs into consistent belief pairs. One can check that the restriction of $\mathcal{D}_T$ to consistent belief pairs coincides with the operator on consistent belief pairs considered in [DMT99]. Thus, the approach developed in this paper, admitting the possibility of inconsistent belief pairs, is a generalization of the approach from [DMT99].

As we noticed earlier, Proposition 3.9 implies that the operator $\mathcal{D}_T$ has a unique $\preceq_{pr}$-least fixpoint. We denote it by $KK(T)$ and refer to it as the *Kripke-Kleene fixpoint* for T. Similarly, we call the semantics it defines *Kripke-Kleene semantics* for T. In this semantics a formula φ has logical value v (where v is from the Belnap lattice) if $\mathcal{H}^4_{KK(T)}(K\varphi) = v$. This choice of terms is motivated by a close analogy between the least fixpoint of the operator $\mathcal{D}_T$ and the Kripke-Kleene semantics for logic programs (see Section 8).

The Kripke-Kleene fixpoint has a clear constructive flavor. It can be obtained by iterating the operator $\mathcal{D}_T$, starting at the least informative belief pair, $(\mathcal{A}, \emptyset)$.

Example 2.1 (cont'd). We will now find the least fixpoint of the operator $\mathcal{D}_{T_p}$ (the least partial expansion of T_p). We start by computing $\mathcal{D}_{T_p}(\mathcal{A}_p, \emptyset)$. Let us observe that

$$\mathcal{H}^2_{(\emptyset,\mathcal{A}_p),I}(\neg Kp) = \neg\mathcal{H}^2_{(\mathcal{A}_p,\emptyset),I}(Kp) = \mathbf{t}.$$

Thus,

$$\mathcal{D}^l_{T_p}(\mathcal{A}_p, \emptyset) = \{I\colon \mathcal{H}^2_{(\emptyset,\mathcal{A}_p),I}(\neg Kp \vee p) = \mathbf{t}\} = \mathcal{A}_p.$$

In a similar fashion,

$$\mathcal{H}^2_{(\mathcal{A}_p,\emptyset),I}(\neg Kp) = \neg\mathcal{H}^2_{(\emptyset,\mathcal{A}_p),I}(Kp) = \mathbf{f}.$$

Consequently,

$$\mathcal{D}^u_{T_p}(\mathcal{A}_p, \emptyset) = \{I\colon \mathcal{H}^2_{(\mathcal{A}_p,\emptyset),I}(\neg Kp \vee p) = \mathbf{t}\} = \mathcal{I}_p.$$

Thus, $\mathcal{D}_{T_p}(\mathcal{A}_p, \emptyset) = (\mathcal{A}_p, \mathcal{I}_p)$. We already showed earlier that $\mathcal{D}_{T_p}(\mathcal{A}_p, \mathcal{I}_p) = (\mathcal{A}_p, \mathcal{I}_p)$ (see Tables 4 and 5). Thus, $(\mathcal{A}_p, \mathcal{I}_p)$ is the least fixpoint (Kripke-Kleene fixpoint) of $\mathcal{D}_{T_p}$. □

We now summarize basic properties of the fixpoint $KK(T)$ as a corollary to Propositions 3.5, 3.6, and 3.9.

Corollary 3.10 *Let T be a modal theory.*

1. *The fixpoint $KK(T)$ is consistent.*
2. *For every partial expansion B of T, $KK(T) \preceq_{pr} B$.*
3. *For every partial expansion B of T,*

$$kn(KK(T)) \subseteq kn(B) \quad \textit{and} \quad ig(KK(T)) \subseteq ig(B).$$

4. *If $KK(T)$ is a complete belief pair, then it is the unique consistent partial expansion of T. Moreover the possible-world structure P such that $KK(T) = (P, P)$ is the unique expansion of T.*

Corollary 3.10 has important epistemological consequences. It states that the Kripke-Kleene fixpoint is a consistent belief pair that approximates belief sets that are formalized as fixpoints of the operator $\mathcal{D}_T$. In other words, the iterative approximation process is sound. Next, it demonstrates how the knowledge and ignorance of the Kripke-Kleene fixpoint approximates that of all other partial expansions of T. Lastly, it implies that the Kripke-Kleene semantics provides sufficient conditions for the uniqueness of an expansion of T. Corollary 3.10 has also computational implications. We discuss them later in Section 7.

4 Autoepistemic logic — extensions and the well-founded semantics

In this section we show that the theory of belief pairs allows us to introduce new semantics for autoepistemic logic. Given a modal theory T, we use the operator $\mathcal{D}_T$ to define two additional operators: the operator D^{st}_T defined on the lattice $\mathcal{W}$, and the operator $\mathcal{D}^{st}_T$ defined on the lattice $\mathcal{B}$. They give rise to semantics for autoepistemic logic that are closely related to the semantics

of extensions for default logic. One of them, the semantics obtained by means of fixpoints of the operator D^{st}_T, is a perfect match to Reiter's semantics of extensions for default logic, an object long sought after in the autoepistemic logic. The operator $\mathcal{D}^{st}_T$ is $\preceq_{pr}$- monotone, and its least fixpoint gives rise to the well-founded semantics for autoepistemic logic.

Let S be a possible-world structure, that is, $S \in \mathcal{W}$. For a possible-world structure $P \in \mathcal{W}$ we define

$$D_{S,T}(P) = \mathcal{D}^l_T(P, S).$$

The operator $D_{S,T}$ is a monotone operator on the lattice $(\mathcal{W}, \sqsubseteq)$. Indeed, if $P_1 \sqsubseteq P_2$, then $(P_1, S) \preceq_{pr} (P_2, S)$. By the $\preceq_{pr}$-monotonicity of $\mathcal{D}_T$, $\mathcal{D}_T(P_1, S) \preceq_{pr} \mathcal{D}_T(P_2, S)$. Thus,

$$D_{S,T}(P_1) = \mathcal{D}^l_T(P_1, S) \sqsubseteq \mathcal{D}^l_T(P_2, S) = D_{S,T}(P_2).$$

By the theorem of Tarski and Knaster, the operator $D_{S,T}$ has a least fixpoint. We define

$$D^{st}_T(S) = \mathit{lfp}(D_{S,T}) = \mathit{lfp}(\mathcal{D}^l_T(\cdot, S)).$$

Intuitively, $D^{st}_T(S)$ can be viewed as a preferred conservative estimate of what is believed given a fixed S (that is, given a fixed liberal estimate on beliefs)[3].

In a similar way as for $\mathit{lfp}(\mathcal{D}^u_T(\cdot, S))$, we argue that $\mathit{lfp}(\mathcal{D}^u_T(P, \cdot))$ can be regarded as a preferred liberal estimate of what is believed, given a fixed conservative point of view. Let us notice that by the symmetry of the operator $\mathcal{D}_T$,

$$\mathcal{D}^u_T(P, S) = \mathcal{D}^l_T(S, P).$$

Thus,

$$\mathit{lfp}(\mathcal{D}^u_T(P, \cdot)) = \mathit{lfp}(\mathcal{D}^l_T(\cdot, P)) = D^{st}_T(P).$$

Having defined the operator D^{st}_T on possible world structures, we now define an operator $\mathcal{D}^{st}_T$ on belief pairs as follows:

$$\mathcal{D}^{st}_T(P, S) = (D^{st}_T(S), D^{st}_T(P)).$$

From our earlier discussion it follows that the operator $\mathcal{D}^{st}_T$ provides yet another way of revising belief pairs: A belief pair (P, S) is replaced by the belief pair $\mathcal{D}^{st}_T(P, S) = (P', S')$, where P' is a conservative estimate of what is believed given an old liberal estimate S, and S' is a liberal estimate on what is believed given an old conservative estimate P.

Clearly, D^{st}_T is an operator on the lattice $(\mathcal{W}, \sqsubseteq)$. We refer to possible-world structures that are fixpoints of the operator D^{st}_T (and also to their theories) as *extensions*. The choice of the term is not arbitrary. We show in Section 6 that extensions of modal theories can be regarded as generalizations of extensions of default theories. We call fixpoints of the operator $\mathcal{D}^{st}_T$, defined on the lattice $(\mathcal{B}, \preceq_{pr})$ of belief pairs, *partial extensions*, as they can be viewed as "belief-pair" versions of extensions. Indeed, we have the following property relating fixpoints of the operators D^{st}_T and $\mathcal{D}^{st}_T$.

Theorem 4.1 *For every modal theory T, a possible-world structure P is a fixpoint of D^{st}_T if and only if a belief pair (P, P) is a fixpoint of $\mathcal{D}^{st}_T$.*

[3] A similar construction, in the context of logic programming, was introduced by Przymusinski [Prz90].

Proof: The statement follows immediately from the definition of the operator $\mathcal{D}^{st}_T$. □

Example 2.1 (cont'd). We will determine the operator $D^{st}_{T_p}$. Let us first observe that for every $S \in \mathcal{W}_p$ and each I:

$$\mathcal{H}^2_{(S,\mathcal{A}_p),I}(\neg Kp) = \neg\mathcal{H}^2_{(\mathcal{A}_p,S),I}(Kp) = \mathbf{t}$$

and hence,

$$\mathcal{H}^2_{(S,\mathcal{A}_p),I}(Kp \Rightarrow p) = \mathbf{t}.$$

Consequently, for each S, it holds that

$$\mathcal{D}^l_{T_p}(\mathcal{A}_p, S) = \{I\colon \mathcal{H}^2_{(S,\mathcal{A}_p),I}(Kp \Rightarrow p) = \mathbf{t}\} = \mathcal{A}_p.$$

It follows that for every $S \in \mathcal{W}_p$, $\mathcal{D}^l_{T_p}(\mathcal{A}_p, S) = \mathcal{A}_p$. Thus, $\mathcal{A}_p$ is the least fixpoint of the operator $\mathcal{D}^l_{T_p}(\cdot, S)$ (let us recall that $\mathcal{A}_p$ is the least element of the lattice $(\mathcal{W}_p, \sqsubseteq)$ on which the operator $\mathcal{D}^l_{T_p}(\cdot, S)$ is defined). That is,

$$D^{st}_{T_p}(S) = \mathcal{A}_p,$$

for every $S \in \mathcal{W}_p$.

It follows that the theory T_p has exactly one extension, $\mathcal{A}_p$ (operator $D^{st}_{T_p}$ has exactly one fixpoint). It is also easy to see that $(\mathcal{A}_p, \mathcal{A}_p)$ is the only partial extension of T_p (the only fixpoint of the operator $\mathcal{D}^{st}_{T_p}$). □

The circular dependence allowing the agent to accept p to the belief set just on the basis of this agent believing in p, allowed under the semantics of expansions, is eliminated in the case of extensions. For instance, as we observed before, the theory $\{Kp \Rightarrow p\}$ has two expansions. One of them is determined by the possible-world structure consisting of all interpretations, the other one — by the possible-world structure consisting of all interpretations in which p is true. It is this second expansion that suffers from the circular-argument problem: the belief in p is the only justification for having p in this expansion. At the same time, the theory $\{Kp \Rightarrow p\}$ has exactly one extension, the one given by the possible-world structure consisting of all interpretations. The atom p is not known in it and, hence, circular arguments are not used in the construction of this expansion.

