A Konolige bridge between default logic and autoepistemic logic

MSc Thesis (*Afstudeerscriptie*)

written by

Martijn Pennings
(born February 12, 1981 in Sittard, the Netherlands)

under the supervision of **Dr. Sujata Ghosh** and **Prof.dr. Frank Veltman**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:	**Members of the Thesis Committee:**
August 25, 2008	Prof.dr. Peter van Emde Boas
	Dr. Sujata Ghosh
	Prof.dr. Michiel van Lambalgen
	Prof.dr. Frank Veltman

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract

Default Logic and Autoepistemic Logic are two forms of nonmonotonic logic originating from the works of Reiter in 1980 and Moore in 1985, respectively. Konolige showed that via a translation from defaults to modal formulas default logic can be expressed in autoepistemic logic. However, no translation has been found that relates the originally used semantics for default logic (extensions) precisely to the originally used semantics for autoepistemic logic (expansions).

In a series of papers, Denecker, Marek and Truszczyński revealed a uniform semantics for default logic and autoepistemic logic, which we will explain. They defined several operators on the lattices of possible world structures relating to various semantics for default logic and autoepistemic logic. This way they defined four connected semantics for each of the two logics, among those the two semantics of extensions and expansions. Hence, they revealed the connection between the two semantics (using the Konolige translation from defaults to modal formulas), but also the difference.

We will also examine a form of prioritized versions of default logic and autoepistemic logic inspired by Rintanen, in which preference structures on defaults or formulas relate to preference structures on extensions or expansion. We will explain how the Konolige translation can be used to relate the prioritized version of default logic to the prioritized version of autoepistemic logic.

Acknowledgements

There are a few people I would like to thank for their help in making this thesis or helping me during the making of this thesis. They are truly some of the nicest people I have met in nine years of living and studying in Amsterdam.

I would like to thank my first first supervisor Sujata Ghosh, for her enthusiasm, her guidance and her patience. Also, for giving me a bit of confidence by assuring me that what I did was all right and to tell me not to give up. And last but not least for getting me interested in the subject of non-monotonic logic in the first place.

I would like to thank my second first supervisor Frank Veltman, for guiding me through the last months of the process, especially for letting me know that my thesis was good enough to graduate and giving me a deadline. That really pushed me over the last hurdle.

I would like to thank Benedikt Löwe for "luring" me into the Master of Logic programme a few years ago. He is a great teacher and in just a few conversations about theses, studying and life in general has shown to be a great mentor. As for not starting a Master of Logic thesis with Aristotle, I did it anyway.

I am very thankful for the existence of the "scriptieklasje". After hitting a low in the progress of writing this thesis, this group of struggling students gave me exactly what I needed: structure, confidence, and the feeling that I wasn't alone in my struggles of writing a thesis. I would recommend it to anyone.

I thank my parents, my sisters, my friends in Amsterdam, my friends in Limburg, my fellow students at the ILLC. For their support, their faith in me, and just being there.

Amsterdam
June 25, 2008

Contents

1 Introduction

Probably the most widely used examples of logical reasoning are Aristotle's syllogisms, such as

$$\frac{\begin{array}{l}\text{All men are mortal}\\ \text{Socrates is a man}\end{array}}{\text{Socrates is mortal}}$$

There's no denying that this conclusion, given the two premisses, is irrefutable. Even more so, *whenever* one draws a conclusion from a set of premises in classical logic, this conclusion will not be refuted by any additional information. This property of classical logic is called *monotonicity of entailment*:

Definition 1.1. *An entailment relation* $\vDash$ *is* monotone *if whenever* $\Pi \vDash \varphi$ *then also* $\Pi \cup A \vDash \varphi$.

It seems very reasonable for an entailment relation. However, every-day reasoning is usually not that simple; we simplify, generalize and jump to conclusions all the time. One of the most used example in non-monotonic reasoning is that of Tweety:

$$\frac{\begin{array}{l}\text{Birds fly}\\ \text{Tweety is a bird}\end{array}}{\text{Tweety flies}}$$

This seems reasonable, but if we would add the following information, our conclusion above is refuted.

$$\frac{\begin{array}{l}\text{Penguins don't fly}\\ \text{Tweety is a penguin}\end{array}}{\text{Tweety doesn't fly}}$$

Knowing that penguins are also birds, what should we conclude from these argumentations? Should we use the first line of argument and conclude that Tweety flies or use the second and conclude that Tweety doesn't? The statements "Birds fly" and "Penguins don't fly" are of course generalizations and thus we could say that these logical deductions are simply incorrect, because the premises are incorrect. However, in every day life such jumps to conclusions are made all the time and even necessary. This is because we live in a state of *incomplete information*.

As an example, suppose we would want to talk to someone at the Universiteit van Amsterdam (UvA). We don't know anything about this person, except that we meet him at this Dutch university; obviously we are in a state of incomplete information. We figure that since people in the Netherlands are usually Dutch we will address him in Dutch. However, approaching him we see a name tag of the ILLC (the Institute of Logic, Language and Computation at the UvA) on his sweater after which we decide to address him in English, since that is the going language at this internationally orientated institute. Either way, we have

to make some assumptions, some generalizations, some jumps to conclusion, otherwise the conversation will never start off. As well as the Tweety example, which we will use a lot in the following text, the above is an example of non-monotonic reasoning; after adding some information (seeing the name tag of the ILLC), our previous conclusion (that we should address him in Dutch) gets refuted.

To model the behavior of the examples above calls for a non-classical approach. It can not be dealt with by classical logics, because (as Moore explains in [15]):

> "As Minsky [14] has pointed out, standard logics are always monotonic, because their inference rules make every axiom permissive. That is, the inference rules are always of the form "P is a theorem if $Q_1, \ldots, Q_n$ are theorems", so new axioms can only make more theorems derivable; they can never result in a previous theorem being invalidated.
> $(\ldots)$
> The general idea is to allow axioms to be restrictive as well as permissive, by employing inference rules of the form "P is a theorem if $Q_1, \ldots, Q_n$ are not theorems". The inference that birds can fly handled by having, in effect, a rule that says that for any X, "X can fly" is a theorem if "X is a bird" is a theorem and "X cannot fly" is not a theorem. If all we are told about Tweety is that he is a bird, then we will not be able to derive "Tweety cannot fly", and the inference to "Tweety can fly" will go through. If we are told that Tweety is a penguin and we know that no penguin can fly, we will be able to derive the fact that Tweety cannot fly, and the inference that Tweety can fly will be blocked."

Default logic and autoepistemic logic are two forms of non-monotonic logics dealing with these kinds of problems. They originate from papers by Raymond Reiter [17] in 1980 and Robert C. Moore [15] in 1985 respectively and since have been studied extensively, exposing both the similarities and the differences between the two formalisms.

The thesis is organized as follows: chapter 2 contains the preliminaries of default logic and autoepistemic logic. The usual semantics for these logics are treated: Reiter's extensions for default logic and Moore's expansions for autoepistemic logic. The principle properties of the two formalisms are proven and a link between the two is given in the form of the Konolige translation.

Chapter 3 explains mostly the writings of Denecker, Marek and Truszczyński, as it explains how default logic and autoepistemic logic can be treated in terms of possible world structures. Characterizations of extensions for default logic and expansions for autoepistemic logic in terms of possible world structures are given with new, straightforward and detailed proofs. For both formalisms four semantics are treated, among which weak extensions for default logic and extensions for autoepistemic logic. It is explained how to link the four semantics of

these logics together with the Konolige translation. A short description is presented of the theory Denecker et al. used for their work, namely Approximation Theory. The chapter is closed of by a short description of the link with Logic Programming. A translation of logic programming with negation to default logic is presented and how to link different semantics for logic programming to the notions of (weak) extensions and expansions.

In chapter 4 the notion of prioritization is explained, after which Rintanen's lexicographic prioritization is explained for default logic and expanded to autoepistemic logic. Again, the Konolige translation is used to show the link between prioritized default logic and prioritized autoepistemic logic.

The thesis is closed off by a chapter on topics of future research.

2 Preliminaries on default and autoepistemic logic

For any set of formulas T, we will denote the set of classical consequences of T by $Th(T)$. This is not to be confused with $Th_L(Q)$ for possible world structures Q, which will all be defined later on.

2.1 Default logic

Reiter's default logic contains, apart from the usual propositional language, denoted by $\mathcal{L}$, inference rules called *default rules* or *defaults*. Default rules are rules-of-thumb, such as "birds fly" in the Tweety example. When using such a rule-of-thumb, we can explain it in numerous ways, such as

- "birds usually fly", or
- "most birds fly", or
- "when encountering a bird we will by default assume that it flies", or
- "all birds fly, except for penguins, ostriches, Big Bird, ...".

Reiter chose to go with the following approach:

$$\frac{bird_{Tweety} : flies_{Tweety}}{flies_{Tweety}}$$

which is read as "if we know that Tweety is a bird and it is consistent to assume that Tweety flies, then conclude that Tweety flies". It is specified below what it means to "know" that Tweety flies and with what exactly $flies_{Tweety}$ has to be consistent. Formally, we define defaults as follows:

Definition 2.1. *A* default rule *(or* default*)* δ *is of the form*

$$\frac{\varphi : \psi_1, \ldots, \psi_n}{\chi},$$

with $\varphi, \psi_1, \ldots, \psi_n$ *and* χ *propositional formulas. We refer to* φ *as the* prerequisite of δ *and denote this by* $pre(\delta)$*. Also,* $\{\psi_1, \ldots, \psi_n\}$ *is the set of* justifications of δ *(denoted by* $just(\delta)$*) and* χ *the* conclusion of δ *(denoted by* $con(\delta)$*). A* default theory Δ *is of the form* (D, W)*, where* W *is a a set of propositional formulas (the set of facts) and* D *is a set of defaults.*

Reiter defined the semantics of default logic in terms of *extensions*. Given a set of facts W and a set of default rules D these extensions are meant to formally specify an agent's set of beliefs. The extensions will be interpreted as acceptable sets of beliefs that one may hold about the incompletely specified world W. Reiter defined these belief sets in such a way that

1. they include the set of facts about the world W,

2. they are deductively closed, i.e. closed under propositional consequence (hence the agents we would be dealing with are perfect rational agents),

3. they are closed under the default rules in D (in a sense to be explained below).

Formally, we have the following definition of extensions:

Definition 2.2. *Let $\Delta = (D, W)$ be a default theory. For any set of formulas S, let $\Gamma_\Delta(S)$ be the least set of formulas such that*

D1 $W \subseteq \Gamma_\Delta(S)$,

D2 $Th(\Gamma_\Delta(S)) = \Gamma_\Delta(S)$ *(Recall that by $Th(T)$ we mean the set of classical consequences of T),*

D3 for any default $\frac{\varphi:\psi_1,\ldots,\psi_n}{\chi} \in D$, if $\varphi \in \Gamma_\Delta(S)$ and $S \nvdash \neg\psi_i$ for all $i \in \{1,\ldots,n\}$ then $\chi \in \Gamma_\Delta(S)$.

A set of formulas E is an extension *of Δ if and only if it is a fixpoint of the operator Γ_Δ, i.e. if and only if $\Gamma_\Delta(E) = E$.*

For a set to be "closed under the default rules in D" thus means that every default who's prerequisite is in the set and who's justifications are consistent with it has also it's conclusion in the set.

Note that the definition of extensions is not a constructive one, since there is a self-reference in **D3**. This means that we have to *guess* a set E and subsequently check wether it fulfills the requirements **D1** to **D3**. Alternative constructive characterizations for extensions have been proposed in (among others) [17] and [1].

Example 2.3. To illustrate the definitions, consider the Tweety example. The given facts are that Tweety is a bird and that Tweety is a penguin; these are described in the set of facts $W\{bird, penguin\}$. The rules-of-thumb are:

- "if Tweety is a bird and it is consistent to assume that Tweety flies, then we conclude that Tweety flies",
- "if Tweety is a penguin and it is consistent to assume that Tweety doesn't fly, then we conclude that Tweety doesn't fly".

These rules-of-thumb are described by the default rules in the set of defaults

$$D = \{\delta_1 = \frac{bird : flies}{flies}, \delta_2 = \frac{penguin : \neg flies}{\neg flies}\}.$$

The default theory we are dealing with is $\Delta = (D, W)$. Since we know that Tweety is a bird and a penguin, but we doubt whether he flies or not, we consider the following sets as possible extensions: $E_1 = Th(\{bird, penguin, flies\})$ and $E_2 = Th(\{bird, penguin, \neg flies\})$. The set of facts W is clearly included in both

E_1 and E_2 and also both sets are clearly deductively closed, so items **D1** and **D2** are obviously satisfied. Now let's look at **D3**. Since $pre(\delta_1) = bird \in E_1$ and $E_1 \nvdash \neg flies = \neg just(\delta_1)$, item **D3** requires that $con(\delta_1) = flies \in E_1$, which is satisfied. Moreover, since $E_1 \vdash flies = \neg just(\delta_2)$, δ_2 vacuously satisfies **D3** for E_1. Since there is no smaller set than E_1 itself that satisfies all items we can conclude that E_1 is indeed a fixpoint of Γ_Δ and thus is an extension. We can also conclude that E_2 is an extension, since now δ_2 satisfies **D3** by "application", whereas δ_1 satisfies it "vacuously".

Just to see where it leads to, consider W as a possible extension. Then by **D1** and **D2** $\Gamma_\Delta(W)$ would have to include $Th(W)$, thus in any case $pre(\delta_1)$ and $pre(\delta_2)$ are included in $\Gamma_\Delta(W)$. Then, since $W \nvdash \neg flies$ and $W \nvdash \neg\neg flies$, both $flies$ and $\neg flies$ must be included in $\Gamma_\Delta(W)$ by **D3**, thus $\Gamma_\Delta(W) = \mathcal{L}$ and thus W is not an extension. □

This example immediately illustrates that default logic can't always lead to straight answers: we still don't know whether Tweety flies or not! And indeed, we will not know until we somehow decide which default to prefer over the other. We will come to this subject in the chapter 4 on prioritizing default logic and autoepistemic logic.

2.1.1 Properties of default logic

The following notable properties of default logic provide for a better understanding of default logic and will be used in later sections. The most notable property of default logic is of course it's nonmonotonic nature. As an example, the default theory $(\{\frac{true:a}{a}\}, \emptyset)$ has the extension $Th(\{a\})$, whereas the extended default theory $(\{\frac{true:a}{a}\}, \{\neg a\})$ has the extension $Th(\{\neg a\})$; *the addition of new information can cause old information to be refuted.* As we mentioned in the previous section, definition 2.2 is non-constructive, hence extensions must be guessed and tested in stead of constructed. However, the following theorem from [17] can help in deciding which sets to choose as possible extensions.

Theorem 2.4. *Let $E \subseteq \mathcal{L}$ be a set of formulas and $\Delta = (W, D)$ a default theory. Then E is an extension of Δ iff $E = \bigcup_n E_n$, where $E_0 = W$ and*

$$E_{n+1} = Th(E_n) \cup \{con(\delta) \mid \delta \in D \wedge pre(\delta) \in E_n \wedge \forall \psi_i \in just(\delta)(\neg\psi_i \notin E)\}.$$

Proof: Firstly, it is easy to see that $\bigcup_n E_n$ satisfies:

D*i* $W \subseteq \bigcup_n E_n$,

D*ii* $Th(\bigcup_n E_n) = \bigcup_n E_n$, and

D*iii* for any default $\frac{\varphi:\psi_1,\ldots,\psi_m}{\chi}$, if $\varphi \in \bigcup_n E_n$ and $E \nvdash \neg\psi_i$ for all $i \in \{1,\ldots,m\}$ then $\chi \in \bigcup_n E_n$.

Since $\Gamma_\Delta(E)$ is the minimal set with the properties **D*i*** to **D*iii***, we get $\Gamma_\Delta(E) \subseteq \bigcup_n E_n$. Now, for the proof of the theorem:

"⇒": Suppose E is an extension of Δ, i.e. $\Gamma_\Delta(E) = E$. We need to show that $E = \bigcup_n E_n$, but by $E = \Gamma_\Delta(E) \subseteq \bigcup_n E_n$ we only have left to show that $\bigcup_n E_n \subseteq \Gamma_\Delta(E) = E$, which we will do by induction on n. Firstly it is obvious (by requirement **D1**) that $E_0 = W \subseteq \Gamma_\Delta(E)$. Now assume that $E_n \subseteq E$ and let $\xi \in E_{n+1}$. If $\xi \in Th(E_n)$ then by assumption $\xi \in Th(E) = E$. Otherwise ξ is the conclusion of a default $\frac{\alpha:\beta_1,\ldots,\beta_p}{\xi} \in D$ such that $\alpha \in E_n$ and $\beta_1, \ldots, \beta_p \notin E$. By assumption $\alpha \in E_n \subseteq E$ and so by **D3** also $\xi \in E$. This concludes $\bigcup_n E_n \subseteq \Gamma_\Delta(E)(= E)$.