The following result collects most important properties of the operators D^{st}_T and $\mathcal{D}^{st}_T$.

Theorem 4.2 *Let T be a modal theory. The operator D^{st}_T is $\sqsubseteq$-antimonotone. The operator $\mathcal{D}^{st}_T$ is $\preceq_{pr}$-monotone and symmetric. Moreover, for every consistent belief pair B, $\mathcal{D}^{st}_T(B)$ is also consistent.*

Proof: We will use the following additional basic property of operators on lattices. An element x of a lattice L is a *prefixpoint* of an operator $O : L \to L$ if $O(x) \leq x$. A proof of the theorem by Tarski and Knaster shows that on a complete lattice L the least prefixpoint of a monotonic operator O exists and, in fact, is equal to the least fixpoint of O. Thus for each prefixpoint x of O, $lfp(O) \leq x$.

Let us consider two possible-world structures $P, S \in \mathcal{W}$ such that $P \sqsubseteq S$. We set $P' = D^{st}_T(P)$ and $S' = D^{st}_T(S)$. We have that $P' = D^{st}_T(P) = lfp(\mathcal{D}^l_T(\cdot, P))$. Therefore, $P' =$

$\mathcal{D}^l_T(P', P)$. Since $(P', S) \preceq_{pr} (P', P)$, by $\preceq_{pr}$-monotonicity of $\mathcal{D}_T$ we obtain that $\mathcal{D}^l_T(P', S) \sqsubseteq \mathcal{D}^l_T(P', P) = P'$. Consequently, P' is a prefixpoint of $\mathcal{D}^l_T(\cdot, S)$. By our earlier remarks, $S' = D^{st}_T(S) = lfp(\mathcal{D}^l_T(\cdot, S)) \sqsubseteq P'$.

The $\preceq_{pr}$-monotonicity of the operator $\mathcal{D}^{st}_T$ is an immediate consequence of the $\sqsubseteq$-antimonotonicity of D^{st}_T. The symmetry of the operator $\mathcal{D}^{st}_T$ follows directly from its definition.

The last part of the assertion follows from Proposition 3.5. □

Theorem 4.2 implies, in particular, that the operator $\mathcal{D}^{st}_T$ has a least fixpoint with respect to the ordering $\preceq_{pr}$. We will refer to this fixpoint as the *well-founded fixpoint* of T or *well-founded partial extension* of T. We will denote this fixpoint by $WF(T)$. Our choice of the term again is not accidental. The semantics specified by the well-founded fixpoint $WF(T)$ is closely related to the well-founded semantics for default logic [BS91] and logic programming [VRS91].

The well-founded partial extension for a modal theory can be used to approximate all partial extensions of T. It also provides a sufficient condition for the uniqueness of an extension. We have the following result, analogous to Corollary 3.10. It follows from Theorem 4.2 and Propositions 3.5 and 3.6.

Corollary 4.3 *Let T be a modal theory.*

1. *The fixpoint $WF(T)$ is consistent.*
2. *For every partial extension B of T, $WF(T) \preceq_{pr} B$.*
3. *For every partial extension B of T,*

$$kn(WF(T)) \subseteq kn(B) \quad and \quad ig(WF(T)) \subseteq ig(B).$$

4. *If $WF(T)$ is a complete belief pair, then it is the unique partial extension of T. Moreover the possible-world structure P such that $WF(T) = (P, P)$ is the unique extension of T.*

Well-founded semantics has a constructive flavor. It can be obtained by iterating the operator $\mathcal{D}^{st}_T$ over the belief pair $(\mathcal{A}, \emptyset)$. We will discuss an algorithm for computing $WF(T)$ in Section 7. We will also show there that the problem of computing the well-founded semantics is in the class Δ^2_P.

The next result connects expansions and the Kripke-Kleene semantics with extensions and the well-founded semantics. It shows that the well-founded semantics is stronger than the Kripke-Kleene semantics and that (partial) extensions of T are (partial) expansions of T. Moreover extensions satisfy an additional minimality condition with respect to the ordering $\sqsubseteq$ in $\mathcal{B}$.

This minimality condition is expressed in terms of the following order on belief pairs: we define $(P, S) \sqsubseteq (P', S')$ if $P \sqsubseteq P'$ and $S \sqsubseteq S'$. We briefly mentioned it in Section 3 and referred to it as the *knowledge* ordering.

Theorem 4.4 *Let T be a modal theory. Then:*

1. *$KK(T) \preceq_{pr} WF(T)$.*
2. *Every extension of T is a $\sqsubseteq$-minimal expansion of T.*

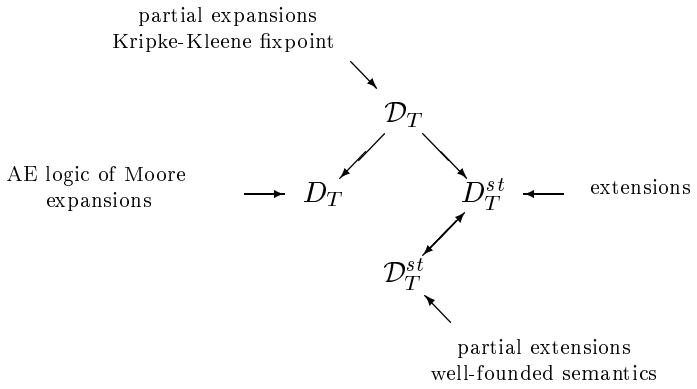

Figure 4: Operators associated with autoepistemic logic

3. *Every partial extension (P, S) of T is a $\sqsubseteq$-minimal partial expansion of T: for every partial expansion (P', S'), if $P' \sqsubseteq P$ and $S' \sqsubseteq S$, then $P = P'$ and $S = S'$.*

Proof: (1) Since $KK(T)$ is the $\preceq_{pr}$-least fixpoint of $\mathcal{D}_T$, it suffices to show that each fixpoint of $\mathcal{D}_T^{st}$ is a fixpoint of $\mathcal{D}_T$. Let (P, S) be a fixpoint of $\mathcal{D}_T^{st}$. Then $P = D_T^{st}(S) = lfp(\mathcal{D}_T^l(\cdot, S))$ and consequently, $\mathcal{D}_T^l(P, S) = P$. Similarly, $S = lfp(\mathcal{D}_T^l(\cdot, P))$ and, hence, $S = \mathcal{D}_T^l(S, P) = \mathcal{D}_T^u(P, S)$. It follows that $\mathcal{D}_T(P, S) = (P, S)$.

Since (2) is a special case of (3), it suffices to prove (3). Let (P, S) be a partial extension of T. Let us assume that (P', S') is a fixpoint of $\mathcal{D}_T$ such that $(P', S') \sqsubseteq (P, S)$. Since $P' \sqsubseteq P$, it follows that $(S', P) \preceq_{pr} (S', P')$. The $\preceq_{pr}$-monotonicity of $\mathcal{D}_T$ implies that $\mathcal{D}_T^l(S', P) \sqsubseteq \mathcal{D}_T^l(S', P') = S'$. Thus, S' is a prefixpoint of the operator $\mathcal{D}_T^l(\cdot, P)$. Since S is the least fixpoint of $\mathcal{D}_T^l(\cdot, P)$, it follows that $S \sqsubseteq S'$ (we refer the reader to the comment we made in the proof of Theorem 4.2). Now, however, due to the assumption that $(P', S') \sqsubseteq (P, S)$, it follows that $S' \sqsubseteq S$. Consequently, S and S' are identical. By a similar argument, we prove that P and P' are identical. Thus, $(P', S') = (P, S)$ which, in turn, implies that (P, S) is a $\sqsubseteq$-minimal fixpoint of $\mathcal{D}_T$. □

Example 2.1 (cont'd). In the case of the theory T_p, $KK(T_p) = (\mathcal{A}_p, \mathcal{I}_p)$ and $WF(T_p) = (\mathcal{A}_p, \mathcal{A}_p)$. Thus, we indeed have $KK(T_p) \preceq_{pr} WF(T_p)$ (see Figure 2). Let us also note that the partial extension $(\mathcal{A}_p, \mathcal{A}_p)$ is indeed a $\sqsubseteq$-minimal partial expansion of T_p. □

We can give now a schematic illustration of the panorama of semantics for autoepistemic logic (Figure 4). The central position is occupied by the operator $\mathcal{D}_T$. Its fixpoints yield the semantics of partial expansions and its least fixpoint yields the Kripke-Kleene semantics. Restriction of the operator $\mathcal{D}_T$ to complete belief pairs leads to the operator D_T, originally introduced by Moore, and results in the semantics of expansions. The operator $\mathcal{D}_T$ also gives rise to the operators D_T^{st} and $\mathcal{D}_T^{st}$ that yield new semantics for autoepistemic logic: the semantics of extensions, the semantics of partial extensions and the well-founded semantics.

5 Default logic

While possible-world semantics played a prominent role in the study of autoepistemic logics [Moo84, Lev90, DMT99] it has not, up to now, had a similar impact on default logic. In this section we will introduce a comprehensive semantic treatment of default logic in terms of possible-world structures and belief pairs. Our approach will follow closely that used in the preceding sections.

We observed earlier that autoepistemic logic can be viewed as the logic of the operator $\mathcal{D}_T$. Its fixpoints, and fixpoints of the operators that can be derived from $\mathcal{D}_T$, determine all major semantics for autoepistemic logic. We will now develop a similar treatment of default logic.

We start by recalling basic concepts in default logic. For more details we refer the reader to [MT93]. A *default* is an expression of the form

$$\frac{\alpha: \beta_1, \ldots, \beta_k}{\gamma},$$

where α, $\beta_1, \ldots, \beta_k$ and γ are propositional formulas from the language $\mathcal{L}$. The formula α is called the *prerequisite* of the default. The formulas $\beta_1, \ldots, \beta_k$ are called its *justifications*. Finally, the formula γ is called the *consequent* of the default.

A *default theory* is a pair (D, W), where D is a set of defaults and W is a set of formulas from $\mathcal{L}$. To define a semantics for a default theory $\Delta = (D, W)$, Reiter introduced an operator Γ_Δ on sets of propositional formulas [Rei80]. Given a set of formulas X, we say that a default d is *X-applicable* if for every justification β of d, $X \not\vdash \neg\beta$ (intuitively, a default is X-applicable if none of its justifications is outright contradicted by X). For a set X of propositional formulas, Reiter defined $\Gamma_\Delta(X)$ to be a least set of formulas Y such that:

1. Y is closed under propositional provability.
2. $W \subseteq Y$.
3. For every X-applicable default $d \in D$, if the prerequisite of d is in Y then so is the consequent of d.

It is easy to see that the least set of formulas satisfying conditions (1) - (3) exists. Thus, the operator Γ_Δ is well defined. A set of formulas E is an *extension* of a default theory Δ if $E = \Gamma_\Delta(E)$ [Rei80].

While the notion of an extension received most attention, over the years several other classes of theories were proposed as alternative semantics of default theories. One of them, the semantics of *weak extensions* [MT89b], is especially relevant to our considerations. Let us define $\Gamma^w_\Delta(X)$ to be a least set of formulas Y such that:

1. Y is closed under propositional provability.
2. $W \subseteq Y$.
3. For every X-applicable default $d \in D$, if the prerequisite of d is in X then the consequent of d is in Y.

As before, it is easy to see that the operator Γ^w_Δ is well defined. A set of formulas E is a *weak extension* of a default theory Δ if $E = \Gamma^w_\Delta(E)$. The concepts of extension and weak extension are closely related (not surprisingly, given that their definitions are so similar, differing only in the third condition). We refer the reader to [MT93] for a detailed discussion of default logic and properties of extensions and weak extensions.

We will introduce now an approach to default logic based on the semantic notions of a possible-world structure and of a belief pair. As before, we start with a two-valued truth function that gives a conservative estimate of the logical value of a formula or a default with respect to a belief pair (P, S) and an interpretation I.

For a propositional formula φ, we define

$$\mathcal{H}^{dl}_{(P,S),I}(\varphi) = I(\varphi).$$

For a default $d = \frac{\alpha:\beta_1,\ldots,\beta_k}{\gamma}$, we set

$$\mathcal{H}^{dl}_{(P,S),I}(d) = \mathbf{t}$$

if at least one of the following conditions holds:

1. There is $J \in S$ such that $J(\alpha) = \mathbf{f}$.
2. There is i, $1 \leq i \leq k$, such that for every $J \in P$, $J(\beta_i) = \mathbf{f}$.
3. $I(\gamma) = \mathbf{t}$.

We set $\mathcal{H}^{dl}_{(P,S),I}(d) = \mathbf{f}$, otherwise. Clearly, the definition of $\mathcal{H}^{dl}_{(P,S),I}(d)$ agrees with the intuitive reading of a default d: it is true (according to a conservative point of view) if its prerequisite is false (even with respect to a liberal view captured by S) or if at least one of its justifications is perceived as impossible (it is false according to a conservative point of view captured by P) or if its consequent is true (with respect to I). As before, we can also argue that $\mathcal{H}^{dl}_{(S,P),I}(d)$ provides a liberal estimate for a truth value of d with respect to (P, S) (the roles of P and S are reversed).

This truth function $\mathcal{H}^{dl}_{(P,S),I}$ satisfies a monotonicity property analogous to that satisfied by the truth function $\mathcal{H}^{2}_{(P,S),I}$ in the case of autoepistemic logic (see Proposition 3.2).

Proposition 5.1 *Let (P, S) and (P', S') be belief pairs from $\mathcal{B}$ such that $(P, S) \preceq_{pr} (P', S')$. Then, for each default d, for each propositional formula φ and for each interpretation $I \in \mathcal{A}$, $\mathcal{H}^{dl}_{(P,S),I}(d) \leq \mathcal{H}^{dl}_{(P',S'),I}(d)$ and $\mathcal{H}^{dl}_{(P,S),I}(\varphi) = \mathcal{H}^{dl}_{(P',S'),I}(\varphi)$.*

Proof: To prove the first part of the assertion let us consider a default $d = \frac{\alpha:\beta_1,\ldots,\beta_k}{\gamma}$ and let us assume that $\mathcal{H}^{dl}_{(P,S),I}(d) = \mathbf{t}$ (the case when $\mathcal{H}^{dl}_{(P,S),I}(d) = \mathbf{f}$ is trivial). By the definition of $\mathcal{H}^{dl}_{(P,S),I}(d)$, there are three cases to consider.

1. There is $J \in S$ such that $J(\alpha) = \mathbf{f}$. Since $S \subseteq S'$, it follows that $\mathcal{H}^{dl}_{(P',S'),I}(d) = \mathbf{t}$.
2. There is i, $1 \leq i \leq k$, such that for every $J \in P$, $J(\beta_i) = \mathbf{f}$. Since $P' \subseteq P$, it again follows that $\mathcal{H}^{dl}_{(P',S'),I}(d) = \mathbf{t}$.
3. We have $I(\gamma) = \mathbf{t}$. In this case, clearly, $\mathcal{H}^{dl}_{(P',S'),I}(d) = \mathbf{t}$, as well.

The second part of the assertion is straightforward as $\mathcal{H}^{dl}_{(P,S),I}(\varphi) = I(\varphi) = \mathcal{H}^{dl}_{(P',S'),I}(\varphi)$. □

Let $\Delta = (D, W)$ be a default theory. We use the truth function $\mathcal{H}^{dl}_{(P,S),I}$ to define an operator $\mathcal{E}_\Delta$ on the lattice $\mathcal{B}$ of belief pairs:

$$\mathcal{E}_\Delta(P, S) = (\mathcal{E}^l_\Delta(P, S), \mathcal{E}^u_\Delta(P, S)),$$

where

$$\mathcal{E}^l_\Delta(P, S) = \{I\colon \mathcal{H}^{dl}_{(S,P),I}(\Delta) = \mathbf{t}\} \text{ and } \mathcal{E}^u_\Delta(P, S) = \{I\colon \mathcal{H}^{dl}_{(P,S),I}(\Delta) = \mathbf{t}\}$$

($\mathcal{H}^{dl}_{B,I}(\Delta) = \mathbf{t}$ stands for the statement that $\mathcal{H}^{dl}_{B,I}(d) = \mathbf{t}$ for every element (formula or default) $d \in D \cup W$). This definition can be justified similarly as that of the operator $\mathcal{D}_T$ in Section 4. We now have the following key property of the operator $\mathcal{E}_\Delta$.

Proposition 5.2 *The operator $\mathcal{E}_\Delta$ is $\preceq_{pr}$-monotone and symmetric.*

Proof: By the definition, we have $\mathcal{E}^l_\Delta(P, S) = \mathcal{E}^u_\Delta(S, P)$. Thus, $\mathcal{E}_\Delta$ is a symmetric operator. To prove the $\preceq_{pr}$-monotonicity of the operator $\mathcal{E}_\Delta$, let us consider two belief pairs (P, S) and (P', S') such that $(P, S) \preceq_{pr} (P', S')$. We need to prove that

$$\mathcal{E}^l_\Delta(P', S') \subseteq \mathcal{E}^l_\Delta(P, S) \text{ and } \mathcal{E}^u_\Delta(P, S) \subseteq \mathcal{E}^u_\Delta(P', S').$$

Let $I \in \mathcal{E}^l_\Delta(P', S')$. Then, $\mathcal{H}^{dl}_{(S',P'),I}(\Delta) = \mathbf{t}$. Since $(P, S) \preceq_{pr} (P', S')$, it follows that $(S', P') \preceq_{pr} (S, P)$. Thus, by Proposition 5.1, $\mathcal{H}^{dl}_{(S,P),I}(\Delta) = \mathbf{t}$ and, consequently, $I \in \mathcal{E}^l_\Delta(P, S)$. The second inclusion can be proved in the same manner. □

Let Q be a possible-world structure. We define

$$E_\Delta(Q) = \mathcal{E}^l_\Delta(Q, Q)$$

(or, equivalently, $E_\Delta(Q) = \mathcal{E}^u_\Delta(Q, Q)$). Clearly,

$$\mathcal{E}_\Delta(Q, Q) = (E_\Delta(Q), E_\Delta(Q)).$$

As we will show later, fixpoints of the operators E_Δ and $\mathcal{E}_\Delta$ correspond to fixpoints of the operators D_T and $\mathcal{D}_T$. Thus, we will call them *expansions* and *partial expansions* of Δ, respectively. It is clear that

$$Q = E_\Delta(Q) \text{ if and only if } (Q, Q) = \mathcal{E}_\Delta(Q, Q).$$

Thus, those partial expansions of Δ that are complete correspond precisely to expansions of Δ.

To the best of our knowledge, the operator E_Δ has not appeared explicitly in the literature before. Its fixpoints, however, did. It turns out that they correspond to weak extensions [MT89a]. Hence, the semantics given by the operator E_Δ (by its fixpoints, to be precise) is the semantics of weak extensions.

Theorem 5.3 *Let Δ be a default theory. If a possible-world structure Q is an expansion of Δ then $\{\varphi \in \mathcal{L}\colon I(\varphi) = \mathbf{t}$, for every $I \in Q\}$ is a weak extension of Δ. Conversely, if E is a weak extension of Δ then $Q = \{I \in \mathcal{A}\colon I(\varphi) = \mathbf{t}$, for every $\varphi \in E\}$ is an expansion of Δ.*

Proof: Let us assume that Δ is of the form (D, W). Let Q be an expansion of Δ. Directly from the definition of an expansion it follows that

$$Q = \{I \in \mathcal{A}\colon \mathcal{H}^{dl}_{(Q,Q),I}(\Delta) = \mathbf{t}\}. \tag{1}$$

Let us set

$$Y_Q = \{\varphi \in \mathcal{L}\colon I(\varphi) = \mathbf{t}, \text{ for every } I \in Q\}.$$

We will show that Y_Q is a weak extension of (D, W). To this end, we will show that $Y_Q = \Gamma^w_\Delta(Y_Q)$.

First, it follows from the definition of Y_Q that it is closed under propositional provability. Further, from (1) it follows that for every $I \in Q$ and for every $\varphi \in W$, $I(\varphi) = \mathbf{t}$. Hence, $W \subseteq Y_Q$.

Next, let us consider a Y_Q-applicable default

$$d = \frac{\alpha\colon \beta_1, \ldots, \beta_k}{\gamma}$$

and let us assume that $\alpha \in Y_Q$. Let $I \in Q$. The equation (1) implies that $\mathcal{H}^{dl}_{(Q,Q),I}(d) = \mathbf{t}$. Since $\alpha \in Y_Q$, there is no $J \in Y_Q$ such that $J(\alpha) = \mathbf{f}$. By Y_Q-applicability of d, for every i, $1 \leq i \leq k$, $Y_Q \not\vdash \neg\beta_i$. Since Y_Q is closed under propositional provability, we have that $\neg\beta_i \notin Y_Q$. Thus, for every i, $1 \leq i \leq k$, there is a valuation $J \in Q$ such that $J(\beta_i) = \mathbf{t}$. Consequently, it follows that $I(\gamma) = \mathbf{t}$. Since I is an arbitrary valuation from Q, we find that $\gamma \in Y_Q$.

To summarize, we proved that Y_Q satisfies all three conditions on a set Y in the definition of $\Gamma^w_\Delta(X)$, with $X = Y_Q$. Since $\Gamma^w_\Delta(Y_Q)$ is the least of all sets satisfying these conditions, it follows that $\Gamma^w_\Delta(Y_Q) \subseteq Y_Q$.