"⇐": Suppose that $E = \bigcup_n E_n$, then we need show that E is an extension of Δ, but by $\Gamma_\Delta(E) \subseteq \bigcup_n E_n$ we only have left to show that $\bigcup_n E_n \subseteq \Gamma_\Delta(E)$, which we will again do by induction on n. Again it is obvious (by requirement **D1**) that $E_0 = W \subseteq \Gamma_\Delta(E)$. Now assume that $E_n \subseteq \Gamma_\Delta(E)$ and let $\xi \in E_{n+1}$. If $\xi \in Th(E_n)$ then by assumption $\xi \in Th(\Gamma_\Delta(E)) = \Gamma_\Delta(E)$. Otherwise ξ is the conclusion of a default $\frac{\alpha:\beta_1,\ldots,\beta_p}{\xi} \in D$ such that $\alpha \in E_n$ and $\beta_1, \ldots, \beta_p \notin E$. Then again by assumption $\alpha \in E_n \subseteq \Gamma_\Delta(E)$ and by **D3** then also $\xi \in \Gamma_\Delta(E)$. This concludes $(E =) \bigcup_n E_n \subseteq \Gamma_\Delta(E)$

The above shows the claim. □

Consider the following example as an illustration of theorem 2.4:

Example 2.5. Let $\Delta = \{D, W\}$ be a default theory consisting of the set of facts $W = \{\neg p \vee \neg q\}$ and the set of defaults $D = \{\frac{true:p}{p}, \frac{true:q}{q}\}$. For any extension E of Δ, E_0 as defined in the theorem would of course be equal to W. Now, since the prerequisites of both defaults are *true* and no negations of justifications are in E_0, it is definitely safe to add any of the two consequences to the candidate extension E. For instance, we could choose $E_1 = Th(E_0) \cup \{p\}$. In the next step however, we could not add the consequence of the second default. Namely, E_1 contains $\neg p \vee \neg q$ and p, and so E_2 would contain $\neg q$. It is clear that the second default can then not be applied. Similarly, if we choose $E_1 = Th(E_0) \cup \{q\}$ by applying the second default, the first default can subsequently not be applied. Thus the extensions are $Th(W \cup \{p\})$ and $Th(W \cup \{q\})$. □

Lemma 2.6. *Let E and F be extensions of a default theory (D, W). If $E \subseteq F$ then $E = F$.*

Proof: Let $E = \bigcup_n E_n$ and $F = \bigcup_n F_n$ as given by theorem 2.4 and assume that $E \subseteq F$. Inductively on n we will show that $F_n \subseteq E_n$ and thus that $F \subseteq E$. Trivially $F_0 \subseteq E_0 \subseteq E$. Suppose $F_n \subseteq E_n$ and let $\xi \in F_{n+1}$. If $\xi \in Th(F_n)$ then by the assumption $\xi \in Th(E_n) \subseteq E_{n+1}$. Otherwise ξ is the conclusion of a default $\frac{\alpha:\beta_1,\ldots,\beta_p}{\xi} \in D$ such that $\alpha \in F_n$ and $\beta_1, \ldots, \beta_p \notin F$. Then again by the assumptions $F_n \subseteq E_n$ and $E \subseteq F$ we get that $\alpha \in E_n$ and $\beta_1, \ldots, \beta_p \notin E$, hence $\xi \in E_{n+1}$. This concludes $F_{n+1} \subseteq E_{n+1}$. □

Lemma 2.7. *A default theory $\Delta = (D, W)$ has an inconsistent extension if and only if W is inconsistent.*

Proof: If W is inconsistent then $Th(W) = \mathcal{L}$ is obviously equal to $\Gamma_\Delta(Th(W))$, hence $\mathcal{L}$ is an inconsistent extension of Δ. Conversely, let E be an inconsistent extension of Δ. Then $E = \bigcup_n E_n$ as given by theorem 2.4, and thus there is a minimal n such that E_n is inconsistent. Suppose this $n > 0$, so E_{n-1} is consistent. Let $\delta \in D$. Then $just(\delta) \in E$, since E is inconsistent (and thus equal to $\mathcal{L}$). Therefore $con(\delta)$ will not be included in E_n and so $E_n = Th(E_{n-1})$, which is consistent. This concludes that n must be 0, i.e. $E_0 = W$ is inconsistent. □

Corollary 2.8. *If a default theory has an inconsistent extension then this is its only extension.*

Proof: Immediate from lemma 2.6 and lemma 2.7. □

If E is an extension of a default theory $\Delta = (D, W)$ then the set $GD(E, \Delta) = \{\delta \in D \mid pre(\delta) \in E \text{ and } just(\delta) \notin E\}$ is the set of *generating defaults* for E with respect to Δ.

Theorem 2.9. *If E is an extension of a default theory $\Delta = (D, W)$ then $E = Th(W \cup \{con(\delta) \mid \delta \in GD(E, \Delta)\})$.*

Proof: Let $E = \bigcup_n E_n$ as given by theorem 2.4. Inductively on n we will show that $E_n \subseteq Th(W \cup \{con(\delta) \mid \delta \in GD(E, \Delta)\})$. This is obvious for $E_0 = W$, so assume that the inclusion holds for E_n and let $\chi \in E_{n+1}$. Then if $\chi \in Th(E_n)$ it is obviously included in $Th(W \cup \{con(\delta) \mid \delta \in GD(E, \Delta)\})$ by the assumption. If not then there must be a default $\frac{\varphi : \psi_1, \ldots, \psi_m}{\chi} \in D$ with $\varphi \in E$ and $\psi_1, \ldots, \psi_m \notin E$. But then obviously $\frac{\varphi : \psi_1, \ldots, \psi_m}{\chi}$ is a generating default, hence $\chi \in \{con(\delta) \mid \delta \in GD(E, \Delta)\})$, which concludes the inclusion $E \subseteq Th(W \cup \{con(\delta) \mid \delta \in GD(E, \Delta)\})$.

As for the reverse inclusion, it is obvious from the definition of extensions that if δ is a generating default for an extension E then $con(\delta) \in E$. Then, for the other inclusion, it follows that $\{con(\delta) \mid \delta \in GD(E, \Delta)\} \subseteq E$. Subsequently it is obvious that $Th(W \cup \{con(\delta) \mid \delta \in GD(E, \Delta)\}) \subseteq E$. □

2.2 Autoepistemic logic

In contrast to Reiter's default logic, Moore's autoepistemic logic doesn't contain non-standard default rules, but contains a modal operator L, which stands for *knowing* something, in order for an agent to *reason about his own knowledge*[1]. We will denote the autoepistemic language, i.e. the language $\mathcal{L}$ of propositional logic plus the modal operator L, by $\mathcal{L}_{ae}$. Formulas of $\mathcal{L}_{ae}$ are called *ae-formulas*

[1] In some literature (for example that of Marek and Truszczyński) the letter K is used in stead of L. Also, M is often used as the dual of L or K; we will simply write $\neg L \neg$ instead.

and sets of ae-formulas are called *ae-theories*. The semantics of autoepistemic logic are, similar to default logic, defined in knowledge sets which are here called *stable expansions*, or simply *expansions*. The key property of these knowledge sets are *positive and negative introspection*. This means that if you know φ then you know that you know φ (positive introspection) and if you don't know φ then you know that you don't know φ (negative introspection). Given some set of facts T, we thus want expansions of T (the knowledge induced by T) to include LT, $L\neg T^c$, LLT, and so on. (For any set of formulas $S \subseteq \mathcal{L}_{ae}$ we write S^c for $\mathcal{L}_{ae} \backslash S$.)

Definition 2.10. *Let T and E be sets of ae-formulas. Denote $\{L\varphi \mid \varphi \in E\}$ by LE and $\{\neg L\varphi \mid \varphi \notin E\}$ by $\neg LE^c$. Also, let $\Omega_T(E) = \{\varphi \mid T \cup LE \cup \neg LE^c \vDash \varphi\}$. Then E is called an* expansion of T *if and only if $E = \Omega_T(E)$.*

Note that the definition of expansions, like the definition of extensions, is not a constructive one, since there is a self-reference in the requirement that $E = \Omega_T(E) = \{\varphi \mid T \cup LE \cup \neg LE^c \vDash \varphi\}$, i.e. we have to *guess* a set E and subsequently check wether it fulfills the requirements. We will give some examples is the following section.

2.2.1 Properties of autoepistemic logic

The *degree* of an ae-formula φ (written $deg(\varphi)$) is the maximal depth of L-nesting occurring in φ:

- if $\varphi \in \mathcal{L}$ then $deg(\varphi) = 0$,
- $deg(\neg\varphi) = deg(\varphi)$,
- $deg(\varphi \wedge \psi) = \max(\varphi, \psi)$,
- $deg(\varphi \vee \psi) = \max(\varphi, \psi)$,
- $deg(L\varphi) = deg(\varphi) + 1$.

For an ae-theory T we denote the set of formulas in T with degree less than or equal to n by T_n. The nonmodal or propositional part of T, i.e. T_0, is called the *kernel* of T. The set $\mathcal{L}_{ae_0}$ is simply the set of propositional, or non-modal, formulas $\mathcal{L}$.

The following notion, that of *stable sets*, is very important for autoepistemic logic, especially because of theorems 2.12 and 2.14.

Definition 2.11. *An ae-theory E is called* stable *if and only if*

- *E is deductively closed, i.e. $Th(E) = E$,*
- *if $\varphi \in E$ then also $L\varphi \in E$,*
- *if $\varphi \notin E$ then also $\neg L\varphi \in E$.*

Theorem 2.12. *Let T and E be ae-theories. Then the following are equivalent:*

(a) *E is an expansion of T*

(b) *E is stable, contains T and satisfies $E \subseteq \Omega_T(E)$.*

Proof: by definition 2.10, (a) implies (b). Conversely, we only have to show that $\Omega_T(E) \subseteq E$. By stability of E, we have that $Th(E) = E$ and that $LE \cup \neg LE^c \subseteq E$. Since by assumption E also contains T, we get that $\Omega_T(E) = Th(T \cup LE \cup \neg LE^c) \subseteq E$. □

Theorem 2.14 states that two stable sets are the same if their kernels are the same. In particular this also holds for expansions, as we now know by the previous theorem that expansions are stable sets. The theorem uses the following lemma.

Lemma 2.13. *Every ae-formula φ is equivalent to a formula φ' (with the same degree as φ) of the form $\varphi_0 \wedge \ldots \wedge \varphi_n$, where each φ_i is of the form*

$$\varphi_i = \varphi_{i,0} \vee L\varphi_{i,1} \vee \ldots \vee L\varphi_{i,p_i} \vee \neg L\varphi_{i,p_i+1} \vee \ldots \vee \neg L\varphi_{i,q_i}, \tag{1}$$

where $\varphi_{i,0} \in \mathcal{L}$ and $deg(\varphi_{i,j}) < deg(\varphi_i)$. Formulas of the form of φ' are said to be in normal form*.*

Proof: The proof is very well-know and will be omitted. For a reference, see [1]. □

Theorem 2.14. *Let E and F be stable sets. Then $E_0 = F_0$ implies $E = F$.*

Proof: By induction on the degree of L-nesting, we will prove that $\varphi \in E$ iff $\varphi \in F$. For formulas with degree 0 this is given by the assumption $E_0 = F_0$. Now assume that the claim holds for formulas with degree less or equal to d and let the degree of φ be $d + 1$. As we explained before, φ is equivalent to $\varphi_0 \wedge \ldots \wedge \varphi_n$, where

$$\varphi_i = \psi_{i,0} \vee L\psi_{i,1} \vee \ldots \vee L\psi_{i,p_i} \vee \neg L\psi_{i,p_i+1} \vee \ldots \vee \neg L\psi_{i,q_i},$$

where $\psi_{i,0}$ is a non-modal formula and $deg(\psi_{i,j}) < d + 1$ for $j = 1, \ldots j = q_i$. Since E and F are deductively closed, it suffices to prove $\varphi_i \in E$ iff $\varphi_i \in F$ for an arbitrary i.

Case 1: Suppose $\psi_{i,j} \in E$ for some $j \in \{1, \ldots, p_i\}$. Then by the induction hypothesis also $\psi_{i,j} \in F$, since $deg(\psi_{i,j}) \leq d$. By stability of E and F it follows that $L\psi_{i,j} \in E, F$ and thus $\varphi_i \in E, F$.

Case 2: Suppose $\psi_{i,k} \notin E$ for some $k \in \{p_i + 1, \ldots q_i\}$. Then again by the induction hypothesis $\psi_{i,k} \notin F$. By stability then $\neg L\psi_{i,k} \in E, F$ and thus $\varphi_i \in E, F$.

Case 3: Suppose case 1 and 2 do not hold, i.e. that $\psi_{i,j} \notin E$ for all $j \in \{1, \ldots, p_i\}$ and $\psi_{i,k} \in E$ for all $k \in \{p_i + 1, \ldots, q_i\}$. By induction hypothesis and stability $\neg L\psi_{i,1}, \ldots \neg L\psi_{i,p_i}, L\psi_{i,p_i+1}, \ldots L\psi_{i,q_i} \in E, F$. By deductive closure and the assumption $E_0 = F_0$ it follows that $\varphi_i \in E$ iff $\psi_{i,0} \in E$ iff $\psi_{i,0} \in F$ iff $\varphi_i \in F$. □

As we explained, the definition of expansions of an ae-theory T are non-constructive. The following theorem gives a somewhat more constructive way of finding expansions, similar to theorem 2.4 for finding extensions. Also, it proves that every deductively closed first order theory is the kernel of an expansion. In the theorem we write $T \models_E \varphi$ for $T \cup LE \cup \neg LE^c \models \varphi$. In [1] the entailment relation $T \models_E \varphi$ is called "deduction of φ from T with belief set E". Note that E is an expansion of T if and only if $E = \{\varphi \mid T \models_E \varphi\}$.

Theorem 2.15. *Let $T \subseteq \mathcal{L}$ be a deductively closed set of non-modal formulas. Then $E = \bigcup_n E(n)$ is a stable set with kernel T, where $E(n)$ is inductively defined as follows:*

$$\begin{aligned} &E(0) = T, \\ &E(n+1) = \{\varphi \in \mathcal{L}_{ae_{n+1}} \mid T \models_{E(n)} \varphi\}. \end{aligned}$$

Consequently, E is an expansion of T.