To prove the converse inclusion, let us consider a valuation $I \in \mathcal{A}$ such that for every $\varphi \in \Gamma^w_\Delta(Y_Q)$, $I(\varphi) = \mathbf{t}$. We will show that $I \in Q$. By (1), it will suffice to show that $\mathcal{H}^{dl}_{(Q,Q),I}(\Delta) = \mathbf{t}$. Since $W \subseteq \Gamma^w_\Delta(Y_Q)$, it follows that for every $\varphi \in W$, $\mathcal{H}^{dl}_{(Q,Q),I}(\varphi) = I(\varphi) = \mathbf{t}$. Thus, let us consider a default

$$d = \frac{\alpha\colon \beta_1, \ldots, \beta_k}{\gamma}$$

from D. Let us assume that

(i) for every $J \in Q$, $J(\alpha) = \mathbf{t}$, and

(ii) for every i, $1 \leq i \leq k$, there is $J_i \in Q$ such that $J_i(\beta_i) = \mathbf{t}$

(if any of these two assumptions does not hold, we obtain right away $\mathcal{H}^{dl}_{(Q,Q),I}(d) = \mathbf{t}$). From (i), it follows that $\alpha \in Y_Q$. Furthermore, (ii) and the definition of Y_Q imply that for every i, $1 \leq i \leq k$, $Y_Q \not\vdash \beta_i$. Consequently, d is Y_Q-applicable and, by the definition of $\Gamma^w_\Delta(Y_Q)$, we get that $\gamma \in \Gamma^w_\Delta(Y_Q)$. Hence, $I(\gamma) = \mathbf{t}$ and $\mathcal{H}^{dl}_{(Q,Q),I}(d) = \mathbf{t}$.

We proved that $I \in Q$. That is, we have that

$$\{I \in \mathcal{A}\colon I(\varphi) = \mathbf{t}, \text{ for every } \varphi \in \Gamma^w_\Delta(Y_Q)\} \subseteq Q.$$

By the definition of Y_Q, we have that for every $I \in Q$ and for every $\varphi \in Y_Q$, $I(\varphi) = \mathbf{t}$. Consequently,

$$Q \subseteq \{I \in \mathcal{A}\colon I(\varphi) = \mathbf{t}, \text{ for every } \varphi \in Y_Q.\}$$

Thus,

$$\{I \in \mathcal{A}\colon I(\varphi) = \mathbf{t},\ \text{for every } \varphi \in \Gamma^w_\Delta(Y_Q)\} \subseteq \{I \in \mathcal{A}\colon I(\varphi) = \mathbf{t}\ \ \text{for every } \varphi \in Y_Q\}$$

or, equivalently, $Y_Q \subseteq \Gamma^w_\Delta(Y_Q)$ (this last step depends on the fact that Y_Q and $\Gamma^w_\Delta(Y_Q)$ are closed under propositional provability).

The second part of the assertion can be proved by means of a similar argument and so we omit it. □

The key property of the operator $\mathcal{E}_\Delta$ is its $\preceq_{pr}$-monotonicity. In particular, $\mathcal{E}_\Delta$ has a least fixpoint. We call it the *Kripke-Kleene fixpoint* or *Kripke-Kleene expansion* of Δ. We denote it by $KK(\Delta)$. We refer to the corresponding semantics as the *Kripke-Kleene semantics* for Δ.

As in the case of the autoepistemic logic, the Kripke-Kleene semantics approximates the skeptical reasoning with partial expansions and provides a test for uniqueness of a partial expansion. Indeed, we have the following corollary[4] to Propositions 5.2, 3.5 and 3.6. This result is a counterpart to Corollary 3.10.

Corollary 5.4 *Let Δ be a default theory.*

1. *The fixpoint $KK(\Delta)$ is consistent.*
2. *For every partial expansion B of Δ, $KK(\Delta) \preceq_{pr} B$.*
3. *For every partial expansion B of Δ,*

$$kn(KK(\Delta)) \subseteq kn(B) \ \ \textit{and} \ \ ig(KK(\Delta)) \subseteq ig(B).$$

4. *If $KK(\Delta)$ is a complete belief pair, then it is a unique consistent partial expansion of Δ. Moreover the possible-world structure P such that $KK(\Delta) = (P, P)$ is the unique expansion of Δ.*

So far we have not yet reconstructed the concept of an extension. In order to do so, we will now derive from $\mathcal{E}_\Delta$ two other operators related to default logic. Let us consider a belief pair (P, S). We want to revise it to a belief pair (P', S'). We might do it by fixing S and taking for P' a preferred revision of P, and by fixing P and taking for S' a preferred revision of S.

It is easy to see that $\preceq_{pr}$-monotonicity of $\mathcal{E}_\Delta$ implies that the operator $\mathcal{E}^l_\Delta(\cdot, S)$ is $\sqsubseteq$-monotone operator on $\mathcal{W}$. Consequently, it has a least fixpoint. This fixpoint can be taken as the preferred way to revise P given S. Thus, we define

$$E^{st}_\Delta(S) = lfp(\mathcal{E}^l_\Delta(\cdot, S)).$$

As in the case of autoepistemic logic, one can see that E^{st}_Δ also specifies the preferred way to revise S given P. Combining these two revisions, we define the operator on $\mathcal{B}$ as follows:

$$\mathcal{E}^{st}_\Delta(P, S) = (E^{st}_\Delta(S), E^{st}_\Delta(P)).$$

The operator $\mathcal{E}^{st}_\Delta$ describes a way to revise belief pairs.

We start our discussion of the properties of the operators E^{st}_Δ and $\mathcal{E}^{st}_\Delta$ with the following straightforward result relating their fixpoints.

[4] As a matter of fact, the proof of part (3) requires additional arguments. However, it is easy to derive this claim from Corollary 3.10 and Theorem 6.3, proved in the next section. Thus, we omit the direct argument here.

Proposition 5.5 *Let Δ be a default theory. For every possible-world structure P, P is a fixpoint of E^{st}_{Δ} if and only if (P, P) is a fixpoint of $\mathcal{E}^{st}_{\Delta}$.*

The operator E^{st}_{Δ} allows us to reconstruct the notion of an extension as defined by Reiter. Namely, we have the following theorem.

Theorem 5.6 *Let Δ be a default theory. If a possible-world structure Q is a fixpoint of E^{st}_{Δ} then the theory $E = \{\varphi \in \mathcal{L}\colon I(\varphi) = \mathbf{t},$ for every $I \in Q\}$ is an extension of Δ. Conversely, if a theory E is an extension of Δ then $Q = \{I \in \mathcal{A}\colon I(\varphi) = \mathbf{t},$ for every $\varphi \in E\}$ is a fixpoint of E^{st}_{Δ}.*

Proof: The proof is very similar to that of Theorem 5.3. Let us recall that we proved there the first assertion (a counterpart of the first assertion of the present theorem) and omitted the proof of the second statement. Here we proceed the other way around. We omit the proof of the first assertion and provide an argument for the second assertion only.

In the proof, we will use the concept of a *generating default* [Rei80, MT93]. Let E be a theory closed under propositional consequence. A default $d = \frac{\alpha:\beta_1,\ldots,\beta_k}{\gamma}$ is *generating for* E if $\alpha \in E$ and for every i, $1 \leq i \leq k$, $\neg\beta_i \notin E$. Extensions can be characterized by means of generating defaults. Namely, we have the following result [Rei80]: if E is an extension of a default theory (D, W) then

$$E = Cn(W \cup CGD_E), \tag{2}$$

where CGD_E is the set of the consequents of all those defaults in D that are generating for E.

Let us consider an extension E of a default theory Δ (we will assume that $\Delta = (D, W)$). We define $Q = \{I \in \mathcal{A}\colon I(\varphi) = \mathbf{t},$ for every $\varphi \in E\}$ (in other words, Q is the set of all models of E). Since E is closed under propositional provability, $E = \{\varphi \in \mathcal{L}\colon I(\varphi) = \mathbf{t},$ for every $I \in Q\}$. Therefore, to prove the second assertion it suffices to show that $Q = E^{st}_{\Delta}(Q)$. To this end we will prove that Q is the least fixpoint of the operator $\mathcal{E}^{l}_{\Delta}(\cdot, Q)$. We will do so by showing that Q is a pre-fixpoint of $\mathcal{E}^{l}_{\Delta}(\cdot, Q)$ (that is, satisfies $\mathcal{E}^{l}_{\Delta}(Q, Q) \sqsubseteq Q$) and that for any fixpoint Q' of the operator $\mathcal{E}^{l}_{\Delta}(\cdot, Q)$, $Q \sqsubseteq Q'$.

We will first prove that $\mathcal{E}^{l}_{\Delta}(Q, Q) \sqsubseteq Q$. Let us recall that $\mathcal{E}^{l}_{\Delta}(Q, Q) = \{I \in \mathcal{A}\colon \mathcal{H}^{dl}_{(Q,Q),I}(\Delta) = \mathbf{t}\}$. Let us consider a valuation $I \in Q$. Since E is an extension of (D, W), $W \subseteq E$. Thus, for every $\varphi \in W$, $\mathcal{H}^{dl}_{(Q,Q),I}(\varphi) = I(\varphi) = \mathbf{t}$. Next, let us consider a default $d = \frac{\alpha:\beta_1,\ldots,\beta_k}{\gamma}$ from D. If d is a generating default for E, then $\gamma \in E$ and $I(\gamma) = \mathbf{t}$. Thus, $\mathcal{H}^{dl}_{(Q,Q),I}(d) = \mathbf{t}$. If d is not a generating default or E, then either we have (1) $\alpha \notin E$, or (2) there is i, $1 \leq i \leq k$, such that $J(\beta_i) = \mathbf{f}$, for every $J \in Q$. In either case, it follows that $\mathcal{H}^{dl}_{(Q,Q),I}(d) = \mathbf{t}$, as well. Consequently, $\mathcal{H}^{dl}_{(Q,Q),I}(\Delta) = \mathbf{t}$ and $I \in \mathcal{E}^{l}_{\Delta}(Q, Q)$. Thus, we get $Q \subseteq \mathcal{E}^{l}_{\Delta}(Q, Q)$, or equivalently $\mathcal{E}^{l}_{\Delta}(Q, Q) \sqsubseteq Q$.

Let us now consider a fixpoint Q' of $\mathcal{E}^{l}_{\Delta}(\cdot, Q)$ and let us define $E' = \{\varphi \in \mathcal{L}\colon I(\varphi) = \mathbf{t},$ for every $I \in Q'\}$. Clearly, E' is closed under propositional provability. Since Q' is a fixpoint of $\mathcal{E}^{l}_{\Delta}(\cdot, Q)$,

$$Q' = \mathcal{E}^{l}_{\Delta}(Q', Q) = \{I \in \mathcal{A}\colon \mathcal{H}^{dl}_{(Q,Q'),I}(\Delta) = \mathbf{t}\}.$$

Thus, for every $\varphi \in W$ and for every $I \in Q'$, $I(\varphi) = \mathbf{t}$. In other words, $W \subseteq E'$. Next, let us consider a default $d = \frac{\alpha:\beta_1,\ldots,\beta_k}{\gamma}$ from D. Let us assume that $\alpha \in E'$ and that for every i,

$1 \leq i \leq k$, $E \not\vdash \neg\beta_i$. It follows that for every $J \in Q'$, $J(\alpha) = \mathbf{t}$ and, since E is closed under propositional provability, that for every i, $1 \leq i \leq k$, there is $J_i \in Q$ such that $J_i(\beta_i) = \mathbf{t}$. Let $I \in Q'$. Since $\mathcal{H}^{dl}_{(Q,Q'),I}(d) = \mathbf{t}$, it follows that $I(\gamma) = \mathbf{t}$. Thus, $\gamma \in E'$.

We have just proved that E' satisfies the three requirements from the definition of $\Gamma_\Delta(E)$. Thus, $E = \Gamma_\Delta(E) \subseteq E'$. Consequently, $Q' \subseteq Q$ or, equivalently, $Q \sqsubseteq Q'$. We proved that Q is a pre-fixpoint of $\mathcal{E}^l_\Delta(\cdot, Q)$ and that $Q \sqsubseteq Q'$ for any fixpoint Q' of $\mathcal{E}^l_\Delta(\cdot, Q)$. It follows that Q is the least fixpoint of $\mathcal{E}^l_\Delta(\cdot, Q)$ □

In view of Theorem 5.6, we refer to the fixpoints of the operator E^{st}_Δ as *extensions* of Δ. Further, in view of Proposition 5.5, we call fixpoints of the operator $\mathcal{E}^{st}_\Delta$, *partial extensions* of Δ. One can show that consistent partial extensions of a default theory Δ are in one-to-one correspondence with stationary extensions of Δ defined in [PP94].

We also note that the operator E^{st}_Δ coincides with the operator Σ^Δ defined on sets of interpretations and proposed by Guerreiro and Casanova [GC90, Lif90, MT93]. Guerreiro and Casanova simply rephrased the original definition of the operator Γ_Δ (which works on theories and can be restricted, without the loss of generality, to theories closed under propositional provability) in terms of sets of interpretations (possible-world structures) that are models of such theories. One of our contributions is that we derive this operator in a systematic and purely algebraic (thus, not relying on any particular properties of defaults) fashion from an operator $\mathcal{E}_\Delta$ defined on the lattice of belief pairs.

Our next result describes monotonicity properties of the operators E^{st}_Δ and $\mathcal{E}^{st}_\Delta$. It is analogous to Theorem 4.2 and can be proved in the same way.

Theorem 5.7 *Let Δ be a default theory. Then, the operator E^{st}_Δ is $\sqsubseteq$-antimonotone and the operator $\mathcal{E}^{st}_\Delta$ is $\preceq_{pr}$-monotone and symmetric.*

Theorem 5.7 implies that the operator $\mathcal{E}^{st}_\Delta$ has a least fixpoint. We will denote it by $WF(\Delta)$ and refer to it as the *well-founded fixpoint* (or the *well-founded extension*) of Δ. We will call the semantics it implies the *well-founded semantics* of Δ. One can show that the well-founded semantics of Δ, as we introduced it here, coincides with the well-founded semantics of default logic introduced in [BS91].

The well-founded semantics allows us to approximate skeptical reasoning with extensions and yields a sufficient condition for the uniqueness of an extension. As before, the following result is a simple corollary to the fact that $\mathcal{E}^{st}_\Delta$ is symmetric and $\preceq_{pr}$-monotone (Theorem 5.7), and to Propositions 3.5 and 3.6. We also note that, as in the case of Corollary 5.4, part (3) of the assertion requires additional arguments and can be derived, for instance, from Corollary 4.3 and Theorem 6.3.

Corollary 5.8 *Let Δ be a default theory.*

1. *The fixpoint $WF(\Delta)$ is consistent.*
2. *For every partial extension B of Δ, $WF(\Delta) \preceq_{pr} B$.*
3. *For every partial extension B of Δ,*

$$kn(WF(T)) \subseteq kn(B) \quad \textit{and} \quad ig(WF(T)) \subseteq ig(B).$$

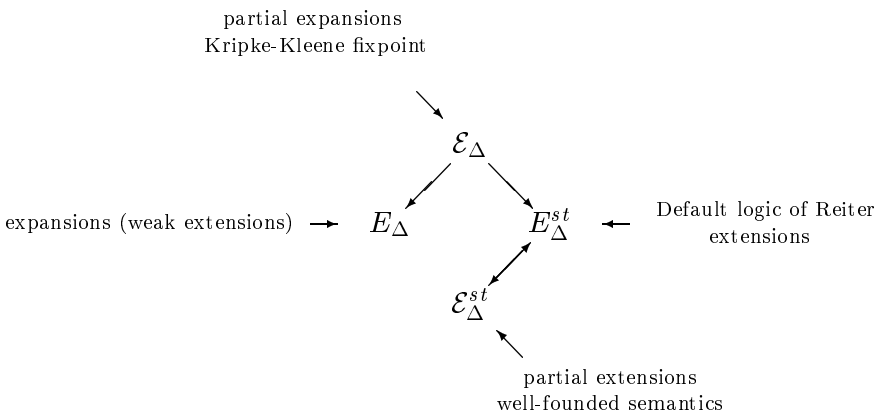

Figure 5: Operators associated with default logic

4. *If $WF(T)$ is a complete belief pair, then it is the unique consistent partial extension of Δ. Moreover the possible-world structure P such that $WF(T) = (P, P)$ is the unique extension of Δ.*

Finally, let us note connections between (partial) expansions and (partial) extensions, and between the Kripke-Kleene and well-founded semantics for default logic. The proof of this result follows closely the lines of the proof of Theorem 4.4.

Theorem 5.9 *Let Δ be a default theory. Then:*

1. *$KK(\Delta) \preceq_{pr} WF(\Delta)$.*
2. *Every extension of Δ is a $\sqsubseteq$-minimal expansion of Δ.*
3. *Every partial extension (P, S) of Δ is a $\sqsubseteq$-minimal partial expansions of Δ: for every partial expansions (P', S'), if $P' \sqsubseteq P$ and $S' \sqsubseteq S$, then $P = P'$ and $S = S'$.*

In summary, default logic can be viewed as the logic of the operator $\mathcal{E}_\Delta$. That is, the fixpoints of $\mathcal{E}_\Delta$ define the semantics of partial expansions. The least fixpoint of $\mathcal{E}_\Delta$ defines the Kripke-Kleene fixpoint. The operator $\mathcal{E}_\Delta$ gives rise to the operator E_Δ, whose fixpoints are expansions (also referred to as weak extensions). Kripke-Kleene semantics provides an approximation for the skeptical reasoning under the semantics of expansions. The operator $\mathcal{E}_\Delta$ also leads to the operator $\mathcal{E}^{st}_\Delta$. Consistent fixpoints of this operator yield partial extensions (stationary extensions in the terminology of [PP94]). Fixpoints of the related operator E^{st}_Δ correspond to extensions by Reiter. The least fixpoint of the operator $\mathcal{E}^{st}_\Delta$ results in the well-founded semantics for default logic and approximates the skeptical reasoning under the semantics of extensions. The relationships between the operators of default logic are illustrated in Figure 5. They are parallel to those for the autoepistemic logic (Figure 4).

6 Default logic versus autoepistemic logic

The results of this paper shed new light on the relationship between default and autoepistemic logics. The nature of this relationship was the subject of extensive investigations since the

time both systems were introduced in the early 80s. Konolige [Kon88] proposed to encode a default $d = \frac{\alpha:\beta_1,\ldots,\beta_k}{\gamma}$ by the modal formula

$$m(d) = K\alpha \wedge \neg K\neg\beta_1 \wedge \ldots \neg K\neg\beta_k \Rightarrow \gamma,$$

and to represent a default theory $\Delta = (D, W)$ by a modal theory

$$m(\Delta) = W \cup \{m(d)\colon d \in D\}.$$

Despite the fact that the encoding is intuitive it does not provide a correspondence between default logic as defined by Reiter and autoepistemic logic as defined by Moore. Let us consider a default theory Δ with $W = \emptyset$ and $D = \{\frac{p:q}{p}\}$, where p and q are two different atoms. Then Δ has exactly one extension, $Cn(\emptyset)$. Applying the translation of Konolige to Δ yields the theory $m(\Delta) = \{Kp \wedge \neg K\neg q \Rightarrow p\}$. The theory $m(\Delta)$ has two expansions. One of them is generated by the theory $Cn(\emptyset)$ (equivalently, it is the possible-world structure $\mathcal{A}$) and corresponds to the only extension of Δ. The other expansion is generated by the theory $Cn(\{p\})$ (equivalently, it is the possible-world structure that consists of just one valuation of $\{p, q\}$, the one in which p is true and q is false). Thus, the Konolige's translation does not give a one-to-one correspondence between extensions of default theories and expansions of their modal encodings. Another example can be obtained from the default theory $\Delta = (D, W)$ where $W = \emptyset$, and $D = \{\frac{p:}{p}\}$ (yielding the modal theory $\{Kp \Rightarrow p\}$ that we used as a running example). Our concept of extension of an autoepistemic theory eliminated the unwanted expansion of Δ, leaving only the desired modal theory that corresponds to Reiter's extension.

This mismatch can now be explained within the semantic framework introduced in this paper. Konolige's translation does not establish a correspondence between extensions and expansions because they are associated with different operators. Expansions are associated with fixpoints of the operator D_T. Its counterpart on the side of default logic is the operator E_Δ. Fixpoints of this operator are not extensions but expansions (weak extensions, in the terminology of [MT89a]) of Δ. Extensions of Δ are associated with the operator E^{st}_Δ. Its counterpart on the side of autoepistemic logic is the operator D^{st}_T, introduced in Section 4. This operator, to the best of our knowledge, has not appeared in the literature.

In this section, we show that once we properly align concepts from default logic with those from autoepistemic logic, Konolige's translation works! This alignment is illustrated in Figure 6 and is formally described by Theorem 6.3. To prove it, we will need the following lemma.

Lemma 6.1 *Let (P, S) be a belief pair. For every interpretation $I \in \mathcal{A}$ and every default d we have*

$$\mathcal{H}^{dl}_{(P,S),I}(d) = \mathcal{H}^{2}_{(P,S),I}(m(d)).$$

Proof: Let us assume that

$$d = \frac{\alpha : \beta_1, \ldots, \beta_k}{\gamma}.$$

The modal translation $m(d)$ of d can be equivalently written as

$$\neg K\alpha \vee K\neg\beta_1 \vee \ldots \vee K\neg\beta_k \vee \gamma.$$

We now show the desired equality.

$\mathcal{H}^{dl}_{(P,S),I)}(d) = \mathbf{f}$ if and only if the following three conditions hold:

1. For every $J \in S$, $J(\alpha) \neq \mathbf{f}$.
2. For every i, $1 \leq i \leq k$, there exists $J \in P$, $J(\beta_i) \neq \mathbf{f}$.
3. $I(\gamma) \neq \mathbf{t}$.

These three conditions are equivalent to the conjunction of the following three conditions:

1. $\mathcal{H}^2_{(P,S),I}(\neg K\alpha) = \mathbf{f}$.
2. For every i, $1 \leq i \leq k$, $\mathcal{H}^2_{(P,S),I}(K\neg\beta_i) = \mathbf{f}$.
3. $\mathcal{H}^2_{(P,S),I}(\gamma) = \mathbf{f}$.