Proof: Firstly we show by induction on n that $E(n) = E(n+1)_n$. As the base case,

$$\begin{aligned} E(1)_0 &= \{\varphi \in \mathcal{L}_{ae_1} \mid T \models_{E(0)} \varphi\}_0 \\ &= \{\varphi \in \mathcal{L}_{ae_0} \mid T \models_{E(0)} \varphi\} \\ &= \{\varphi \in \mathcal{L}_{ae_0} \mid T \models \varphi\} \\ &= T = E(0). \end{aligned}$$

(The second and third step are validated by the fact that T is a deductively closed set of first order formulas.) Next, assume as an induction hypothesis that $E(n) = E(n+1)_n$. Then

$$\begin{aligned} E(n+2)_{n+1} &= \{\varphi \in \mathcal{L}_{ae_{n+2}} \mid T \models_{E(n+1)} \varphi\}_{n+1} \\ &= \{\varphi \in \mathcal{L}_{ae_{n+1}} \mid T \models_{E(n+1)} \varphi\} \\ &= \{\varphi \in \mathcal{L}_{ae_{n+1}} \mid T \models_{E(n+1)_n} \varphi\} \\ &= \{\varphi \in \mathcal{L}_{ae_{n+1}} \mid T \models_{E(n)} \varphi\} \\ &= E(n+1). \end{aligned}$$

The third step is validated by the fact that $\{\varphi \mid T \cup LE(n+1) \cup \neg LE(n+1)^c \models \varphi\}$ restricted to $\mathcal{L}_{(n+1)}$ is the same as $\{\varphi \mid T \cup LE(n) \cup \neg LE(n)^c \models \varphi\}$. In particular, we have that $E(n) \subseteq E(n+1)$. It is obvious that E is deductively closed and that $T = E_0$, so we only have to prove that E is stable. Let $\varphi \in E$. Then

for some n, $\varphi \in E(n)$. Then also $L\varphi \in E(n+1)$ and so $L\varphi \in E$. Also, if $\varphi \notin E$ and $\varphi \in \mathcal{L}_{ae_{n+1}}$ then $\varphi \notin E(n)$, hence $T \vDash_{E(n)} \neg L\varphi$ and so $\neg L\varphi \in E$. This proves stability of E. □

2.3 Link between default and autoepistemic logic

Given the two nonmonotonic formalisms, one wonders about the link between the two. As we saw, problems can be attacked by both formalisms with similar outcomes, for example the Tweety example. Konolige proposed in [12] to translate a default

$$\frac{\varphi : \psi_1, \ldots, \psi_n}{\chi} \tag{2}$$

into the ae-formula

$$L\varphi \wedge \neg L\neg\psi_1 \wedge \ldots \wedge \neg L\neg\psi_n \rightarrow \chi. \tag{3}$$

Certainly this seems legitimate, if one recalls that the default $\frac{\varphi:\psi_1,\ldots,\psi_n}{\chi}$ would be explained as "if we know that φ and it is consistent to assume that $\psi_1, \ldots \psi_n$, then conclude χ" which can be interpreted as "if we know that φ and we *don't* know that *not* $\psi_1, \ldots \psi_n$, then conclude χ". The latter is easily interpreted in autoepistemic logic as $L\varphi \wedge \neg L\neg\psi_1 \wedge \ldots \wedge \neg L\neg\psi_n \rightarrow \chi$. We call this translation from default logic to autoepistemic logic the *Konolige translation*; for a default δ of the form (2), let its Konolige translation $kon(\delta)$ be the ae-formula (3). For a set of defaults D, let $kon(D) = \{kon(\delta) \mid \delta \in D\}$. For a default theory $\Delta = (D, W)$, its Konolige translation $kon(\Delta)$ is then defined as the ae-theory $W \cup kon(D)$.

There is a problem, though. Although the extensions of default theories correspond to expansions of their Konolige translation, the opposite is not guaranteed. In other words, the set of extensions corresponds to a *subset of* the set of expansions under Konolige translation.

Example 2.16. As an example, consider the default theory

$$\Delta = (\emptyset, \{\frac{p : true}{p}\})$$

and the Konolige translation

$$kon(\Delta) = \{Lp \wedge \neg L\neg true \rightarrow p\}.$$

The autoepistemic theory $kon(\Delta)$ has two expansions; with kernels $Th(\emptyset)$ and $Th(\{p\})$ respectively. Namely, if we decide to believe in nothing, this turns out to be a stable set containing $kon(\Delta)$, since $Lp \wedge \neg L\neg true \rightarrow p$ is satisfied vacuously, and if we decide to believe in p, this also turns out to be a stable set containing $kon(D)$, since $Lp \wedge \neg L\neg true \rightarrow p$ is satisfied. Now consider the default theory Δ. The set $Th(\emptyset)$ (which is the kernel of one of the expansions of $kon(\Delta)$) is an extension of Δ, since it obviously satisfies **D1**-**D3** and thus is a fixpoint of the operator Γ_Δ. The set $Th(p)$ however is not an extension of Δ, for in this case the least set $\Gamma_\Delta(Th(p))$ such that

D1 $\emptyset \subseteq \Gamma_\Delta(Th(p))$,

D2 $Th(\Gamma_\Delta(Th(p))) = \Gamma_\Delta(Th(p))$

D3 if $p \in \Gamma_\Delta(Th(p))$ and $Th(p) \nvdash \neg true$ then $p \in \Gamma_\Delta(Th(p))$

is $Th(\emptyset)$; $Th(p)$ evidently is not a fixpoint of Γ_Δ and thus not an extension. □

The "*deciding* to believe in p" that we used in the autoepistemic example is an option default logic doesn't have; the belief in p has to be justified by either the set of facts or by another default. In this case: since the prerequisite p of the only default will never be known (since it is not provided by the set of facts, nor by any other default), the default can never be applied in an extension. This is how Denecker et al. explain that

> "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" [8].

Many have tried since, but eventually

> "Gottlob [11] 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*" [8].

However, Konolige's translation has turned out to be very useful after all. As Konolige himself proposed, default logic can be embedded into a version of autoepistemic logic, in which extensions relate to *strongly grounded expansions* under Konolige translation [12]. To be more precise, he proved that if Δ is a default theory and $T = kon(\Delta)$ its Konolige translation, then the extensions of Δ are in one-to-one correspondence to the strongly grounded expansions of T. Moreover, Marek and Truszczyński [13] proposed the concept of *weak extensions*, an alternative semantics for default logic, which correspond to expansions under Konolige translation. That is, the weak extensions of a default theory Δ are in one-to-one correspondence with the expansions of the ae-theory $T = kon(\Delta)$. However, the link between default logic and autoepistemic logic does not stop here.

An underlying, unifying structure of default logic and autoepistemic logic has been extensively studied by Denecker, Marek and Truszczyński. It was found with help of *possible world structures*; collections of two-valued interpretations each representing a state of the world that is possible according to the agents beliefs. Moore [16] already used this alternative framework for autoepistemic logic and provided for an alternative characterization, given in section 3.2. These works of Denecker, Marek and Truszczyński are explained in the next section.

3 Possible world semantics

In this chapter we will explain how Denecker, Marek and Truszczyński united default logic and autoepistemic logic with use of possible world structures. Next to Reiter's extensions they described three other semantics for default logic: weak extensions, partial extensions and partial expansions. Also, next to Moore's expansions they described three other semantics for autoepistemic logic: extensions[2], partial extensions and partial expansions. Their main contribution is proving that the Konolige translation connects precisely the extensions, weak extensions, partial extensions and partial expansions of default logic with the extensions, expansions, partial extensions and partial expansions of autoepistemic logic, respectively.

3.1 An example of possible world semantics

We will consider the autoepistemic language $\mathcal{L}_{ae}$ (the propositional language $\mathcal{L}$ with one modal operator L; in examples we will use appropriate indices in stead of the index ae). A *two-valued interpretation* or simply *interpretation* is a function assigning to each propositional letter in the given propositional language $\mathcal{L}$ one of the values $true$ and $false$. The set of all interpretations of $\mathcal{L}$ will be denoted by $\mathcal{A}$ (with appropriate index). A *possible world structure* is a set of interpretations (it can be seen as a universal Kripke model with total accessibility relation with the possibility of an empty set of interpretations). The possible world structures represent an agent's knowledge about the actual world. The set of all possible world structures $2^{\mathcal{A}}$ will be denoted by $\mathcal{W}$ (with appropriate index).

To get an idea of how possible world structures work, we will look at a simplified version of the Tweety example from the introduction.

Example 3.1. Since the facts that Tweety is a bird and a penguin are given in the example, we consider the autoepistemic language $\mathcal{L}_f$ with only f (for "Tweety flies") as an atom. Then there are just two possible interpretations, namely:

I_f **:** $f \mapsto true$

$I_{\neg f}$ **:** $f \mapsto false$

We will denote the set of interpretations $\{I_f, I_{\neg f}\}$ by $\mathcal{A}_f$. The four subsets of $\mathcal{A}_f$ are the possible world structures in this example: $\emptyset$, $\{I_f\}$, $\{I_{\neg f}\}$ and $\mathcal{A}_f$ itself. They represent an agent's knowledge about the world. For instance, in the possible world structure $\{I_f\}$ the agent knows that Tweety flies, whereas in the possible world structure $\mathcal{A}_f$ the agent holds it possible that Tweety flies (I_f) or not ($I_{\neg f}$); all options are left open. In the possible world structure $\emptyset$ there are no options at all; every assertion about the ability to fly is considered

[2]The extensions for autoepistemic logic are what Konolige [12] called *strongly grounded expansions*.

true and thus the agent has an inconsistent theory about the world. □

In general, the more interpretations we include in a possible world structure, the more the agent holds possible, so the less he actually knows for sure. In particular in the structure $\mathcal{A}$ itself (or $\mathcal{A}_f$ in the example above) everything is possible and nothing is certain. Therefore, the agents *knowledge* about the actual world decreases when including more interpretations. That justifies the following order: given a collection of possible world structures $\mathcal{W}$, it can be ordered by reverse set inclusion, also called the *knowledge ordering*. For all $Q, Q' \in \mathcal{W}$, let

$$Q \sqsubseteq Q' \text{ if and only if } Q' \subseteq Q.$$

The structure $\langle \mathcal{W}, \sqsubseteq \rangle$ is a complete lattice: if $Q, Q' \in \mathcal{W}$ then $Q \cup Q'$ is obviously the least upper bound of Q and Q' with respect to the subset ordering $\subseteq$, so $Q \cup Q'$ is the greatest lower bound of Q and Q' with respect to the knowledge ordering $\sqsubseteq$. Similarly, $Q \cap Q'$ is the $\subseteq$-greatest lower bound and the $\sqsubseteq$-least upper bound of Q and Q'.

We will first give a two-valued truth function $\mathcal{H}_{Q,I} : \mathcal{L}_{ae} \to \{true, false\}$, given a possible world structure $Q \subseteq \mathcal{W}$ and an interpretation $I \in \mathcal{A}$ by induction on formulas. This truth function will give the truth values of ae-formulas relative to Q and I. The possible world structure Q represents the set of possible states that the agent believes the actual world to be in as we explained above; the interpretation I represents the actual world.

Definition 3.2. *The truth function $\mathcal{H}_{Q,I}$ will give truth values to formulas relative to the knowledge state Q of the agent and the actual world I as follows:*

1. $\mathcal{H}_{Q,I}(p) = I(p)$, *if p is a propositional letter,*
2. $\mathcal{H}_{Q,I}(\varphi \wedge \psi) = true$ *if and only if* $\mathcal{H}_{Q,I}(\varphi) = \mathcal{H}_{Q,I}(\psi) = true$,
3. $\mathcal{H}_{Q,I}(\varphi \vee \psi) = true$ *if and only if* $\mathcal{H}_{Q,I}(\varphi) = true$ *or* $\mathcal{H}_{Q,I}(\psi) = true$,
4. $\mathcal{H}_{Q,I}(\neg\varphi) = true$ *if and only if* $\mathcal{H}_{Q,I}(\varphi) = false$,
5. $\mathcal{H}_{Q,I}(L\varphi) = true$ *if and only if for every* $J \in Q, \mathcal{H}_{Q,J}(\varphi) = true$.

Note that the value of $L\varphi$ entirely depends on Q (knowing something will indeed only depend on the agent's knowledge state, not on the actual world); we will illustrate this by writing $\mathcal{H}_Q(L\varphi)$ in stead of $\mathcal{H}_{Q,I}(L\varphi)$. The *theory of a possible world structure* Q is defined as

$$Th_L(Q) = \{\varphi \mid \mathcal{H}_Q(L\varphi) = true\}.$$

This means that $Th_L(Q)$ contains the formulas that the agent believes to be true. For instance in example 3.1 above the theory of the possible world structure $\{I_f\}$ contains f. Namely,

$$\begin{aligned} f \in Th_L(\{I_f\}) \quad &\text{iff} \quad \mathcal{H}_{\{I_f\}}(Lf) = true \\ &\text{iff} \quad \mathcal{H}_{\{I_f\},J}(f) = true \text{ for every } J \in \{I_f\} \\ &\text{iff} \quad I_f(f) = true \text{ (which is certainly the case.)} \end{aligned}$$

But then also $LLf \in Th(\{I_f\})$ namely

$$\begin{array}{lll} Lf \in Th_L(\{I_f\}) & \text{iff} & \mathcal{H}_{\{I_f\}}(LLf) = true \\ & \text{iff} & \mathcal{H}_{\{I_f\},J}(Lf) = true \text{ for every } J \in \{I_f\} \\ & \text{iff} & \mathcal{H}_{\{I_f\}}(Lf) = true \\ & & \quad \text{(since the value of } Lf \text{ only depends on } \{I_f\}) \\ & \text{iff} & \mathcal{H}_{\{I_f\},J}(f) = true \text{ for every } J \in \{I_f\} \\ & \text{iff} & I_f(f) = true. \end{array}$$

Notice that this is an instance of positive introspection. We will indeed see that $\mathcal{H}$ plays an important role in the definition of expansions (and other semantics) in terms of possible world structures.

3.2 Possible world semantics for autoepistemic logic

In this section we will firstly present the operator D_T on the lattice $\langle \mathcal{W}, \sqsubseteq \rangle$ and prove that its fixpoints correspond exactly to the expansions of an autoepistemic theory T. Then we will present the notion of approximations of possible world structures, which are certain pairs of possible world structures (P, S). Next, we will present partial expansions, which are the fixpoints of the operator $\mathcal{D}_T$ on pairs of possible world structures. This operator can be seen as a generalization of D_T to pairs of possible world structures. Then, from $\mathcal{D}_T$, we will derive another operator on $\langle \mathcal{W}, \sqsubseteq \rangle$ denoted by D_T^{st}. The fixpoints of D_T^{st} will later be proven to correspond precisely to the extensions of default logic. That is, if Δ is a default theory and $T = kon(\Delta)$ its Konolige translation, then the fixpoints of D_T^{st} correspond precisely to the extensions of Δ. Hence we will call these fixpoints the extensions of T[3].

All those operators represent an agent's *revising* his belief set. This means that given a possible world structure Q that represents his belief set, the agent can *revise* his belief set with help of such an operator, such that the revision is a "better" belief set ("better" in ways we will explain).

3.2.1 Characterization of expansions

Firstly, to clear up any confusion that might occur, we should again explicitly distinguish the two operators with similar notation: $Th()$ and $Th_L()$. The first is simply logical consequence and will be taken over sets of formulas, the second is defined over possible world structures as follows: $Th_L(Q) = \{\varphi \mid \mathcal{H}_Q(L\varphi) = true\}$.

Let T be an ae-theory and Q a possible world structure. Given the brief preview above, we define the operator $D_T : \mathcal{W} \to \mathcal{W}$ as follows:

$$D_T(Q) = \{I \mid \mathcal{H}_{Q,I}(\varphi) = true \text{ for all } \varphi \in T\}.$$

[3]As we mentioned before, note that the extensions of T are sometimes also called the *strongly grounded expansions* of T

Moore [16] described the following characterization of expansions, which we provide with a much more detailed proof.

Theorem 3.3. *Let $T \subseteq \mathcal{L}_{ae}$. A theory $E \subseteq \mathcal{L}_{ae}$ is an expansion of T if and only if $E = Th_L(Q)$ for some possible world structure Q such that $Q = D_T(Q)$.*

Proof:

"⇐" Let Q be a possible world structure such that $Q = D_T(Q)$ and let $E = Th_L(Q) = \{\varphi \mid \mathcal{H}_Q(L\varphi) = true\}$. Then we need to show that E is an expansion of T, i.e. that $E = Th(T \cup LE \cup \neg LE^c)$.

1. "$T \subseteq E$": let $\varphi \in T$. Then for all $I \in D_T(Q), \mathcal{H}_{Q,I}(\varphi) = true$. Since $Q = D_T(Q)$, it follows that for all $I \in Q, \mathcal{H}_{Q,I}(\varphi) = true$, hence $\mathcal{H}_Q(L\varphi) = true$, so $\varphi \in Th_L(Q) = E$. This proves that $T \subseteq E$.
2. "$LE \subseteq E$": let $\varphi \in E$, i.e. $\mathcal{H}_Q(L\varphi) = true$. Since $\mathcal{H}_Q(L\varphi)$ is an abbreviation of $\mathcal{H}_{Q,I}(L\varphi)$ (because $\mathcal{H}_{Q,I}(L\varphi)$ only depends on Q, not on I), this means that $\mathcal{H}_{Q,I}(L\varphi) = true$ for all I. But then in particular $\mathcal{H}_{Q,I}(L\varphi) = true$ for all $I \in Q$, so $\mathcal{H}_Q(LL\varphi) = true$, hence $L\varphi \in E$.
3. "$\neg LE^c \subseteq E$": let $\varphi \notin E$, i.e. $\mathcal{H}_Q(L\varphi) = false$. This means, again because the truth value $\mathcal{H}_{Q,I}(L\psi)$ only depends on Q, not on I, that $\mathcal{H}_{Q,I}(L\varphi) = false$ for all I, and thus that $\mathcal{H}_{Q,I}(\neg L\varphi) = true$ for all I. In particular this is so for all $I \in Q$, hence we can conclude $\mathcal{H}_Q(L\neg L\varphi) = true$, and thus that $\neg L\varphi \in E$.
4. "$Th(E) \subseteq E$": let $\psi_0 \wedge \ldots \wedge \psi_n \vdash \varphi$ and let $\psi_0, \ldots \psi_n \in E$, i.e. $\mathcal{H}_Q(L\psi_i) = true$ for each $i \in \{0, \ldots, n\}$. Then we need to show that $\varphi \in E$. By an easy induction on n it follows that $\mathcal{H}_{Q,I}(L(\psi_0 \wedge \ldots \wedge \psi_n)) = true$ for all $I \in Q$. (Namely, because $\mathcal{H}_Q(L\psi_i) = true$ for each i, we know that for each $I \in Q$, $\mathcal{H}_{Q,I}(\psi_i)$ for each i, and thus for each $I \in Q$, $\mathcal{H}_{Q,I}(\psi_0 \wedge \ldots \wedge \psi_n)$, and thus that that $\mathcal{H}_Q(L(\psi_0 \wedge \ldots \wedge \psi_n))$.) Also, $\psi_0 \wedge \ldots \wedge \psi_n \rightarrow \varphi$ is a tautology and since all interpretations make tautologies true it follows that $\mathcal{H}_{Q,I}(\psi_0 \wedge \ldots \wedge \psi_n \rightarrow \varphi) = true$ for all $I \in Q$. It is easy to see that $\mathcal{H}_{Q,I}(\varphi) = true$ for all $I \in Q$, and thus that $\varphi \in E$. (Namely, $\mathcal{H}_{Q,I}((\psi_0 \wedge \ldots \wedge \psi_n) \rightarrow \varphi) = true$ iff $\mathcal{H}_{Q,I}(\varphi) = true$ whenever $\mathcal{H}_{Q,I}(\psi_0 \wedge \ldots \wedge \psi_n) = true$.)
5. "$E \subseteq Th(T \cup LE \cup LE^c)$: suppose the opposite, i.e. let $\varphi \in E$ and $\varphi \notin Th(T \cup LE \cup LE^c)$. Note that T cannot be inconsistent, for then φ would be included. Then by deductive closure $\neg\varphi \in Th(T \cup LE \cup LE^c)$and hence, by steps 1 to 4, $\neg\varphi \in E$. This means that by 4 $E = For$. Now, since Q is such that $E = Th_L(Q)$ and E is inconsistent, it follows that $Q = \emptyset$, since for no interpretation I, $\mathcal{H}_{Q,I}(\bot) = true$. But then by the consistency of T, $D_T(Q)$ cannot be empty, hence $Q \neq D_T(Q)$.

"⇒" : Let E be an expansion of T, i.e. $E = Th(T \cup LE \cup \neg LE^c)$. Then we need to show that there is a possible world structure Q such that $Q = D_T(Q)$ and $E = Th_L(Q)$. Stable sets, and thus expansions in particular, are completely determined by their kernel. Moreover, any expansion E of T is also an expansion of its kernel E_0, i.e. $E = Th(E_0 \cup LE \cup \neg LE^c)$. Now, if φ is a propositional formula (without occurrences of the modal operator L) then the truth value $\mathcal{H}_{Q,I}(\varphi)$ depends entirely on the interpretation I and not on Q, so that the following possible world structure is well defined:

$$\begin{aligned} Q &:= \{I \mid I(\varphi) = true \text{ for all } \varphi \in E_0\} \\ &= \{I \mid \mathcal{H}_{Q,I}(\varphi) = true \text{ for all } \varphi \in E_0\} \\ &= D_{E_0}(Q) \end{aligned}$$

We will show that $E = Th_L(Q)$ and that $Q = D_T(Q)$:

- "$E \subseteq Th_L(Q)$": We can construct the expansion E as follows: $E = \bigcup_n E(n)$, where $E(0) := E_0$ and $E(n+1) := \{\varphi \in \mathcal{L}_{ae_{n+1}} \mid E(0) \models_{E(n)} \varphi\}$(Theorem 9.5 in [1]). Now we can prove by induction on n that $E(n) \subseteq Th_L(Q)$, thus concluding that $E \subseteq Th_L(Q)$. For $n = 0$, it is obvious that $E(0) \subseteq Th_L(Q)$, since by our definition of Q, for all $\varphi \in E_0$ and all $I \in Q$, $\mathcal{H}_{Q,I}(\varphi) = true$ and so $E_0 \subseteq Th_L(Q)$. Now suppose that $E(n) \subseteq Th_L(Q)$ and that $E_0 \models_{E(n)} \varphi$ with $\varphi \in \mathcal{L}_{ae_{n+1}}$. The only relevant case is if $\varphi = L\psi$, with $\psi \in E(n)$ (since all sets $E(k)$ we are talking about are deductively closed and $\mathcal{H}_{Q,I}$ respects logical consequence). But then by assumption $\psi \in Th_L(Q)$, hence for all $I \in Q$, $\mathcal{H}_{Q,I}(\psi) = true$, so also $\mathcal{H}_Q(L\psi) = true$ and thus $L\psi \in Th_L(Q)$.
- "$E \supseteq Th_L(Q)$": we will show the equivalent statement $E^c \subseteq (Th_L(Q))^c$. Suppose $\varphi \notin E$. Then, since $\neg LE^c \in E$, $\neg L\varphi \in E$. Since we just proved that $E \subseteq Th_L(Q)$, it follows that $\mathcal{H}_Q(L\neg L\varphi) = true$, hence $\mathcal{H}_{Q,I}(\neg L\varphi) = true$. It follows that there is some $I \in Q$ such that $\mathcal{H}_{Q,I}(\varphi) = false$, and thus that $\varphi \notin Th_L(Q)$.
- "$Q \subseteq D_T(Q)$": this follows from the above, in the following way. Since $E = Th_L(Q)$, it follows that $\mathcal{H}_{Q,I}(\varphi) = true$ for all $I \in Q$ if and only if $\varphi \in E$. Now, since $T \subseteq E$, it follows that for all $I \in Q, \mathcal{H}_{Q,I}(\varphi) = true$ for all $\varphi \in T$, and so that each $I \in Q$ is also in $D_T(Q)$.
- "$Q \supseteq D_T(Q)$": let I be such that $\mathcal{H}_{Q,I}(T) = true$. Then we need to show that $I \in Q$, i.e. that $\mathcal{H}_{Q,I}(E_0) = true$ as well. Since we showed that $E = Th_L(Q)$, i.e. $\varphi \in E$ iff $\mathcal{H}_Q(L\varphi) = true$, we know that both $\mathcal{H}_{Q,I}(LE) = true$ and that $\mathcal{H}_{Q,I}(\neg LE^c) = true$. Now, since E is the set of logical consequences of $T \cup LE \cup \neg LE^c$ and the truth function $\mathcal{H}_{Q,I}$ respects logical consequence, $\mathcal{H}_{Q,I}(E) = true$, so also $\mathcal{H}_{Q,I}(E_0) = true$. □

As theorem 3.3 is stated, it gives no definite answer on whether the possible world structure given in the proof, namely $Q = D_{E_0}(Q)$ is the only structure that satisfies the theorem. The following proposition shows that it is in fact unique.

Proposition 3.4. *Let Q, Q' be two different fixpoints of D_T, i.e. $Q \neq Q'$, $D_T(Q) = Q$ and $D_T(Q') = Q'$. Then $Th_L(Q) \neq Th_L(Q')$. Consequently, if E is an expansion of T then the possible world structure P such that $P = D_T(P)$ and $E = Th_L(P)$ given by theorem 3.3 is unique.*

Proof: Suppose for a contradiction that $Th_L(Q) = Th_L(Q')$, i.e.

$$\{\varphi \mid \mathcal{H}_Q(L\varphi) = true\} = \{\varphi \mid \mathcal{H}_{Q'}(L\varphi) = true\}. \tag{4}$$

Then an easy induction will show that precisely the same interpretations I deliver the same truth values in $\mathcal{H}_{Q,I}(\varphi)$ and $\mathcal{H}_{Q',I}(\varphi)$, namely:

- if p is a propositional letter then $\mathcal{H}_{Q,I}(p) = I(p) = \mathcal{H}_{Q',I}(p)$,

When we assume as an induction hypothesis that $\mathcal{H}_{Q,I}(\varphi_i) = \mathcal{H}_{Q',I}(\varphi_i)$ for $i = 1, 2$ then it's obvious by definition 3.2 that

- $\mathcal{H}_{Q,I}(\varphi_1 \wedge \varphi_2) = \mathcal{H}_{Q',I}(\varphi_1 \wedge \varphi_2)$,
- $\mathcal{H}_{Q,I}(\varphi_1 \vee \varphi_2) = \mathcal{H}_{Q',I}(\varphi_1 \vee \varphi_2)$,
- $\mathcal{H}_{Q,I}(\neg\varphi_1) = \mathcal{H}_{Q',I}(\neg\varphi_1)$,
- $\mathcal{H}_Q(L\varphi) = \mathcal{H}_{Q'}(L\varphi)$ (by 4).

It follows directly that $\{I \mid \mathcal{H}_{Q,I}(\varphi) = true\} = \{I \mid \mathcal{H}_{Q',I}(\varphi) = true\}$ and so that $D_T(Q) = D_T(Q')$. This is a contradiction with the facts that $Q \neq Q'$ and $Q = D_T(Q)$ and $D_T(Q')$, thus the possible world structure talked about in theorem 3.3 is unique. □

To illustrate the above, consider the following example.

Example 3.5. Consider the ae-language $\mathcal{L}_p$ with the only atom p and let $T = \{\neg Lp\}$, i.e., the only thing the agent knows is that he doesn't know p. The relevant interpretations are of course $I_p : p \mapsto true$ and $I_{\neg p} : p \mapsto false$. The four possible world structures are then $\mathcal{A}_p = \{I_p, I_{\neg p}\}, \mathcal{I}_p = \{I_p\}, \mathcal{I}_{\neg p} = \{I_{\neg p}\}$ and $\emptyset$. Let us look at what the operator D_T does with a possible world structure Q:

$$\begin{aligned} D_T(Q) &= \{I \mid \mathcal{H}_{Q,I}(\neg Lp) = true \text{ for all } I \in Q\} \\ &= \{I \mid \mathcal{H}_{Q,I}(Lp) = false \text{ for all } I \in Q\} \end{aligned}$$

Now, since the falsity of Lp does not depend on I, but solely on Q, we get that

$$\begin{aligned} D_T(Q) &= \mathcal{A}_p \text{ if } \mathcal{H}_Q(Lp) = false, \text{ and} \\ D_T(Q) &= \emptyset \text{ if } \mathcal{H}_Q(Lp) = true. \end{aligned}$$

Thus, we get the following table, in which we can read that the expansions of T are $\mathcal{A}_p$ and $\emptyset$:

Q	$\mathcal{A}_p$	$\mathcal{I}_{\neg p}$	$\mathcal{I}_p$	$\emptyset$
$D_T(Q)$	$\mathcal{A}_p$	$\mathcal{A}_p$	$\emptyset$	$\emptyset$

Indeed, to come to a stable set with the only knowledge that you don't know p, you can either decide to believe in nothing ($\mathcal{A}_p$) or decide to believe in everything ($\emptyset$), which is of course inconsistent. □

3.2.2 Approximating possible world structures

Now that we got interested in fixpoints of the operator D_T, it's time to look for ways of finding them. Denecker et al. [8] used the concept of approximations of possible world structures. If someone would have to give an estimate for which formulas are included in some given ae-theory T, a *conservative* guess would be an *underestimate*, i.e. one would only include formulas of which he was certain, while a *liberal* guess would be an *overestimate*. An *approximation of* T is a tuple of ae-theories (V, W) such that V is a conservative estimate (an underestimate) of T and W a liberal one (overestimate), i.e. $V \subseteq T \subseteq W$. Similarly, if Q is a possible world structure, an *approximation of* Q is a tuple of possible world structures (P, S) such that $P \sqsubseteq Q \sqsubseteq S$. We use the knowledge ordering $\sqsubseteq$ for the reasons explained in section 3.1: recall that $Q \sqsubseteq Q'$ if and only if $Q' \subseteq Q$ and that the larger a set of interpretations is, the more the agent holds possible and the less he knows for certain. This means that when $P \sqsubseteq Q \sqsubseteq S$, the knowledge represented by S is more than that of Q and thus is an overestimate. Similarly, P represents an underestimate of the knowledge represented by Q. A pair of possible world structures (P, S) is called a *belief pair*; if also $P \sqsubseteq S$ then we call (P, S) a *consistent* belief pair; a pair (P, P) is called a *complete* belief pair. When a term refers to a belief pair the indices $(\ )_1$ and $(\ )_2$ refer to the first and second component of the pair respectively; for example if $T = (P, S)$ then $T_1 = P$ and $T_2 = S$. The set of belief pairs is denoted by $\mathcal{B}$ (with appropriate index). In particular a consistent belief pair (P, S) approximates all Q such that $P \sqsubseteq Q \sqsubseteq S$. Some argue that *only* consistent pairs are in order in this context, since inconsistent pairs do not approximate anything, and its hard to imagine what an inconsistent belief pair would mean for an agent. For an account of approximation theory restricted to consistent belief pairs, see [9]. We order the set of belief pairs with the so-called *precision order* $\preceq_{pr}$ as follows:

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

Another order on $\mathcal{B}$ is the *knowledge ordering* $\sqsubseteq$; it is the component-wise extension of the knowledge ordering $\sqsubseteq$ on $\mathcal{W}$ (and consequently has got the same notation and terminology):

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

The knowledge ordering plays only a very small role in this thesis.

To illustrate the precision ordering, suppose that (P, S) and (P', S') approximate some Q, i.e. both $P \sqsubseteq Q \sqsubseteq S$ and $P' \sqsubseteq Q \sqsubseteq S'$. If P' is a better (more accurate) $\sqsubseteq$-underestimate to Q than P is (so $P \sqsubseteq P' \sqsubseteq Q$), and S' is a better $\sqsubseteq$-overestimate to Q than S (so $Q \sqsubseteq S' \sqsubseteq S$), we say that (P', S') is a *more precise* approximation of Q than (P, S). That is

$$(P, S) \preceq_{pr} (P', S') \text{ iff } (P', S')P \sqsubseteq P' \sqsubseteq Q \sqsubseteq S' \sqsubseteq S,$$

(so note that $\sqsubseteq$-*larger* pairs are the *more precise* ones).

The following lemma will often be used further on.

Lemma 3.6. *$(P, S) \preceq_{pr} (P', S')$ if and only if $(S', P') \preceq_{pr} (S, P)$.*

Proof: $(P, S) \preceq_{pr} (P', S')$ if and only if $P' \sqsubseteq P$ and $S \sqsubseteq S'$ if and only if $(S', P') \preceq_{pr} (S, P)$ □

Lemma 3.7. *$\langle \mathcal{B}, \preceq_{pr} \rangle$ is a complete lattice*

Proof: for a set of belief pairs $B = \{(P_i, S_i) \in \mathcal{B} \mid i \in I\}$ the following hold

- $(\bigsqcup_{i \in I} P_i, \bigsqcap_{i \in I} S_i)$ is the $\preceq_{pr}$-greatest lower bound, since $\bigsqcup_{i \in I} P_i \sqsubseteq P_i$ for all i and is the $\sqsubseteq$-least to do so and $S_i \sqsubseteq \bigsqcap_{i \in I} S_i$ for all i and is the $\sqsubseteq$-greatest to do so. Consequently, $(\bigsqcup_{i \in I} P_i, \bigsqcap_{i \in I} S_i)$ is the $\preceq_{pr}$-greatest lower bound of B,
- analogously, $(\bigsqcap_{i \in I} P_i, \bigsqcup_{i \in I} S_i)$ is the $\preceq_{pr}$-least upper bound of B. □

The relevance of the fact that $\langle \mathcal{B}, \sqsubseteq \rangle$ is a complete lattice is in the well-known theorem by Tarksi [21], inspired by Knaster [2], which states that a monotone operator on a complete lattice has a least fixpoint. We will namely define a monotone operator $\mathcal{D}_T$ on the lattice $\mathcal{B}$, whose fixpoints are the partial expansions. The Knaster-Tarski theorem will then assure us that there is indeed at least one partial expansion. Also, it states that the fixpoints of a monotone operator on a complete lattice form a complete lattice itself. We denote the least fixpoint of an operator O by *lfp*(O) and the greatest fixpoint by *gfp*(O). The proof is very well-known and will be omitted.