This latter set of conditions is equivalent to $\mathcal{H}^2_{(P,S),I}(m(d)) = \mathbf{f}$. Thus, $\mathcal{H}^{dl}_{(P,S),I)}(d) = \mathbf{f}$ if and only if $\mathcal{H}^2_{(P,S),I}(m(d)) = \mathbf{f}$, and the argument is complete. □

Corollary 6.2 *Let $\Delta = (D, W)$ be a default theory. Then, for every belief pair (P, S),*

$$\{I : \mathcal{H}^{dl}_{(P,S),I}(\Delta) = \mathbf{t}\} = \{I : \mathcal{H}^2_{(P,S),I}(m(\Delta)) = \mathbf{t}\}.$$

We are now ready to prove the main result of this section.

Theorem 6.3 *Let Δ be a default theory and let $T = m(\Delta)$. Then the following pairs of operators coincide and, thus, have the same fixpoints:*

1. *$\mathcal{E}_\Delta = \mathcal{D}_T$ (partial expansions for Δ and T, including Kripke-Kleene fixpoints, coincide).*
2. *$E_\Delta = D_T$ (expansions for Δ and T coincide).*
3. *$E^{st}_\Delta = D^{st}_T$ (extensions for Δ and T coincide).*
4. *$\mathcal{E}^{st}_\Delta = \mathcal{D}^{st}_T$ (partial extensions for Δ and T, including the well-founded fixpoints, coincide).*

Proof: (1) By Corollary 6.2, $\mathcal{E}_\Delta = \mathcal{D}_{m(\Delta)}$.
(2) We recall that $E_\Delta(P) = Q$ if and only if $\mathcal{E}_\Delta(P, P) = (Q, Q)$. By (1), this is equivalent to $\mathcal{D}_{m(\Delta)}(P, P) = (Q, Q)$, that is, $D_{m(\Delta)}(P) = Q$. Thus, $E_\Delta = D_{m(\Delta)}$.
(3) Since $\mathcal{E}_\Delta = \mathcal{D}_{m(\Delta)}$, we have, for every possible-world structure S,

$$\mathcal{E}^l_\Delta(\cdot, S) = \mathcal{D}^l_{m(\Delta)}(\cdot, S).$$

Therefore, for each S,

$$lfp(\mathcal{E}^l_\Delta(\cdot, S)) = lfp(\mathcal{D}^l_{m(\Delta)}(\cdot, S)).$$

But this just means that $E^{st}_\Delta = D^{st}_{m(\Delta)}$.
(4) This assertion follows immediately from (3) as $\mathcal{D}^{st}_T(P, S) = (D^{st}_T(S), D^{st}_T(P))$ and, likewise, $\mathcal{E}^{st}_\Delta(P, S) = (E^{st}_\Delta(S), E^{st}_\Delta(P))$. □

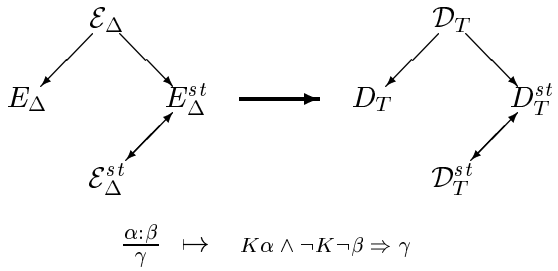

Figure 6: Embedding default logic into autoepistemic logic

7 Computing the well-founded semantics

In this section we discuss methods to compute Kripke-Kleene and well-founded fixpoints. We will focus on the case of autoepistemic logic. Since, as we demonstrated in the previous section, default logic can be viewed as a fragment of autoepistemic logic, our methods and results will apply to default logic, as well.

To compute $\mathcal{D}^l_T(P,S)$ we need to compute all interpretations I such that $\mathcal{H}^2_{(P,S),I}(\varphi) = \mathbf{t}$ for every $\varphi \in T$. The number of such interpretations may be exponential in the number of atoms in T. The key to our approach is a simple observation that these interpretations are determined by some propositional theory of much smaller size. Indeed, the logical values of modal atoms (formulas of the form $K\psi$) occurring in φ do not depend on I. In particular, the logical values of *maximal modal atoms* (modal atoms not within the scope of any other occurrence of the modal operator) are determined by the belief pair (P,S). Once these values are established, we substitute them for the corresponding modal atoms. In this way, we obtain a formula, say $\varphi_{(P,S)}$, in the propositional language with two special symbols $\mathbf{t}$ and $\mathbf{f}$ that represent truth and falsity, respectively. These special symbols are interpreted in a standard way (in fact, we use the same notation for these two special elements of the language as for the corresponding truth values). We denote this language by $\mathcal{L}^e$. The key property of the formula $\varphi_{(P,S)}$ is that $\mathcal{H}^2_{(P,S),I}(\varphi) = \mathbf{t}$ if and only if $I(\varphi_{(P,S)}) = \mathbf{t}$. Thus, the set of interpretations $\mathcal{D}^l_T(P,S)$ is represented by the set of formulas $\{\varphi_{(P,S)}\colon \varphi \in T\}$. A similar representation can be obtained for the set $\mathcal{D}^u_T(P,S)$. This leads to two questions: (1) how to compute these representations, and (2) how to use them instead of belief pairs in the process of computing the operator $\mathcal{D}$ and other related operators. The rest of this section is devoted to these issues.

For every interpretation $I \in \mathcal{A}$, we define an interpretation I^e of $\mathcal{L}^e$ by setting $I^e(\mathbf{t}) = \mathbf{t}$, $I^e(\mathbf{f}) = \mathbf{f}$, $I^e(p) = I(p)$ for every atom p in $\mathcal{L}$, and by extending it to the whole language $\mathcal{L}^e$ in the standard way. For a theory $X \subseteq \mathcal{L}^e$, we define

$$Mod(X) = \{I \in \mathcal{A}\colon I^e(\varphi) = \mathbf{t}, \text{ for every } \varphi \in X\}.$$

Let W be a possible-world structure. We call each propositional theory $X \subseteq \mathcal{L}^e$ such that $Mod(X) = W$, a *representation* of W. Similarly, given a belief pair (P,S), a pair of theories (X,Y) such that $Mod(X) = P$ and $Mod(Y) = S$ is called a *representation* of (P,S). We extend the notation Mod to pairs of theories and define $Mod(X,Y) = (Mod(X), Mod(Y))$.

Let X and Y be two theories in the language $\mathcal{L}^e$. For every modal formula $\varphi \in \mathcal{L}_K$ we define a formula $\varphi_{(X,Y)} \in \mathcal{L}^e$ by induction as follows:

1. If φ is modal-free, $\varphi_{(X,Y)} = \varphi$.
2. If $\varphi = \neg\psi$, $\varphi_{(X,Y)} = \neg\psi_{(Y,X)}$ (one should note that X and Y are reversed on the right hand side).
3. If $\varphi = \psi' \vee \psi''$, $\varphi_{(X,Y)} = \psi'_{(X,Y)} \vee \psi''_{(X,Y)}$.
4. If $\varphi = \psi' \wedge \psi''$, $\varphi_{(X,Y)} = \psi'_{(X,Y)} \wedge \psi''_{(X,Y)}$.
5. If $\varphi = K\psi$, $\varphi_{(X,Y)} = \mathbf{t}$ if $X \vdash \psi_{(X,Y)}$. Otherwise, $\varphi_{(X,Y)} = \mathbf{f}$.

Next, for a modal theory T we define

$$T_{(X,Y)} = \{\varphi_{(X,Y)} : \varphi \in T\}.$$

It is easy to see that the inductive definition given above yields an algorithm to compute $\varphi_{(X,Y)}$. If we count each call to the propositional provability oracle as one step (the input for each such call is given by X or Y and a subformula of φ), then this algorithm runs in polynomial time in the size of φ. Since the input for each call to the oracle is formed by one of X or Y and a subformula of φ, it follows that the problem to compute $\varphi_{(X,Y)}$, given a modal formula φ and a pair of theories (X, Y), is in the class Δ^2_P.

Our algorithm to compute the Kripke-Kleene and well-founded semantics for autoepistemic and default logics is based on the following result.

Theorem 7.1 *For every belief pair (P, S) and every propositional theories X and Y in the extended language, if $Mod(X, Y) = (P, S)$, then*

$$\mathcal{D}_T(P, S) = Mod(T_{(Y,X)}, T_{(X,Y)}).$$

Proof: We will first show that for every belief pair (P, S), every two propositional theories X and Y (in the extended language), every modal formula φ and every interpretation $I \in \mathcal{A}$,

$$\mathcal{H}^2_{(P,S),I}(\varphi) = I^e(\varphi_{(X,Y)}). \tag{3}$$

We proceed by induction on the length of φ. The claim is evident in the case when $\varphi \in \mathcal{L}$ (this case establishes, in particular, the basis for the induction). Let us assume that $\varphi = \neg\psi$. Then,

$$\mathcal{H}^2_{(P,S),I}(\varphi) = \neg\mathcal{H}^2_{(S,P),I}(\psi) = \neg I^e(\psi_{(Y,X)}) = I^e(\neg\psi_{(Y,X)}) = I^e(\varphi_{(X,Y)})$$

(the second equality is implied by the induction hypothesis, the last one follows from the inductive definition of $\varphi_{(X,Y)}$). A similar reasoning establishes the inductive step for the cases when the main connective in φ is the disjunction or the conjunction.

The last case to consider is that of $\varphi = K\psi$. Let us assume that $\mathcal{H}^2_{(P,S),I}(\varphi) = \mathbf{t}$. Then, $\mathcal{H}^2_{(P,S),J}(\psi) = \mathbf{t}$ for every $J \in P$. By the induction hypothesis we obtain that $J^e(\psi_{(X,Y)}) = \mathbf{t}$ for every $J \in P$. Since $Mod(X) = P$, it follows that $X \vdash \psi_{(X,Y)}$. Thus, $\varphi_{(X,Y)} = \mathbf{t}$ and $I^e(\varphi_{(X,Y)}) = \mathbf{t}$. Conversely, let us assume that $I^e(\varphi_{(X,Y)}) = \mathbf{t}$. Since $\varphi = K\psi$, $\varphi_{(X,Y)} = \mathbf{t}$ or $\mathbf{f}$. The latter case is impossible as $I^e(\varphi_{(X,Y)}) = \mathbf{t}$. Thus, $\varphi_{(X,Y)} = \mathbf{t}$. It follows that $X \vdash \psi_{(X,Y)}$. Since $Mod(X) = P$, for every interpretation $J \in P$, $J^e(\psi_{(X,Y)}) = \mathbf{t}$. By the

induction hypothesis, for every interpretation $J \in P$, $\mathcal{H}^2_{(P,S),J}(\psi) = \mathbf{t}$. Thus, by the definition of the function $\mathcal{H}^2_{(P,S),I}$, $\mathcal{H}^2_{(P,S),I}(\varphi) = \mathbf{t}$. This completes our inductive argument.