Theorem 3.8. *(Knaster-Tarksi) Let $\langle L, \leq \rangle$ be a complete lattice, $O : L \to L$ a monotone operator on L, i.e. if $x \leq y$ then $O(x) \leq O(y)$. Then*

(a) *lfp$(O) = \bigwedge\{x \mid O(x) \leq x\}$ and*

(b) *the fixpoints of O form a complete lattice with the ordering $\leq$.* □

For any autoepistemic theory T we will define the operator $\mathcal{D}_T : \mathcal{B} \to \mathcal{B}$. It represents the revising of ones approximation of the real world: if an agent has a conservative estimate P of the world and a liberal estimate S, then $\mathcal{D}_T(P, S)$ will give a new (and more precise) conservative estimate P' and a new (and

more precise) estimate S'. In order to do this we will provide ae-formulas with a conservative truth value and a liberal truth value, given a belief pair $(P, S) \in \mathcal{B}$ which represents an approximation for the agents belief set. Fortunately, we only have to define a two-valued *conservative* truth function $\mathcal{H}^2_{(P,S),I}$ in order to do so; the liberal truth value will be found by switching the roles of the underestimate and overestimate in the conservative truth function. Given a belief pair (P, S) and an interpretation I, the function $\mathcal{H}^2_{(P,S),I} : \mathcal{L}_{ae} \to \{true, false\}$ is defined inductively as follows:

1. $\mathcal{H}^2_{(P,S),I}(p) = I(p)$, if p is a propositional letter,
2. $\mathcal{H}^2_{(P,S),I}(\varphi \wedge \psi) = true$ iff $\mathcal{H}^2_{(P,S),I}(\varphi) = true$ and $\mathcal{H}^2_{(P,S),I}(\psi) = true$,
3. $\mathcal{H}^2_{(P,S),I}(\varphi \vee \psi) = true$ iff $\mathcal{H}^2_{(P,S),I}(\varphi) = true$ or $\mathcal{H}^2_{(P,S),I}(\psi) = true$,
4. $\mathcal{H}^2_{(P,S),I}(\neg\varphi) = \neg\mathcal{H}^2_{(S,P),I}(\varphi)$,
5. $\mathcal{H}^2_{(P,S),I}(L\varphi) = true$ iff $\mathcal{H}^2_{(P,S),J}(\varphi) = true$ for all $J \in P$.

Steps 1, 2 and 3 are straightforward. The key steps are 4 and 5, since these ensure the *conservative nature* of $\mathcal{H}^2_{(P,S),I}$. Consider a formula $\varphi \in \mathcal{L}_{ae}$. Step 5 explains that $L\varphi$ is evaluated only with respect to the conservative belief set P. If we consider the negated formula $\neg L\varphi$ however, step 4 explains that we should switch the roles of the conservative and the liberal belief set; $\neg L\varphi$ is true with respect to a conservative approach if $L\varphi$ is false with respect to a liberal approach. Or in other words, to obtain the conservative estimate for the truth value of a formula φ, modal subformulas that appear positively in φ must be evaluated according to the conservative point of view (i.e. with respect to P) while modal subformulas that appear negatively in φ must be evaluated according to the liberal point of view (i.e. with respect to S).

The following lemma states that the conservative truth value and the liberal truth value with respect to a complete belief pair (P, P) are the same, and that they are equal to the truth value with respect to the only possible world structure that (P, P) approximates, namely P.

Lemma 3.9. *Let $P \in \mathcal{W}, \varphi \in \mathcal{L}_{ae}, I \in \mathcal{A}$. Then $\mathcal{H}^2_{(P,P),I}(\varphi) = \mathcal{H}_{P,I}(\varphi)$.*

Proof: The statement follow directly from the definitions of $\mathcal{H}_{P,I}$ and $\mathcal{H}^2_{P,I}$.□

Let $\leq$ be the ordering of the truth values $true$ and $false$, where $false \leq true$. The following proposition states that the conservative truth value of a formula with respect to a belief pair is as least as true with respect to less precise belief pairs. The opposite holds for the liberal truth value.

Proposition 3.10. *Let $(P, S), (P', S') \in \mathcal{B}$ such that $(P, S) \preceq_{pr} (P', S')$. Let $\varphi \in \mathcal{L}_{ae}, I \in \mathcal{A}$. Then $\mathcal{H}^2_{(P,S),I}(\varphi) \leq \mathcal{H}^2_{(P',S'),I}(\varphi)$.*

Proof: If $\mathcal{H}^2_{(P,S),I}(\varphi) = false$ the inequality holds trivially, so we need to show that if $\mathcal{H}^2_{(P,S),I}(\varphi) = true$ then $\mathcal{H}^2_{(P',S'),I}(\varphi) = true$; we will do so by induction on ae-formulas. If $\varphi \in \mathcal{L}$ then the two values only depend on I and thus are the same. As the induction hypothesis, let the claim hold for φ and ψ. Then the claim holds trivially for $\varphi \wedge \psi$ and $\varphi \vee \psi$. Now let $\mathcal{H}^2_{(P,S),I}(L\varphi) = true$, i.e. $\mathcal{H}^2_{(P,S),J}(\varphi) = true$ for all $J \in P$. By the hypothesis then also $\mathcal{H}^2_{(P',S'),J}(\varphi) = true$ for all $J \in P$. Since $(P,S) \preceq_{pr} (P',S')$, $P \sqsubseteq P'$, i.e. $P' \subseteq P$, so then also $\mathcal{H}^2_{(P',S'),J}(\varphi) = true$ for all $J \in P'$ and so $\mathcal{H}^2_{(P',S'),I}(L\varphi) = true$. To conclude the induction, let $\mathcal{H}^2_{(P,S),I}(\neg\varphi) = true$, i.e. $\mathcal{H}^2_{(S,P),I}(\varphi) = false$. Because of $(P,S) \preceq_{pr} (P',S')$, we also have that $(S',P') \preceq_{pr} (S,P)$, hence by the induction hypothesis $\mathcal{H}^2_{(S',P'),I}(\varphi) \leq \mathcal{H}^2_{(S,P),I}(\varphi) = false$. It follows that $\mathcal{H}^2_{(S',P'),I}(\varphi) = false$ and thus that $\mathcal{H}^2_{(P',S),I}(\neg\varphi) = true$. This concludes the proposition. □

3.2.3 Partial expansions for autoepistemic logic

The most prominent properties of the operator $\mathcal{D}_T$ we are going to study are monotonicity and symmetry. Recall that an operator $O : \mathcal{B} \to \mathcal{B}$ is $\preceq_{pr}$-*monotone* if whenever $(P,S) \preceq_{pr} (P',S')$ then also $O(P,S) \preceq_{pr} O(P',S')$ (for brevity we leave out one pair of brackets in $O((P,S))$). An operator $O : \mathcal{B} \to \mathcal{B}$ is *symmetric* if

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

Proposition 3.11. *Let $O : \mathcal{B} \to \mathcal{B}$ be a $\preceq_{pr}$-monotone and symmetric operator. Then*

(a) *whenever (P,S) is consistent, $O(P,S)$ is also consistent,*

(b) *lfp(O) is consistent,*

(c) *if lfp(O) is complete (i.e. there is some P such that lfp$(O) = (P,P)$) then it is the unique fixpoint of O.*

Proof: If (P,S) is consistent then $P \sqsubseteq S$ and thus $(P,S) \preceq_{pr} (S,P)$. By $\preceq_{pr}$-monotonicity $O(P,S) \preceq_{pr} O(S,P)$. Now, let $O(P,S) = (P',S')$. Then by symmetry $O(S,P) = (S',P')$ and thus $(P',S') \preceq_{pr} (S',P')$. It follows that $P' \sqsubseteq S'$ and that (P',S') is consistent, concluding (a). As for (b), let $(P,S) =$ *lfp*(O) (*lfp*(O) exists by the Knaster-Tarski theorem), so $O(P,S) = (P,S)$. Then also, by symmetry, $O(S,P) = (S,P)$, so (S,P) is also a fixpoint of O. Since (P,S) is the least fixpoint, $(P,S) \preceq_{pr} (S,P)$ and so $P \sqsubseteq S$, which means that (P,S) is consistent. For (c), let *lfp*$(O) = (P,P)$ and let (S,R) be a fixpoint of O. Then again by symmetry (R,S) is a fixpoint of O as well. Since (P,P) is $\preceq_{pr}$-less than both (S,R) and (R,S), it follows that $P \sqsubseteq S$ and $R \sqsubseteq P$ and $P \sqsubseteq R$ and $S \sqsubseteq P$ respectively, thus concluding that $S = R = P$. □

Now, consider the following operator[4]

$$\mathcal{D}_T(P,S) = (\mathcal{D}^1_T(P,S), \mathcal{D}^2_T(P,S))$$

where

$$\mathcal{D}^1_T(P,S) = \{I \mid \mathcal{H}^2_{(S,P),I}(T) = true\}$$

and

$$\mathcal{D}^2_T(P,S) = \{I \mid \mathcal{H}^2_{(P,S),I}(T) = true\},$$

where $\mathcal{H}^2_{(S,P),I}(T) = true$ if and only if $\mathcal{H}^2_{(S,P),I}(\varphi) = true$ for all $\varphi \in T$. The fixpoints of $\mathcal{D}_T$ are called *partial expansions*. We will show that $\mathcal{D}_T$ is a symmetric and $\preceq_{pr}$-monotone operator on belief pairs. This means of course that we can then apply proposition 3.11.

Proposition 3.12. *The operator $\mathcal{D}_T$ is symmetric and $\preceq_{pr}$-monotone.*

Proof: It is obvious from the definitions that $\mathcal{D}^1_T(P,S) = \mathcal{D}^2_T(S,P)$. It follows that

$$\mathcal{D}_T(P,S) = (\mathcal{D}^1_T(P,S), \mathcal{D}^1_T(S,P))$$

and

$$\mathcal{D}_T(S,P) = (\mathcal{D}^1_T(S,P), \mathcal{D}^1_T(P,S))$$

and thus that $\mathcal{D}_T(P,S)$ is symmetric. As for the $\preceq_{pr}$-monotonicity, let $(P,S) \preceq_{pr} (P',S')$, i.e. $P \sqsubseteq P'$ and $S' \sqsubseteq S$, i.e. $P' \subseteq P$ and $S \subseteq S'$. Then we need to show that $\mathcal{D}_T(P,S) \preceq_{pr} \mathcal{D}_T(P',S')$, i.e. $\mathcal{D}^1_T(P,S) \sqsubseteq \mathcal{D}^1_T(P',S')$ and $\mathcal{D}^1_T(S',P') \sqsubseteq \mathcal{D}^1_T(S,P)$, i.e. $\mathcal{D}^1_T(P',S') \subseteq \mathcal{D}^1_T(P,S)$ and $\mathcal{D}^1_T(S,P) \subseteq \mathcal{D}^1_T(S',P')$. Let $I \in \mathcal{D}^1_T(P',S')$, so $\mathcal{H}^2_{(S',P'),I}(T) = true$. Since $(P,S) \preceq_{pr} (P',S')$, also $(S',P') \preceq_{pr} (S,P)$. Then by proposition 3.10, $\mathcal{H}^2_{(S,P),I} = true$ and thus $I \in \mathcal{D}^1_T(P,S)$. Proof for the second inclusion is similar by $(P,S) \preceq_{pr} (P',S')$ and proposition 3.10. □

Applying proposition 3.11 means that $\mathcal{D}_T$ always gives a more precise belief pair if a consistent belief pair is given. Also, it ensures the existence of at least one partial expansion: by simply iterating the operator $\mathcal{D}_T$ over $(\mathcal{A}, \emptyset)$, the $\preceq_{pr}$-least element of $\mathcal{B}$, the least fixpoint of $\mathcal{D}_T$ emerges. This fixpoint is called the *Kripke-Kleene fixpoint* for T and is denoted by $KK(T)$. The terminology 'Kripke-Kleene' was chosen by Denecker et al. to reflect the close analogy of this fixpoint with the Kripke-Kleene semantics in Logic Programming (the link with logic programming is treated briefly in section 3.5.1). The following corollary states that $KK(T)$ approximates all (if any) partial expansions of T, is consistent, and provides a sufficient argument for the uniqueness of a partial expansion.

Corollary 3.13. *Let T be an ae-theory. Then*

[4] A slight difference in notation from [8], which uses $\mathcal{D}^l$ and $\mathcal{D}^u$ in stead of $\mathcal{D}^1$ and $\mathcal{D}^2$ respectively.

(a) *the fixpoint $KK(T)$ is consistent, i.e. $KK(T)_1 \sqsubseteq KK(T)_2$,*

(b) *for every partial expansion (P,S) of T, $KK(T) \preceq_{pr} (P,S)$,*

(c) *if $KK(T)$ is complete and thus $KK(T) = (P,P)$ for some P, then $KK(T)$ is the unique partial expansion of T and P is the unique expansion of T.*

Proof: (a) is a direct consequence of item (b) from proposition 3.11, (b) is a direct result from the fact that $KK(T)$ is the $\preceq_{pr}$-least fixpoint of $\mathcal{D}_T$ and (c) is a direct result from (a) and (b). □

To illustrate the above, consider the following example.

Example 3.14. *(example 3.5 continued)* Recall example 3.5, where we considered the ae-language $\mathcal{L}_p$ with the only atom p and $T = \{\neg Lp\}$. The relevant interpretations are $I_p : p \mapsto true$ and $I_{\neg p} : p \mapsto false$ and the four possible world structures are $\mathcal{A}_p = \{I_p, I_{\neg p}\}, \mathcal{I}_p = \{I_p\}, \mathcal{I}_{\neg p} = \{I_{\neg p}\}$ and $\emptyset$. We established that the expansions of T are $\emptyset$ and $\mathcal{A}_p$. Now, let us look at what operator $\mathcal{D}_T$ does to a belief pair (P,S):

$$\begin{aligned}
\mathcal{D}_T(P,S) &= (\mathcal{D}^1_T(P,S), \mathcal{D}^2_T(P,S)) \\
&= (\{I \mid \mathcal{H}^2_{(S,P),I}(\neg Lp) = true\}, \{I \mid \mathcal{H}^2_{(P,S),I}(\neg Lp) = true\}) \\
&= (\{I \mid \mathcal{H}^2_{(P,S),I}(Lp) = false\}, \{I \mid \mathcal{H}^2_{(S,P),I}(Lp) = false\}).
\end{aligned}$$

Now, the falsity of, for instance, $\mathcal{H}_{(P,S),I}(Lp)$ (in the first entry of $\mathcal{D}_T(P,S)$) only depends on P, namely:

$$\begin{aligned}
\mathcal{H}_{(P,S),I}(Lp) = false \quad &\text{iff} \quad \mathcal{H}_{(P,S),J}(p) = false \text{ for some } J \in P \\
&\text{iff} \quad J(p) = false \text{ for some } J \in P
\end{aligned}$$

Therefore, the first entry of $\mathcal{D}_T(P,S)$ is $\mathcal{A}_p$ if and only if for some $J \in P$, $J(p) = false$ and it is $\emptyset$ if and only if for all $J \in P$, $J(p) = true$. The second entry is the same, now with S in stead if P. Hence we get the following table:

(P,S)	$(\mathcal{A}_p, \emptyset)$	$(\mathcal{A}_p, \mathcal{I}_p)$	$(\mathcal{I}_p, \emptyset)$	$(\mathcal{I}_{\neg p}, \emptyset)$
$\mathcal{D}_T(P,S)$	$(\mathcal{A}_p, \emptyset)$	$(\mathcal{A}_p, \emptyset)$	$(\emptyset, \emptyset)$	$(\mathcal{A}_p, \emptyset)$
(P,S)	$(\mathcal{A}_p, \mathcal{I}_{\neg p})$	$(\mathcal{I}_{\neg p}, \mathcal{I}_p)$	$(\mathcal{A}_p, \mathcal{A}_p)$	$(\mathcal{I}_p, \mathcal{I}_p)$
$\mathcal{D}_T(P,S)$	$(\mathcal{A}_p, \mathcal{A}_p)$	$(\mathcal{A}_p, \emptyset)$	$(\mathcal{A}_p, \mathcal{A}_p)$	$(\emptyset, \emptyset)$
(P,S)	$(\emptyset, \emptyset)$	$(\mathcal{I}_{\neg p}, \mathcal{I}_{\neg p})$	$(\mathcal{I}_p, \mathcal{I}_{\neg p})$	$(\mathcal{I}_{\neg p}, \mathcal{A}_p)$
$\mathcal{D}_T(P,S)$	$(\emptyset, \emptyset)$	$(\mathcal{A}_p, \mathcal{A}_p)$	$(\emptyset, \mathcal{A}_p)$	$(\mathcal{A}_p, \mathcal{A}_p)$
(P,S)	$(\mathcal{I}_p, \mathcal{A}_p)$	$(\emptyset, \mathcal{I}_p)$	$(\emptyset, \mathcal{I}_{\neg p})$	$(\emptyset, \mathcal{A}_p)$
$\mathcal{D}_T(P,S)$	$(\emptyset, \mathcal{A}_p)$	$(\emptyset, \emptyset)$	$(\emptyset, \mathcal{A}_p)$	$(\emptyset, \mathcal{A}_p)$

We see that T has four partial expansions, namely $(\mathcal{A}_p, \emptyset)$, $(\mathcal{A}_p, \mathcal{A}_p)$, $(\emptyset, \emptyset)$ and $(\emptyset, \mathcal{A}_p)$, of which $(\mathcal{A}_p, \emptyset)$ is of course the least and thus the Kripke-Kleene fixpoint for T. It is also obvious, since it is the least element in $\mathcal{B}$, that it approximates the other three partial expansions. As a final observation, note that

iterating $\mathcal{D}_T$ to any belief pair leads to a partial expansion. In this example, applying $\mathcal{D}_T$ just once to any belief pair results in a partial expansion. □

Finally (not unimportantly), we explain how partial expansions are linked to expansions in the following proposition.

Proposition 3.15. *Let T be an ae-theory and P a possible world structure. Then $\mathcal{D}_T(P,P) = (D_T(P), D_T(P))$. Consequently, (P,P) is a fixpoint of $\mathcal{D}_T$ if and only if P is a fixpoint of D_T.*

Proof: By definition $\mathcal{D}_T(P,P) = (\mathcal{D}^1_T(P,P), \mathcal{D}^2_T(P,P))$. Also by definition $\mathcal{D}^1_T(P,P) = \mathcal{D}^2_T(P,P) = \{I \mid \mathcal{H}^2_{(P,P),I}(T) = true\}$ and hence by proposition (3.9) this is equal to $\{I \mid \mathcal{H}_{P,I}(T) = true\} = D_T(P)$. This proves $\mathcal{D}_T(P,P) = (D_T(P), D_T(P))$. Obviously it follows that $\mathcal{D}_T(P,P) = (P,P)$ if and only if $D_T(P) = P$. □

3.2.4 Extensions for autoepistemic logic

In the previous section the operator $\mathcal{D}_T$ provided an ae-theory T with a semantics corresponding to expansions and the newly defined partial expansions. In the present section we will derive from $\mathcal{D}_T$ an additional operator D^{st}_T on the set $\mathcal{W}$ of possible world structures and an operator $\mathcal{D}^{st}_T$ on the set $\mathcal{B}$ of belief pairs. The fixpoint of these operators are called the *extensions* and *partial extensions* of T, respectively. This terminology is of course not arbitrarily chosen; the extensions of a default theory Δ correspond precisely to the fixpoints of $D^{st}_{kon(\Delta)}$. Thus, the operator $\mathcal{D}_T$ establishes a uniform semantics for both default and autoepistemic logic. The $\preceq_{pr}$-least fixed point of $\mathcal{D}^{st}_T$ provides for the so-called *well-founded semantics* of autoepistemic logic and is related to the extensions of T in a similar way as the Kripke-Kleene fixpoint of T is related to the expansions of T.

Consider the function $\mathcal{D}^1_T(P,S) = \{I \mid \mathcal{H}^2_{(S,P),I}(T) = true\}$ which we defined in the previous section and constitutes the first entry of $\mathcal{D}_T(P,S)$. We can view this function as an operator on possible world structures in the first variable as follows:

$$\mathcal{D}^1_T(\cdot, S) : P \mapsto \mathcal{D}^1_T(P,S).$$

Of course we will simply write $\mathcal{D}^1_T(P,S)$ for $\mathcal{D}^1_T(\cdot,S)(P)$. Similarly defined, we will consider the function $\mathcal{D}^2_T(P,S) = \{I \mid \mathcal{H}^2_{(P,S),I}(T) = true\}$ (the second entry of $\mathcal{D}_T(P,S)$) as an operator on possible world structures in the second variable:

$$\mathcal{D}^2_T(P, \cdot) : S \mapsto \mathcal{D}^2_T(P,S).$$

It is easy to see from the definitions that

$$\mathcal{D}^1_T(P,S) = \mathcal{D}^2_T(S,P) \quad \text{and so} \quad \mathcal{D}^1_T(\cdot,S) = \mathcal{D}^2_T(S,\cdot). \tag{5}$$

Lemma 3.16. *Let T be an ae-theory and S a possible world structure. The operators $\mathcal{D}^1_T(\cdot, S)$ and $\mathcal{D}^2_T(S, \cdot)$ are $\sqsubseteq$-monotone.*

Proof: Let $P \sqsubseteq P'$. Then since also $S \sqsubseteq S$ it is obvious that $(P, S) \preceq_{pr} (P', S)$. By the $\preceq_{pr}$-monotonicity of $\mathcal{D}_T$ it follows that $\mathcal{D}_T(P, S) \preceq_{pr} \mathcal{D}_T(P', S)$ and hence $\mathcal{D}^1_T(P, S) \sqsubseteq \mathcal{D}^1_T(P', S)$, which proves $\sqsubseteq$-monotonicity of $\mathcal{D}^1_T(\cdot, S)$. By (5), $\mathcal{D}^2_T(S, \cdot)$ is also $\sqsubseteq$-monotone. □

Now that we have established monotonicity, we will look at the least fixed points provided by the Knaster-Tarski theorem. Let

$$D^{st}_T(S) = lfp(\mathcal{D}^1_T(\cdot, S)) = lfp(\mathcal{D}^2_T(S, \cdot)).$$

Similar to D_T this is a revision operator for possible world structures. The fixpoints of D^{st}_T are called *extensions* of T.

Example 3.17. Recall example 3.5 and its continuation 3.14, in which we considered the ae-language $\mathcal{L}_p$ with the only atom p and let $T = \{\neg Lp\}$, i.e., the only thing the agent knows is that he doesn't know p. The relevant interpretations are of course $I_p : p \mapsto true$ and $I_{\neg p} : p \mapsto false$. The four possible world structures are then $\mathcal{A}_p = \{I_p, I_{\neg p}\}, \mathcal{I}_p = \{I_p\}, \mathcal{I}_{\neg p} = \{I_{\neg p}\}$ and $\emptyset$. We established that to come to a stable set with the only knowledge that you don't know p, you can either decide to believe in nothing ($\mathcal{A}_p$) or decide to believe in everything ($\emptyset$). In default logic, however, we cannot just decide to believe everything. This is illustrated in the extensions of T. By definition $\mathcal{D}^{st}_T(Q) = lfp(\mathcal{D}^1_T(\cdot, Q))$. Now, from the table in example 3.14, we can read off entries of $\mathcal{D}^1_T(\cdot, Q)$ very easily. The results are that for any $Q \subseteq \mathcal{A}_p$ the fixpoints of $\mathcal{D}^1_T(\cdot, Q)$ are $\emptyset$ and $\mathcal{A}_p$, of which the $\sqsubseteq$-least is $\mathcal{A}_p$. Thus the table of D^{st}_T is

Q	$\mathcal{A}_p$	$\mathcal{I}_{\neg p}$	$\mathcal{I}_p$	$\emptyset$
$D^{st}_T(Q)$	$\mathcal{A}_p$	$\mathcal{A}_p$	$\mathcal{A}_p$	$\mathcal{A}_p$

We see that T has only one extension, which is believing in nothing, since that is the only justified belief we can deduce from not knowing p. □

Consider the following operator on $\mathcal{B}$:

$$\mathcal{D}^{st}_T(P, S) = (D^{st}_T(S), D^{st}_T(P)). \tag{6}$$

Similar to $\mathcal{D}_T$, $\mathcal{D}^{st}_T$ is a revision operator for belief pairs. The fixpoints of $\mathcal{D}^{st}_T$ are called *partial extensions*. From (6) it follows immediately that

$$\mathcal{D}^{st}_T(P, P) = (D^{st}_T(P), D^{st}_T(P)). \tag{7}$$

Consequently, (P, P) is a fixpoint of $\mathcal{D}^{st}_T$ if and only if P is a fixpoint of D^{st}_T.

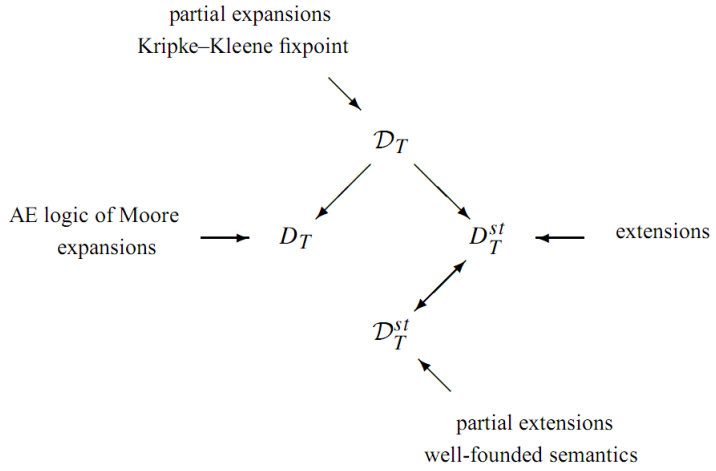


Fig. 1. Operators associated with autoepistemic logic ([8]).

Theorem 3.18. *Let T be an ae-theory. The operator D_T^{st} is anti-monotone and the operator $\mathcal{D}_T^{st}$ is $\preceq_{pr}$-monotone and symmetric. Also, for every consistent belief pair (P, S), $\mathcal{D}_T^{st}(P, S)$ is also consistent.*

Proof: Let $P \sqsubseteq S$ and let $P' = D_T^{st}(P) = lfp(\mathcal{D}_T^1(\cdot, P))$ and $S' = D_T^{st}(S) = lfp(\mathcal{D}_T^1(\cdot, P))$. Clearly (because P' is a fixpoint of $\mathcal{D}_T^1(\cdot, P)$)

$$P' = \mathcal{D}_T^1(P', P). \tag{8}$$

Also, by $\preceq_{pr}$-monotonicity of $\mathcal{D}_T$ and since $(P', S) \preceq_{pr} (P', P)$ it follows that $\mathcal{D}_T(P', S) \preceq_{pr} \mathcal{D}_T(P', P)$ and hence

$$\mathcal{D}_T^1(P', S) \sqsubseteq \mathcal{D}_T^1(P', P). \tag{9}$$

By (8) and (9) $\mathcal{D}_T^1(P', S) \sqsubseteq P'$, i.e. P' is a prefixpoint of $\mathcal{D}_T^1(\cdot, S)$. By lemma 3.16, $\mathcal{D}_T^1(\cdot, S)$ is $\sqsubseteq$-monotone. Then by the Knaster-Tarski theorem, $lfp(mathcal D_T^1(\cdot, S)) \leq x$ for all $\sqsubseteq$-prefixpoints of $mathcal D_T^1(\cdot, S)$. It follows that $S' \sqsubseteq P'$, which concludes the anti-monotonicity of D_T^{st}.

Let $(P, S) \preceq_{pr} (P', S')$. It follows directly from the anti-monotonicity of D_T^{st} that $D_T^{st}(S, P) \preceq_{pr} D_T^{st}(S', P')$, i.e. $\mathcal{D}_T^{st}(P, S) \preceq_{pr} \mathcal{D}_T^{st}(P', S')$, hence concluding monotonicity of $\mathcal{D}_T^{st}$. The symmetry of $\mathcal{D}_T^{st}$ is immediate from the definition.

The last statement is immediate from proposition 3.11. □

Since we now established monotonicity of $\mathcal{D}^{st}$, we will again investigate the least fixpoint, which is called the *well-founded fixpoint* of T and is denoted by $WF(T)$. Denecker et al. chose the name because the semantics specified by the well-founded fixpoint is closely related to the well-founded semantics for default logic and logic programming.

The following corollary states for the well-founded fixpoint similar statements to what corollary 3.13 stated for the Kripke-Kleene fixpoint; $WF(T)$

approximates all (if any) partial extensions of T, is consistent, and provides a sufficient argument for the uniqueness of a partial extension.

Corollary 3.19. *Let T be an ae-theory. Then*

(a) *the fixpoint $WF(T)$ is consistent, i.e. $WF(T)_1 \sqsubseteq WF(T)_2$,*

(b) *for every partial extension (P, S) of T, $WF(T) \preceq_{pr} (P, S)$,*

(c) *if $WF(T)$ is complete and thus $WF(T) = (P, P)$ for some P, then it is the unique partial extension of T and P is the unique extension of T.*

Proof: Since $WF(T)$ can be reached by iterating $\mathcal{D}^{st}_T$ over $(\mathcal{A}, \emptyset)$ (the least, and consistent, element of $\mathcal{B}$) (a) follows directly from theorem 3.18. (b) is a direct result from the fact that $WF(T)$ is the $\preceq_{pr}$-least fixpoint of $\mathcal{D}^{st}_T$ and (c) is a direct result from (a) and (b). □

Example 3.17 (continued) Because of the work in examples 3.5, 3.14 and 3.17 it is now very easy to find the partial extensions of $T = \{\neg Lp\}$. Since the only extension of T is $\mathcal{A}_p$, we know from the definition that $\mathcal{D}^{st}_T(P, S) = (\mathcal{A}_p, \mathcal{A}_p)$ for all P and S. Thus the only partial extension is $(\mathcal{A}_p, \mathcal{A}_p)$. This in particular complies with the corollary above and the theorem below on the relation between the four different operators. □

The following theorem explains the relation between the partial extensions and the partial expansions of a theory T. The order $\sqsubseteq$ on belief pairs in the theorem is the component-wise extension of the knowledge ordering on possible world structures which we briefly mentioned earlier 3.2.2: $(P, S) \sqsubseteq (P', S')$ if and only of $P \sqsubseteq P'$ and $S \sqsubseteq S'$.

Theorem 3.20. *Let T be an ae-theory. Then*

(a) *partial extensions of T are also partial expansions of T,*

(b) *$KK(T) \preceq_{pr} WF(T)$,*

(c) *every partial extension of T is a $\sqsubseteq$-minimal partial expansion of T.*

Proof: For (a), let (P, S) be a partial extension, i.e. a fixpoint of $\mathcal{D}^{st}_T$. Then $P = D^{st}_T(S) = lfp(\mathcal{D}^1_T(\cdot, S))$ and consequently, $\mathcal{D}^1_T(P, S) = P$. Similarly, $S = lfp(\mathcal{D}^1_T(\cdot, P))$ and consequently, $S = \mathcal{D}^1_T(S, P) = \mathcal{D}^2_T(P, S)$, hence concluding that $\mathcal{D}_T(P, S) = (P, S)$, proving (a). Since $KK(T)$ is the least fixpoint of $\mathcal{D}_T$, (b) follows directly from (a).