We will use (3) to prove the assertion of the theorem. We have

$$\mathcal{D}^l_T(P,S) = \{I\colon \mathcal{H}^2_{(S,P),I}(T) = \mathbf{t}\} = \{I\colon I^e(T_{(Y,X)}) = \mathbf{t}\} = Mod(T_{(Y,X)}).$$

Similarly,

$$\mathcal{D}^u_T(P,S) = \{I\colon \mathcal{H}^2_{(P,S),I}(T) = \mathbf{t}\} = \{I\colon I^e(T_{(X,Y)}) = \mathbf{t}\} = Mod(T_{(X,Y)}).$$

Thus, the assertion follows. □

Let us define an operator on pairs of theories from $\mathcal{L}^e$ by

$$\mathcal{S}_T(X,Y) = (T_{(Y,X)}, T_{(X,Y)}).$$

Directly from Theorem 7.1, it follows that for every pair (X,Y) of propositional theories in the extended language we have:

$$\mathcal{D}_T(Mod(X,Y)) = Mod(S_T(X,Y)). \tag{4}$$

Before we state the next corollary, which is the key to our complexity results, we introduce additional notation. Let L be a set, and let $O : L \to L$ be an operator in L. The iterations of the operator O, $\langle O{\uparrow}^n\rangle_{n\in N}$ are defined inductively as follows. $O{\uparrow}^0$ is the identity operator in L. When $O{\uparrow}^n$ is defined, $O{\uparrow}^{n+1}$ is defined by the condition:

$$O{\uparrow}^{n+1}(x) = O(O{\uparrow}^n(x)).$$

We now have the following corollary.

Corollary 7.2 *Let T be a modal theory. For every $n \geq 0$,*

$$\mathcal{D}_T{\uparrow}^n(\mathcal{A},\emptyset) = Mod(\mathcal{S}_T{\uparrow}^n(\{\mathbf{t}\},\{\mathbf{f}\})).$$

Proof: Clearly, $\mathcal{A} = Mod(\mathbf{t})$ and $\emptyset = Mod(\mathbf{f})$. Thus, the assertion holds for $n = 0$. Let us consider an integer $n \geq 0$ and assume that the assertion holds for n. We have

$$\mathcal{D}_T{\uparrow}^{n+1}(\mathcal{A},\emptyset) = \mathcal{D}_T(\mathcal{D}_T{\uparrow}^n(\mathcal{A},\emptyset)) = \mathcal{D}_T(Mod(\mathcal{S}_T{\uparrow}^n(\{\mathbf{t}\},\{\mathbf{f}\}))$$

(the second equality follows by the induction hypothesis). Now, by (4),

$$\mathcal{D}_T(Mod(\mathcal{S}_T{\uparrow}^n(\{\mathbf{t}\},\{\mathbf{f}\}))) = Mod(\mathcal{S}_T(\mathcal{S}_T{\uparrow}^n(\{\mathbf{t}\},\{\mathbf{f}\}))) = Mod(\mathcal{S}_T{\uparrow}^{n+1}(\{\mathbf{t}\},\{\mathbf{f}\})).$$

Thus, the assertion follows by induction. □

Corollary 7.2 implies an algorithm to compute the Kripke-Kleene expansion (fixpoint) for a modal theory T. Since the operator $\mathcal{D}_T$ is $\preceq_{pr}$-monotone, this fixpoint (which is the least fixpoint of $\mathcal{D}_T$) can be computed by iterating $\mathcal{D}_T$ over the belief pair $(\mathcal{A},\emptyset)$. By Corollary 7.2, this fixpoint (or, more precisely, a pair of theories that represents it) can be computed by iterating the operator $\mathcal{S}_T$ over the pair of theories $(\{\mathbf{t}\},\{\mathbf{f}\})$.

The number of iterations necessary to compute the least fixpoint of $\mathcal{D}_T$ (or, equivalently, $\mathcal{S}_T$) is polynomial in the size of T. Indeed, by the monotonicity of $\mathcal{D}_T$, the sequence of sets $B_i = \mathcal{D}_T{\uparrow}^i(\mathcal{A}, \emptyset)$ is ascending, that is,

$$B_0 \preceq_{pr} B_1 \preceq_{pr} B_2 \preceq_{pr} \ldots \ .$$

Moreover, by Proposition 3.5, all B_is are consistent. Thus, by Proposition 3.2, for every modal atom $K\psi$ of T we have

$$\mathcal{H}^4_{B_0}(K\psi) \preceq_{pr} \mathcal{H}^4_{B_1}(K\psi) \preceq_{pr} \mathcal{H}^4_{B_2}(K\psi) \preceq_{pr} \ \ldots .$$

After no more that $2k$ iterations, where k is the number of maximal modal atoms in T, this latter sequence stabilizes (reaches its limit). Indeed, each modal atom can change its value at most twice (from **f** to **u** or **i** and, then, one more time to **t**). By Proposition 3.8 it follows that if no modal atoms change their values when moving from B_i to B_{i+1}, then $B_{i+1} = B_{i+2}$. Thus, $B_{2k+1} = B_{2k+2}$.

We have already argued before that the task to compute a single iteration of the operator $\mathcal{S}_T$ is in the class Δ^2_P. Thus, we obtain the following result.

Theorem 7.3 *The problem of computing the Kripke-Kleene fixpoint for a given finite modal theory T is in the class Δ^2_P.*

This result was first proved in [DMT99]. The method we presented here is a simplification of the approach from [DMT99]. Moreover, we will now extend it to the case of computing the well-founded fixpoint of a modal theory T.

In order to compute the well-founded fixpoint of T we need to design techniques to compute the stable operator $\mathcal{D}^{st}_T$. Let us recall that $\mathcal{D}^{st}_T(P, S) = (D^{st}_T(S), D^{st}_T(P))$, where $D^{st}_T(S) = lfp(D_{S,T})$ (we recall that $D_{S,T} = \mathcal{D}_T(\cdot, S)$). Thus, we will first focus on computing the operator D^{st}_T.

Let Y be a theory in the language $\mathcal{L}^e$. For every theory $X \subseteq \mathcal{L}^e$, we define

$$S_{Y,T}(X) = T_{(Y,X)}.$$

Theorem 7.1 has the following corollary concerning the connection between the operators $D_{S,T}$ and $S_{Y,T}$ (the proof is straightforward and we omit it).

Corollary 7.4 *Let T be a modal theory, let P and S be possible-world structures and let X and Y be theories in $\mathcal{L}^e$. If $P = Mod(X)$ and $S = Mod(Y)$, then $D_{S,T}(P) = Mod(S_{Y,T}(X))$.*

It follows from Corollary 7.4 that the possible-world structure $D^{st}_T(S)$ (or, to be precise, its representation) can be computed by iterating the operator $S_{Y,T}$, where Y is a representation of S.

Corollary 7.5 *Let T be a modal theory. For every possible-world structure S and every theory $Y \subseteq \mathcal{L}^e$, if $S = Mod(Y)$, then for every $n \geq 0$,*

$$D_{S,T}{\uparrow}^n(\mathcal{A}) = Mod(S_{Y,T}{\uparrow}^n(\{\mathbf{t}\})).$$

Proof: We proceed by induction. Since

$$D_{S,T}\uparrow^0(\mathcal{A}) = \mathcal{A} = Mod(\{\mathbf{t}\}) = Mod(S_{Y,T}\uparrow^0(\{\mathbf{t}\})),$$

the case $n = 0$ is settled. Let us assume that the assertion holds for all integers smaller than or equal to some integer $n \geq 0$. We will show that the assertion holds for $n + 1$. Indeed,

$$\begin{aligned} D_{S,T}\uparrow^{n+1}(\mathcal{A}) &= D_{S,T}(D_{S,T}\uparrow^n(\mathcal{A})) = D_{S,T}(Mod(S_{Y,T}\uparrow^n(\{\mathbf{t}\}))) \\ &= Mod(S_{Y,T}(S_{Y,T}\uparrow^n(\{\mathbf{t}\}))) = Mod(S_{Y,T}\uparrow^{n+1}(\{\mathbf{t}\})). \end{aligned}$$

The second equality follows by the induction hypothesis, the third one is implied by Corollary 7.4. □

The sequence $D_{S,T}\uparrow^n(\mathcal{A})$ is ascending and it stabilizes after no more than $2k+1$ iterations, where k is the number of maximal modal atoms occurring in T (a similar argument as the one we used for $\mathcal{D}_T\uparrow^n(\mathcal{A}, \emptyset)$ works in this case, as well). Thus, the sequence $S_{Y,T}\uparrow^n(\{\mathbf{t}\})$ stabilizes after no more than $2k + 1$ iterations, too. Let us denote this limit by $S^{st}_T(Y)$.

The problem of computing the value $S_{Y,T}(X)$ is in the class Δ^2_P (it follows from our earlier remarks on the complexity of computing $\mathcal{S}_T$). Thus, it follows from Corollary 7.5 that we can compute $S^{st}_T(Y)$ (which is a representation of the possible-world structure $D^{st}(S)$) by iterating the operator $S_{Y,T}$ (where Y is a representation of S). Since the number of iterations is polynomial in the size of T and Y, the task of computing (a representation of) $D^{st}(S)$ is in Δ^2_P (assuming that S is given in terms of its representation Y).

For a pair of theories (X, Y) in the language $\mathcal{L}^e$, we define

$$\mathcal{S}^{st}_T(X, Y) = (S^{st}_T(Y), S^{st}_T(X)).$$

Let us consider possible-world structures P and S and theories X and Y in $\mathcal{L}^e$ such that $P = Mod(X)$ and $S = Mod(Y)$. Then, by Corollary 7.4,

$$\mathcal{D}^{st}(P, S) = (D^{st}_T(S), D^{st}_T(P)) = (Mod(S^{st}_T(Y), Mod(S^{st}_T(X))) = Mod(\mathcal{S}^{st}_T(X, Y)).$$

It follows, that the problem of computing a representation of $\mathcal{D}^{st}(P, S)$ (that is, $\mathcal{S}^{st}_T(X, Y)$), given representations X and Y of P and S, respectively, is in the class Δ^2_P.

The well-founded fixpoint of $\mathcal{D}^{st}$ can be computed by iterating the operator $\mathcal{D}^{st}$ starting with the belief pair $(\mathcal{A}, \emptyset)$. Our discussion implies that its representation can be computed by iterating the operator $\mathcal{S}^{st}$ and starting with the pair of theories $(\{\mathbf{t}\}, \{\mathbf{f}\})$. The number of iterations is bounded by $2k + 1$, where k is the number of maximal modal atoms occurring in T (the same argument as in the case of the Kripke-Kleene fixpoint computations applies). Thus, we obtain the following theorem.

Theorem 7.6 *The problem of computing the well-founded fixpoint for a given finite modal theory T is in the class Δ^2_P.*

We proved in the previous section that default theories can be translated into equivalent modal theories (Theorem 6.3). Thus, Theorems 7.3 and 7.6 have the following corollary.