For (c), assume that (P, S) is a partial extension of T. Then (P, S) is a fixpoint of $\mathcal{D}^{st}_T$ and hence $S = lfp(\mathcal{D}^1(\cdot, P))$. Also assume that (P', S') is a partial expansion of T such that $(P', S') \sqsubseteq (P, S)$. Then we have to show that $(P', S') = (P, S)$. Since $P' \sqsubseteq P$, it follows that $(S', P) \preceq_{pr} (S', P')$. By $\preceq_{pr}$-monotonicity of $\mathcal{D}_T$ we get that $\mathcal{D}^1_T(S', P) \sqsubseteq \mathcal{D}^1_T(S', P) = S'$, and thus that S' is

a $\sqsubseteq$-prefixpoint of $\mathcal{D}^1(\cdot, P)$. Since by lemma 3.16 $\mathcal{D}^1_T(\cdot, P)$ is $\sqsubseteq$-monotone, it follows by the Knaster-Tarski theorem that $lfp(\mathcal{D}^1(\cdot, P)) \sqsubseteq x$ for all $\sqsubseteq$-prefixpoints. It follows that $S = lfp(\mathcal{D}^1(\cdot, P)) \sqsubseteq S'$ and thus that $S = S'$. By similar argumentation we get that $P \sqsubseteq P'$ and thus that $P = P'$, proving (c). $\square$

Direct results of theorem 3.20 are that extensions of T are also expansions of T and that every extension of T is a $\sqsubseteq$-minimal expansion of T. This of course complies with the idea that extensions are a stronger concept than expansions.

3.3 Possible world semantics for default logic

In this section we will define the equivalents for default logic of the operators in the previous section. The operators $\mathcal{E}_\Delta$, E_Δ, E^{st}_Δ and $\mathcal{E}^{st}_\Delta$ will be the equivalents of $\mathcal{D}_T$, D_T, D^{st}_T and $\mathcal{D}^{st}_T$, respectively. Their fixpoints will be called *partial expansions*, *weak extensions*, *extensions* and *partial extensions* and we will indeed prove that the extensions of a default theory Δ correspond exactly to the fixpoints of E^{st}_Δ. In the next section we will prove that the connection between these quadruples of operators is that they are actually the same if $T = kon(\Delta)$. We will proceed by firstly defining for a belief pair $(P, S) \in \mathcal{B}$ a truth function $\mathcal{H}^{dl}_{(P,S),I}$ that provides formulas and defaults with a conservative truth value, as $\mathcal{H}^{2}_{(P,S),I}$ does for ae-formulas. Again, the liberal truth value of formulas and defaults will be found by simply switching the roles of the underestimate P and overestimate S in the conservative truth function.

Definition 3.21. *Let for a given belief pair (P, S) and an interpretation I the truth function $\mathcal{H}^{dl}_{(P,S),I}$ be defined as follows:*

1. *$\mathcal{H}^{dl}_{(P,S),I}(\varphi) = I(\varphi)$, if φ is a propositional formula,*
2. *for a default $\frac{\varphi:\psi_1,\ldots,\psi_n}{\chi}$, $\mathcal{H}^{dl}_{(P,S),I}(\frac{\varphi:\psi_1,\ldots,\psi_n}{\chi}) = true$ just in case at least one of the following conditions hold:*
   (a) *there is a $J \in S$ such that $J(\varphi) = false$,*
   (b) *there is an $i \in \{1, \ldots n\}$ such that for all $J \in P$, $J(\psi_i) = false$,*
   (c) *$I(\chi) = true$.*

   Otherwise $\mathcal{H}^{dl}_{(P,S),I}(\frac{\varphi:\psi_1,\ldots,\psi_n}{\chi}) = false$.

The definition clearly describes a *conservative* truth value for defaults given an underestimate P and an overestimate S: the value of a default is *true* if its prerequisite is false even with respect to the liberal view S, or at least one of its justifications are perceived impossible just by the conservative view P (remember that generally $S \sqsubseteq P$, so P is generally a smaller set than S), or if the conclusion is true. Arguably, $\mathcal{H}^{dl}_{(S,P),I}$ (switching the roles of P and S) provides a liberal estimate for a truth value with respect to the belief pair (P, S), since now the value of a default is *true* if the prerequisite is false just with respect tot the conservative view P, or at least one of its justifications are perceived

impossible even by the liberal view S, or if the conclusion is true. For a default theory $\Delta = (D, W)$, we write $\mathcal{H}^{dl}_{(P,S),I}(\Delta) = true$ if and only $\mathcal{H}^{dl}_{(P,S),I}(d) = true$ for every element (formula or default) d in $D \cup W$.

Recall that $\leq$ orders the values $true$ and $false$ as $false \leq true$. The following proposition is the analogy of proposition 3.10 for default logic:

Proposition 3.22. *Let $(P, S), (P', S') \in \mathcal{B}$ be two belief pairs such that $(P, S) \preceq_{pr} (P', S')$. Let δ be a default, $\varphi \in \mathcal{L}$ a propositional formula and $I \in \mathcal{A}$ an interpretation. Then*

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

(b) $\mathcal{H}^{dl}_{(P,S),I}(\delta) \leq \mathcal{H}^{dl}_{(P',S'),I}(\delta)$.

Proof: Item (a) is trivial from the definition, since the truth value of formulas only depends on the interpretation I. As for (b), suppose $\delta = \frac{\varphi:\psi_1,\ldots,\psi_n}{\chi}$ and suppose $\mathcal{H}^2_{(P,S),I}(d) = true$. We need to show that also $\mathcal{H}^2_{(P',S'),I}(d) = true$. Now, at least one of the following three cases holds.

1. There is a $J \in S$ such that $J(\varphi) = false$. Since $S \subseteq S'$, it then follows that $\mathcal{H}^2_{(P',S'),I}(\delta) = true$.

2. There is an $i \in \{1, \ldots k\}$ such that for all $J \in P$, $J(\psi_i) = false$. Since $P' \subseteq P$, it follows that $\mathcal{H}^2_{(P',S'),I}(\delta) = true$.

3. $I(\chi) = true$. Again, since the value only depends on I, it follows that also $\mathcal{H}^2_{(P',S'),I}(\delta) = true$.

Each of the three cases results in $\mathcal{H}^2_{(P',S'),I}(d)$ being $true$. □

3.3.1 Partial expansions for default logic

Analogous to the operator $\mathcal{D}_T$ for an ae-theory T, we define the operator $\mathcal{E}_\Delta$ on the lattice of belief pairs $\langle \mathcal{B}, \preceq_{pr} \rangle$ for a default theory $\Delta = (D, W)$:

$$\mathcal{E}_\Delta(P, S) = (\mathcal{E}^1_\Delta(P, S), \mathcal{E}^2_\Delta(P, S)),$$

where

$$\mathcal{E}^1_\Delta(P, S) = \{I \mid \mathcal{H}^{dl}_{(S,P),I}(\Delta) = true\}$$

and

$$\mathcal{E}^2_\Delta(P, S) = \{I \mid \mathcal{H}^{dl}_{(P,S),I}(\Delta) = true\}.$$

The fixpoints of $\mathcal{E}_\Delta$ are called the *partial expansions of* Δ. It is obvious that $\mathcal{E}^1_\Delta(P, S) = \mathcal{E}^2_\Delta(S, P)$ and thus we have the following analogy to proposition 3.12:

Proposition 3.23. *The operator $\mathcal{E}_\Delta$ is $\preceq_{pr}$-monotone and symmetric.*

Proof: Symmetry of $\mathcal{E}_\Delta$ directly follows from $\mathcal{E}^1_\Delta(P,S) = \mathcal{E}^2_\Delta(S,P)$ and the definitions above as follows:

$$\mathcal{E}_\Delta(P,S) = (\mathcal{E}^1_\Delta(P,S), \mathcal{E}^1_\Delta(S,P))$$

and

$$\mathcal{E}_\Delta(S,P) = (\mathcal{E}^1_\Delta(S,P), \mathcal{E}^1_\Delta(P,S))$$

which shows symmetry of $\mathcal{E}_\Delta$. As for the monotonicity, let $(P,S) \preceq_{pr} (P',S')$. Then we need to show that $\mathcal{E}_\Delta(P,S) \preceq_{pr} \mathcal{E}_\Delta(P',S')$, i.e. that $\mathcal{E}^1_\Delta(P,S) \sqsubseteq \mathcal{E}^1_\Delta(P',S')$ and $\mathcal{E}^2_\Delta(P',S') \subseteq \mathcal{E}^2_\Delta(P,S)$, i.e. that $\mathcal{E}^1_\Delta(P',S') \subseteq \mathcal{E}^1_\Delta(P,S)$ and $\mathcal{E}^2_\Delta(P,S) \subseteq \mathcal{E}^2_\Delta(P',S')$. Let $I \in \mathcal{E}^1_\Delta(P',S')$. Then $\mathcal{H}^{dl}_{(S',P'),I}(\Delta) = true$. Since $(P,S) \preceq_{pr} (P',S')$ it follows that $(S',P') \preceq_{pr} (S,P)$ and thus by proposition 3.22 that $\mathcal{H}^{dl}_{(S',P'),I}(\Delta) \leq \mathcal{H}^{dl}_{(S,P),I}(\Delta)$. Since $\mathcal{H}^{dl}_{(S',P'),I}(\Delta) = true$ it follows that also $\mathcal{H}^{dl}_{(S,P),I}(\Delta) = true$ and thus that $I \in \mathcal{E}^1_\Delta(P,S)$. The second inclusion is proved analogously: let $I \in \mathcal{E}^2_\Delta(P,S)$. Then $\mathcal{H}^{dl}_{(P,S),I}(\Delta) = true$. By $(P,S) \preceq_{pr} (P',S')$ and proposition 3.22 it follows that that $\mathcal{H}^{dl}_{(P',S'),I}(\Delta) \geq \mathcal{H}^{dl}_{(P,S),I}(\Delta) = true$ and so $I \in \mathcal{E}^2_\Delta(P',S')$. □

We define for a possible world structure Q the following operator:

$$E_\Delta(Q) = \mathcal{E}^1_\Delta(Q,Q) = \mathcal{E}^2_\Delta(Q,Q),$$

so that, similar to the statement in proposition 3.15 the following holds.

Corollary 3.24. *Let Q be a possible world structure. Then $\mathcal{E}_\Delta(Q,Q) = (E_\Delta(Q), E_\Delta(Q))$. Consequently, (Q,Q) is a fixpoint of $\mathcal{E}_\Delta$ if and only if Q is a fixpoint of E_Δ.*

Proof: by definition of $\mathcal{E}_\Delta$ and E_Δ. □

The fixpoints of E_Δ are called the *weak extensions of* Δ (also sometimes called the *expansions of* Δ).

Obviously, the monotonicity of $\mathcal{E}_\Delta$ in proposition 3.23 results via the Knaster-Tarski theorem in a unique $\preceq_{pr}$-least fixed point for $\mathcal{E}_\Delta$. The fixpoint is called the *Kripke-Kleene fixpoint for* Δ (denoted by $KK(\Delta)$). Again, as did the Kripke-Kleene fixpoint for autoepistemic logic in corollary 3.13, $KK(\Delta)$ approximates all (if any) partial expansions of Δ, is consistent, and provides for a sufficient argument for the uniqueness of a partial expansion. This is stated in the following corollary. (Recall that when for example $KK(\Delta)$ is a pair (P,S), $KK(\Delta)_1$ is the first component of the pair (P) and $KK(\Delta)$ the second component (S).)

Corollary 3.25. *Let Δ be a default theory. Then*

(a) *the fixpoint $KK(\Delta)$ is consistent, i.e. $KK(\Delta)_1 \sqsubseteq KK(\Delta)_2$,*

(b) *for every partial expansion (P,S) of Δ, $KK(\Delta) \preceq_{pr} (P,S)$,*

(c) *if $KK(\Delta)$ is complete and thus $KK(\Delta) = (P, P)$ for some P, then $KK(\Delta)$ is the unique partial expansion of Δ and P is the unique expansion of Δ.*

Proof: The proofs are exactly the same as those of corollary 3.13. □

3.3.2 Partial extensions for default logic

Consider, similar to $\mathcal{D}^1(P, S)$ in section 3.2.4, the function $\mathcal{E}^1(P, S)$ as an operator on $\mathcal{W}$ as follows:

$$\mathcal{E}^1(\cdot, S) : P \mapsto \mathcal{E}^1(P, S).$$

We have the following analogy to lemma 3.16:

Lemma 3.26. *Let Δ be a default theory and S a possible world structure. The operators $\mathcal{E}^1_\Delta(\cdot, S)$ and $\mathcal{E}^2_\Delta(S, \cdot)$ are $\sqsubseteq$-monotone.*

Proof: Let $P \sqsubseteq P'$. Then since also $S \sqsubseteq S$ it is obvious that $(P, S) \preceq_{pr} (P', S)$. By the $\preceq_{pr}$-monotonicity of $\mathcal{E}_\Delta$ it follows that $\mathcal{E}_\Delta(P, S) \preceq_{pr} \mathcal{E}_\Delta(P', S)$ and hence $\mathcal{E}^1_\Delta(P, S) \sqsubseteq \mathcal{E}^1_\Delta(P', S)$, which proves $\sqsubseteq$-monotonicity of $\mathcal{E}^1_\Delta(\cdot, S)$. Since $\mathcal{E}^1_\Delta(\cdot, S) = \mathcal{E}^2_\Delta(S, \cdot)$, $\mathcal{E}^2_\Delta(S, \cdot)$ is also $\sqsubseteq$-monotone. □

The $\sqsubseteq$-least fixpoint of $\mathcal{E}^1_\Delta(\cdot, S)$ exists by the Tarski-Knaster theorem, so we can define

$$E^{st}_\Delta(S) = \mathit{lfp}(\mathcal{E}^1_\Delta(\cdot, S))$$

and

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

The operators E^{st}_Δ and $\mathcal{E}^{st}_\Delta$ are revision operators for possible world structures and belief pairs and their fixpoints will be called *extensions of* Δ and *partial extensions of* Δ, respectively. It follows immediately that (P, P) is a fixpoint of $\mathcal{E}^{st}_\Delta$ if and only if P is fixpoint of E^{st}_Δ. As we hinted al along (in particular by the terminology), the fixpoints of E^{st}_Δ turn out to correspond exactly to the extensions of Δ, as we defined them in section 2.1.

Theorem 3.27. *Let $\Delta = (D, W)$ be a default theory. If a possible world structure Q is a fixpoint of E^{st}_Δ then the theory $E = \{\varphi \in \mathcal{L} \mid \forall I \in Q : I(\varphi) = true\}$ is an extension of Δ. Conversely, if E is an extension of Δ, then the possible world structure $Q = \{I \in \mathcal{A} \mid \forall \varphi \in E, I(E) = true\}$ is a fixpoint of E^{st}_Δ.*

Proof: For the first part of the theorem, let Q be a fixpoint of E^{st}_Δ, so $Q = \mathit{lfp}(\mathcal{E}^1_\Delta(\cdot, Q))$, so

$$Q = \mathcal{E}^1_\Delta(Q, Q) = \{I \mid \mathcal{H}^{dl}_{(Q,Q),I}(\Delta) = true\}. \tag{10}$$

Let $E = \{\varphi \in \mathcal{L} \mid \forall I \in Q : I(\varphi) = true\}$. We will show that E is an extension of Δ (i.e. $E = \Gamma_\Delta(E)$) in two parts:

"$\Gamma_\Delta(E) \subseteq E$": we need to show that E complies to **D1**, **D2** and **D3** of definition 2.2. Since $\Gamma_\Delta(E)$ is the least set to do so, we get that $\Gamma_\Delta(E) \subseteq E$.

- For **D1**, let $\varphi \in W$. Then by 10, $I(\varphi) = true$ for all $I \in Q$, so then $\varphi \in E$ by definition of E.
- From the definition of E it also follows directly that E is deductively closed, which proves **D2**.
- For **D3**, let $\delta = \frac{\varphi:\psi_1,\ldots\psi_n}{\chi} \in D$ be a default and $I \in Q$ and assume that $\varphi \in E$ and $E \nvdash \psi_i$ for all $i \in \{1,\ldots,n\}$. Then we need to show that $\chi \in E$. By (10), we know that $\mathcal{H}^{dl}_{(Q,Q),I}(\delta) = true$, i.e. at least one of the following three holds:

  (a) $\exists J \in Q$ such that $J(\varphi) = false$,

  (b) $\exists i \in \{1,\ldots,n\}$ such that $J(\psi_i) = false$ for all $J \in Q$,

  (c) $I(\chi) = true$.

  Since we assumed that $\varphi \in E$, we have that $J(\varphi) = true$ for all $J \in Q$, so (a) is ruled out. Also, since we assumed that $E \nvdash \psi_i$ for all $i \in \{1,\ldots,n\}$ and since E is deductively closed, we have that $\neg\psi_i \notin E$ for all $i \in \{1,\ldots n\}$. Therefore, for any $i \in \{1,\ldots,n\}$ there is a $J \in Q$ such that $J(\psi_1) = true$, which rules out (b). Therefore (c) must hold. Since I is an arbitrarily chosen interpretation in Q, we conclude that $J(\chi) = true$ for all $J \in Q$, hence $\chi \in E$.