Corollary 7.7 *The problem to compute the Kripke-Kleene fixpoint (respectively, the well-founded fixpoint) for a given default theory Δ is in the class Δ^2_P.*

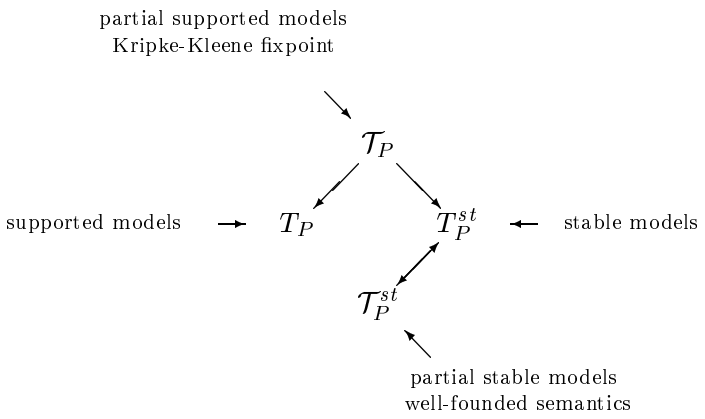

Figure 7: Operators associated with logic programming

The problem to decide whether a default theory Δ has an expansion (respectively, extension) is complete for the class Σ_P^2. The problem to compute an expansion (extension) for Δ is Σ_P^2-hard. These results were obtained by Gottlob [Got92] and, independently, by Stillman [Sti90]. The corresponding problems concerning expansions and extensions of autoepistemic theories have the same complexity [Got92]. Theorems 7.3 and 7.6 and Corollary 7.7 indicate that the complexity of the problems to compute Kripke-Kleene and well-founded fixpoints of autoepistemic and default theories have lower complexity (assuming the polynomial hierarchy does not collapse). Thus, these approximation semantics are computationally more attractive than the semantics of (two-valued) expansions and extensions.

8 Discussion and future work

We presented results uncovering the semantic properties of default and autoepistemic logics. For each logic, we introduced an operator describing how to revise belief pairs when constructing belief sets, and derived from this operator a whole family of semantics. We obtained these semantics by purely algebraic transformations reflecting basic principles of approximating belief sets. Some of these semantics (Kripke-Kleene and well-founded semantics) have a constructive flavor and are more amenable to computational treatment.

We also demonstrated that the modal interpretation of defaults proposed by Konolige establishes a perfect correspondence between the *families* of semantics of default and autoepistemic logics. This elegant picture can be further extended to the case of logic programming with negation. Based on the work of Fitting[Fit02], it was shown in [DMT00b] that all key semantics for logic programs can be similarly obtained from the four-valued operator $\mathcal{T}_P$ generalizing the original van Emden-Kowalski one-step provability operator T_P [vEK76]. The resulting structure of main semantics of logic programs is shown in Figure 7.

The operator $\mathcal{T}_P$ is a counterpart to the operators $\mathcal{D}_T$ and $\mathcal{E}_\Delta$. Indeed, the translation of logic program clauses into default rules proposed in [BF91, MT89b] establishes an embedding of logic programming into default logic that precisely aligns the corresponding semantics (Figure 8).

Let us further note that the approach to semantics of nonmonotonic logics based on the

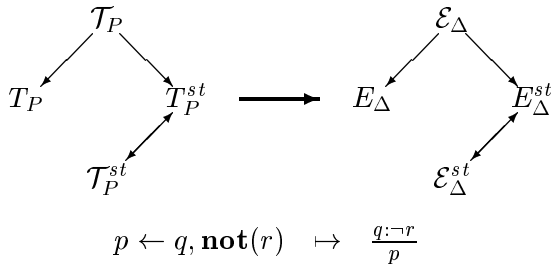

Figure 8: Embedding logic programming in default logic

concept of a belief pair can also be extended to the case of reflexive autoepistemic logic introduced by Schwarz [Sch91]. As in all other cases discussed in this paper, all major semantics for the reflexive autoepistemic logic can be obtained from a single operator on the lattice $\mathcal{B}$ (the definition of the operator remains essentially the same as in the case of autoepistemic logic — what changes is the definition of the truth function $\mathcal{H}^4_{(P,S),I}$).

As pointed in the Section 1, in early 80s, McDermott and Doyle proposed a general approach to define modal nonmonotonic logics [MD80, McD82]. It is known that autoepistemic logic can be obtained within the framework of McDermott and Doyle from the modal logic KD45 [Shv90, MT93]. In [Sch91] Schwarz proved that reflexive autoepistemic logic can similarly be obtained from modal logic SW5. Our results show that both logics can be given an algebraic treatment based on the concept of a belief pair. An interesting question is whether other modal nonmonotonic logics in the McDermott and Doyle's scheme are amenable to such an approach.

Another question concerning the McDermott and Doyle's scheme is whether the semantics of *extensions* for autoepistemic logic can be reconstructed within it as a modal nonmonotonic logic corresponding to some appropriately chosen underlying modal logic. The answer to this question is negative. The modal theory T, where

$$T = \{\neg Kp \Rightarrow p,\ Kp \Rightarrow p\},$$

has no extensions. The easiest way to see it is to observe that $T = m(\Delta)$, where

$$\Delta = \left(\emptyset, \{\frac{p:}{p}, \frac{:\neg p}{p}\}\right).$$

Since Δ has no extensions, Theorem 6.3 implies that $m(T)$ has no extensions either. However, for every modal logic $\mathcal{S}$, T has at least one $\mathcal{S}$-expansion (in the sense of McDermott and Doyle). In particular, the set of consequences of $\{p\}$ in the logic S5 is such an expansion [MT93].

Finally, let us mention that it is possible to develop an abstract, purely algebraic treatment of the concept of an approximation of an operator. It generalizes the approach presented here and the work of Fitting on logic programming semantics. An account of this abstract treatment of approximations can be found in [DMT00a].

Acknowledgments

Marc Denecker was supported in part by the Research Fund of the K.U. Leuven, Belgium. Victor W. Marek and Mirosław Truszczyński were supported by the National Science Foundation under Grants No. 9874764 and 0097278. Any opinions, findings, and conclusions or

recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[Ant97] G. Antoniou. *Nonmonotonic Reasoning.* MIT Press, 1997.

[BS91] C. Baral and V.S. Subrahmanian. Dualities between alternative semantics for logic programming and nonmonotonic reasoning (extended abstract). In A. Nerode, W. Marek, and V.S. Subrahmanian, editors, *Logic programming and non-monotonic reasoning (Washington, DC, 1991)*, pages 69–86, Cambridge, MA, 1991. MIT Press.

[Bes89] P. Besnard. *An Introduction to Default Logic.* Springer-Verlag, Berlin, 1989.

[BS94] P. Besnard and T. Schaub. Possible world semantics for default logic. *Fundamenta Informaticae*, 21:39–66, 1994.

[BF91] N. Bidoit and C. Froidevaux. Negation by default and unstratifiable logic programs. *Theoretical Computer Science*, 78(1, (Part B)):85–112, 1991.

[Che80] B.F. Chellas. *Modal logic. An introduction.* Cambridge University Press, Cambridge-New York, 1980.

[DMT99] M. Denecker, V. Marek, and M. Truszczyński. Fixpoint 3-valued semantics for autoepistemic logic. In *Logical Foundations for Cognitive Agents: Contributions in Honor of Ray Reiter*, pages 113 – 136. Springer-Verlag, 1999.

[DMT00a] M. Denecker, V. Marek, and M. Truszczyński. Unified semantic treatment of default and autoepistemic logics. In *Principles of Knowledge Representation and Reasoning, Proceedings of the Seventh International Conference (KR2000)*, pages 74 – 84. Morgan Kaufmann Publishers, 2000.

[DMT00b] M. Denecker, V. Marek, and M. Truszczyński. Approximating operators, stable operators, well-founded fixpoints and applications in nonmonotonic reasoning. In *Logic-based Artificial Intelligence*, chapter 6, pages 127 – 144. Kluwer Academic Publishers, Boston, 2000.

[Fit02] M. C. Fitting. Fixpoint semantics for logic programming – a survey. *Theoretical Computer Science* 278:25–51, 2002.

[Gel87] M. Gelfond. On stratified autoepistemic theories. In *Proceedings of AAAI-87*, pages 207–211. Morgan Kaufmann, 1987.

[Got92] G. Gottlob. Complexity results for nonmonotonic logics. *Journal of Logic and Computation*, 2(3):397–425, 1992.

[Got95] G. Gottlob. Translating default logic into standard autoepistemic logic. *Journal of the ACM*, 42(4):711–740, 1995.

[GC90] R. Guerreiro and M. Casanova. An alternative semantics for default logic. Preprint. The Third International Workshop on Nonmonotonic Reasoning, South Lake Tahoe, 1990.

[HC84] G.E. Hughes and M.J. Cresswell. *A companion to modal logic*. Methuen and Co., 1984.

[Kon88] K. Konolige. On the relation between default and autoepistemic logic. *Artificial Intelligence*, 35(3):343–382, 1988.

[Lev90] H. J. Levesque. All I know: a study in autoepistemic logic. *Artificial Intelligence*, 42(2-3):263–309, 1990.

[Lif90] V. Lifschitz. On open defaults. In J. Lloyd, editor, *Proceedings of the Symposium on Computational Logic*, pages 80–95. Springer-Verlag, 1990.

[MT89a] W. Marek and M. Truszczyński. Relating autoepistemic and default logics. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning (Toronto, ON, 1989)*, pages 276–288, Morgan Kaufmann, 1989.

[MT89b] W. Marek and M. Truszczyński. Stable semantics for logic programs and default theories. In E.Lusk and R. Overbeek, editors, *Proceedings of the North American Conference on Logic Programming*, pages 243–256. MIT Press, 1989.

[MT93] W. Marek and M. Truszczyński. *Nonmonotonic Logic; Context-Dependent Reasoning*. Springer-Verlag, 1993.

[McD82] D. McDermott. Nonmonotonic logic II: nonmonotonic modal theories. *Journal of the ACM*, 29(1):33–57, 1982.

[MD80] D. McDermott and J. Doyle. Nonmonotonic logic I. *Artificial Intelligence*, 13(1-2):41–72, 1980.

[Moo84] 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.

[Moo85] R.C. Moore. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, 25(1):75–94, 1985.

[PP94] H. Przymusińska and T. Przymusiński. Stationary default extensions. *Fundamenta Informaticae*, 21(1-2):67–87, 1994.

[Prz90] T.C. Przymusinski. The well-founded semantics coincides with the three-valued stable semantics. *Fundamenta Informaticae*, 13(4):445–464, 1990.

[Rei80] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1-2):81–132, 1980.

[Sch91] 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, MIT Press, 1991.

[Shv90] G.F. Shvarts. Autoepistemic modal logics. In R. Parikh, editor, *Theoretical aspects of reasoning about knowledge (Pacific Grove, CA, 1990)*, pages 97–109, Morgan Kaufmann, 1990.

[Sti90] J. Stillman. It is not my default: the complexity of membership problem for restricted propositional default logics. In *Proceedings of AAAI-90*, Los Altos, CA, 1990. American Association for Artificial Intelligence, Morgan Kaufmann.

[Tru91] M. Truszczyński. Modal interpretations of default logic. In *Proceedings of IJCAI-91*, pages 393–398, Morgan Kaufmann, 1991.

[vEK76] M.H. van Emden and R.A. Kowalski. The semantics of predicate logic as a programming language. *Journal of the ACM*, 23(4):733–742, 1976.

[VRS91] 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.