"$\Gamma_\Delta(E) \supseteq E$": Consider an interpretation $I \in \mathcal{A}$ such that for every $\varphi \in \Gamma_\Delta(E)$, $I(\varphi) = true$. We will show that $I \in Q$. By (10) it is sufficient to show that $\mathcal{H}^{dl}_{(Q,Q),I}(\Delta) = true$. Firstly, since $W \subseteq \Gamma_\Delta(E)$, it follows by the assumption that $\mathcal{H}^{dl}_{(Q,Q),I}(\varphi) = I(\varphi) = true$ for all $\varphi \in W$. Secondly, consider a default $\delta = \frac{\varphi:\psi_1,\ldots,\psi_n}{\chi} \in \Delta$. Then we need to show that $\mathcal{H}^{dl}_{(Q,Q),I}(\delta) = true$, i.e. one of the three conditions (a)-(c) that we mentioned above holds. Let's suppose that (a) and (b) do not hold, i.e.

- for every $J \in Q$, $J(\varphi) = true$, and
- for each $i \in \{1,\ldots,n\}$ there is a $J \in Q$ such that $J(\psi_1) = true$.

Then we need to show that (c) holds, that is, $I(\chi) = true$. From (a) it follows that $\varphi \in E$. From (b) and the definition of E it follows that $E \nvdash \neg\psi_i$ for all $i \in \{1,\ldots,n\}$. Consequently, by the definition of $\Gamma_\Delta(E)$, we get that $\chi \in \Gamma_\Delta(E)$. Hence $I(\chi) = true$ and $\mathcal{H}^{dl}_{(Q,Q),I}(\delta) = true$. This concludes that

$$\{I \in \mathcal{A} \mid \forall\varphi \in \Gamma_\Delta(E), I(\varphi) = true\} \subseteq Q.$$

Together with the fact that $Q = \{I \in \mathcal{A} \mid \forall\varphi \in E, I(\varphi) = true\}$ we get that

$$\{I \in \mathcal{A} \mid \forall\varphi \in \Gamma_\Delta(E), I(\varphi) = true\} \subseteq \{I \in \mathcal{A} \mid \forall\varphi \in E, I(\varphi) = true\},$$

or equivalently, $E \subseteq \Gamma_\Delta(E)$. The last step uses the fact that E and $\Gamma_\Delta(E)$ are both deductively closed.

For the second part of the theorem we use the concept of generating defaults as explained in section 2.1.1. Recall that a default $\delta = \frac{\varphi:\psi_1,\ldots,\psi_n}{\chi}$ is generating for E if $\varphi \in E$ and $\neg\psi_i \notin E$ for $i \in \{1,\ldots,n\}$. Also, recall theorem 2.9, which says that if E is an extension of a default theory $\Delta = (D, W)$ then $E = Th(W \cup GD_E)$, where GD_E is the set of consequences of all defaults in D that are generating for E.

let E be an extension of Δ. Then we need to show that the possible world structure $Q = \{I \in \mathcal{A} \mid \forall\varphi \in E, I(\varphi) = true\}$ is a fixpoint of E^{st}_Δ. Since E is deductively closed, $E = \{\varphi \in \mathcal{L} \mid \forall I \in Q : \ I(\varphi) = true\}$. We will show that $\mathcal{E}^1_\Delta(Q, Q) \sqsubseteq Q$ (i.e. Q is a pre-fixpoint of $\mathcal{E}^1_\Delta(\cdot, Q)$) and that $Q \sqsubseteq Q'$ for any fixpoint Q' of $\mathcal{E}^1_\Delta(\cdot, Q)$. By the Knaster-Tarski theorem this proves the assertion that $Q = E^{st}_\Delta(Q) = lfp(\mathcal{E}^1_\Delta(\cdot, Q))$.

To prove that $\mathcal{E}^1_\Delta(Q, Q) \sqsubseteq Q$, recall that $\mathcal{E}^1_\Delta(Q, Q) = \{I \in \mathcal{A} \mid \mathcal{H}^{dl}_{(Q,Q),I}(\Delta) = true\}$. Let $I \in Q$. Since E is an extension of (D, W), $W \subseteq E$. Thus, for every $\varphi \in W$, $\mathcal{H}^{dl}_{(Q,Q),I}(\varphi) = I(\varphi) = true$. Next, let $\delta = \frac{\varphi:\psi_1,\ldots,\psi_n}{\chi} \in D$. If δ is a generating default for E then $\chi \in E$ and $I(\chi) = true$ and so $\mathcal{H}^{dl}_{(Q,Q),I}(\delta) = true$. If δ is not a generating default of E then either we have (1) $\varphi \in E$, or (2) there is an $i \in \{1,\ldots,n\}$ such that $J(\psi_i) = true$ for all $J \in Q$. In either case it follows that $\mathcal{H}^{dl}_{(Q,Q),I}(\delta) = true$, as well. Consequently, $\mathcal{H}^{dl}_{(Q,Q),I}(\Delta) = true$ and $I \in \mathcal{E}^1_\Delta(Q, Q)$. Thus, we get $Q \subseteq \mathcal{E}^1_\Delta(Q, Q)$, or equivalently $\mathcal{E}^1_\Delta(Q, Q) \sqsubseteq Q$.

Let us now consider a fixpoint Q' of $\mathcal{E}^1_\Delta(\cdot, Q)$ and define $E' = \{\varphi \mid \forall I \in Q' : \ I(\varphi) = true\}$. Clearly, E' is deductively closed. It is also obvious that $Q' = \mathcal{E}^1_\Delta(Q', Q) = \{I \in \mathcal{A} \mid \mathcal{H}^{dl}_{(Q,Q'),I}(\Delta) = true$. Thus, for every $\varphi \in W$ and for every $I \in Q'$, $I(\varphi) = true$, i.e. $W \subseteq E'$. Consider again a default $\delta = \frac{\varphi:\psi_1,\ldots,\psi_n}{\chi} \in D$ and assume that $\varphi \in E'$ and for every $i \in \{1,\ldots,n\}$, $E \nvdash \psi_i$. It follows that for every $J \in Q'$, $J(\varphi) = true$ and, since E is deductively closed, that for each $i \in \{1,\ldots,n\}$ there is a $J \in Q$ such that $J(\psi_i) = true$. Let $I \in Q'$. Since $\mathcal{H}^{dl}_{(Q,Q'),I}(\delta) = true$, it follows that $I(\chi) = true$ and thus $\chi \in E'$. We have now 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 $\mathcal{E}^1_\Delta(\cdot, Q)$ and that $Q \sqsubseteq Q'$ for any fixpoint Q' of $\mathcal{E}^1_\Delta(\cdot, Q)$. It follows that Q is the least fixpoint of $\mathcal{E}^1_\Delta(\cdot, Q)$. $\square$

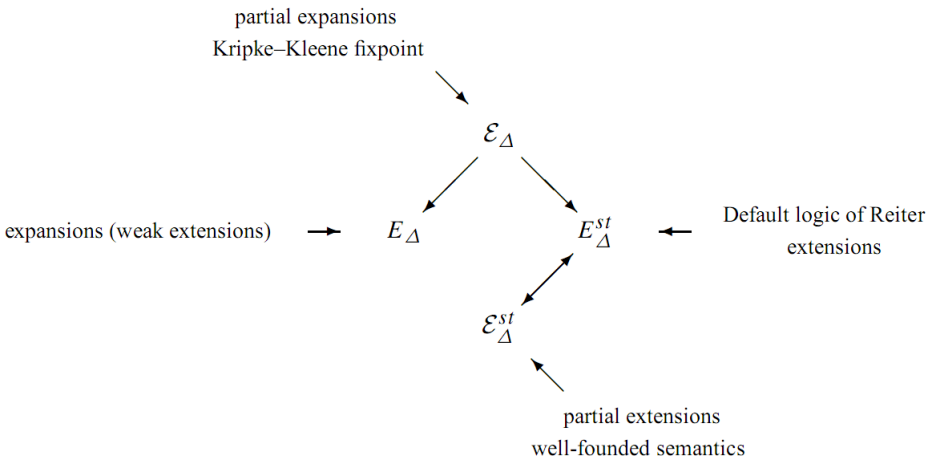


Fig. 2. Operators associated with default logic ([8]).

Analogous to theorem 3.18, we have the following result:

Theorem 3.28. *Let $\Delta = (D, W)$ be a default theory. Then E^{st}_Δ is $\sqsubseteq$-antimonotone and $\mathcal{E}^{st}_\Delta$ is $\preceq_{pr}$-monotone and symmetric. Moreover, for every consistent belief pair (P, S), $\mathcal{E}^{st}_\Delta(P, S)$ is also consistent.*

Proof: The proof is completely analogous to that of theorem 3.18 and will be omitted. □

By $\preceq_{pr}$-monotonicity of $\mathcal{E}^{st}_\Delta$, the Knaster-Tarski theorem ensures the existence of its least fixpoint, which will be called the *well-founded fixpoint of* Δ (denoted by $WF(\Delta)$).

From theorem 3.28 and proposition 3.11, we can conclude the following:

Corollary 3.29. *Let $\Delta = (D, W)$ be a default theory. Then*

(a) *the fixpoint $WF(\Delta)$ is consistent, i.e. $WF(\Delta)_1 \sqsubseteq WF(\Delta)_2$,*

(b) *for every partial extension (P, S) of Δ, $WF(\Delta) \preceq_{pr} (P, S)$,*

(c) *if $WF(\Delta)$ is complete and thus $WF(\Delta) = (P, P)$ for some P, then it is the unique partial extension of Δ and P is the unique extension of Δ.*

Proof: Again, the proof is completely analogous to corollary 3.19. □

Theorem 3.30. *Let Δ be a default theory. Then*

(a) *partial extensions of Δ are also partial expansions of Δ*

(b) *$KK(\Delta) \preceq_{pr} WF(\Delta)$,*

(c) *every partial extension (P,S) of Δ is a $\sqsubseteq$-minimal partial expansion of Δ.*

Proof: Again, the proof is completely analogous to theorem 3.20. □

As might be clear by now, a general approximation theory about these kinds of fixpoint-semantics can be drawn, as has been done by Denecker et al. in [7]. We will sketch this theory very briefly in the following section.

3.4 Approximation theory

In stead of the lattices $\langle \mathcal{W}, \sqsubseteq \rangle$ and $\langle \mathcal{B}, \preceq_{pr} \rangle$ and the specific operators for autoepistemic logic and default logic, we quickly sketch a general theory about approximating operators on arbitrary lattices.

Let $\langle L, \leq \rangle$ be a complete lattice and let $(x, y) \in L^2$ and $(x', y') \in L^2$ be pairs of elements from L. Then L^2 together with the so-called *precision order* defined by

$$(x, y) \leq_{pr} (x', y') \text{ if and only if } x \leq x' \text{ and } y' \leq y$$

forms another complete lattice. A $\leq_{pr}$-monotone and symmetric operator $\mathcal{O} : L^2 \to L^2$ is called an *approximating operator*. If there is an operator $O : L \to L$ such that $\mathcal{O}(x, x) = (O(x), O(x))$ then $\mathcal{O}$ is called an *approximating operator for* O. In particular, since by symmetry $\mathcal{O}(x, x)_1 = \mathcal{O}(x, x)_2$, $\mathcal{O}$ is approximating for $\mathcal{O}(x, x)_1$ and for $\mathcal{O}(x, x)_2$. Consider the operator $\mathcal{O}(\cdot, y)_1 : L \to L$, which takes an element x of L and returns the first component of $\mathcal{O}(x, y)$. We can prove that $\mathcal{O}(\cdot, y)_1$ is $\leq$-monotone and thus has a least fixpoint which constitutes another operator $O^{st}(y) = \mathit{lfp}(\mathcal{O}(\cdot, y)_1)$. The operator $\mathcal{O}^{st}(x, y) = (O^{st}(y), O^{st}(x))$ is called the *stable operator for* $\mathcal{O}$ and can be proven to be $\leq_{pr}$-monotone. The least fixpoints of $\mathcal{O}$ and $\mathcal{O}^{st}$ are called the *Kripke-Kleene fixpoint of* $\mathcal{O}$ and the *well-founded fixpoint of* $\mathcal{O}$, respectively (for now we use the notations $KK(\mathcal{O})$ and $WF(\mathcal{O})$). Moreover, we can prove, among others, the following analogies to theorems and corollaries about the operators $\mathcal{O}, O, O^{st}$ and $\mathcal{O}^{st}$ in the previous sections:

- fixpoints of $\mathcal{O}^{st}$ are also fixpoints of $\mathcal{O}$,
- fixpoints of $\mathcal{O}^{st}$ are $\leq$-minimal fixpoints of $\mathcal{O}$, where $\leq$ is the componentwise extension of the order $\leq$ on the lattice L to the lattice L^2,
- the Kripke-Kleene fixpoint $KK(\mathcal{O})$ approximates all fixpoints of $\mathcal{O}$,
- the well-founded fixpoint $WF(\mathcal{O})$ approximates all fixpoints of $\mathcal{O}^{st}$,
- if $KK(\mathcal{O})$ is complete then it is the only fixpoint of $\mathcal{O}$,
- if $WF(\mathcal{O})$ is complete then it is the only fixpoint of $\mathcal{O}^{st}$,
- $KK(\mathcal{O}) \leq_{pr} WF(\mathcal{O})$,
- $WF(\mathcal{O})$ is consistent.

Obviously, by filling in $\mathcal{D}_T$, D_T, D_T^{st}, $\mathcal{D}_T^{st}$ or $\mathcal{E}_\Delta$, E_Δ, E_Δ^{st}, $\mathcal{E}_\Delta^{st}$ for $\mathcal{O}$, O, O^{st} and $\mathcal{O}^{st}$, we get the approximating theories for autoepistemic logic and default logic as we explained them in detail above.

However, the approximating theories of autoepistemic logic and default logic in particular are connected even more closely by the Konolige translation, as we explain in the following section.

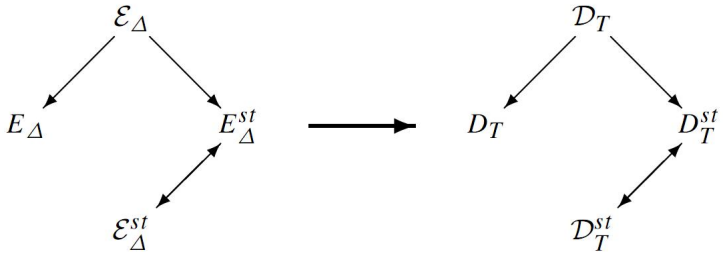
3.5 Link between default and autoepistemic logic

As we explained in section 2.3 of the same name, Konolige's translation does not establish a correspondence between the extensions of a default theory and the expansions of an ae-theory. However, the sections above provide for different semantics for both default and autoepistemic logic between which the Konolige translation *does* establish a direct correspondences. The following theorem is a culmination in the works of Denecker, Marek and Truszczyński.

Theorem 3.31. *Let $\Delta = (D, W)$ be a default theory and $T = kon(\Delta)$ its Konolige translation. Then*

1. $\mathcal{E}_\Delta = \mathcal{D}_T$,
2. $E_\Delta = D_T$,
3. $E_\Delta^{st} = D_T^{st}$,
4. $\mathcal{E}_\Delta^{st} = \mathcal{D}_T^{st}$.

In particular, Δ and T have the same (partial) extensions, (partial) expansions, Kripke-Kleene fixpoint and well-founded fixpoint in terms of possible world structures.



$$\frac{\varphi : \psi_1, \ldots, \psi_n}{\chi} \mapsto L\varphi \wedge \neg L\neg\psi_1 \wedge \ldots \wedge \neg L\neg\psi_n \rightarrow \chi$$

Fig. 3. Embedding default logic into autoepistemic logic ([8].

We will use the following lemma to prove the theorem.

Lemma 3.32. *Let (P, S) be a belief pair. Then for every interpretation $I \in \mathcal{A}$ and every default δ*

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

Proof: Let δ be of the form $\frac{\varphi:\psi_1,\ldots,\psi_n}{\chi}$. Then $kon(\delta) = \neg L\varphi \vee Ł\neg\psi_1 \vee \ldots \vee L\neg\varphi_n \vee \chi$. Now, $\mathcal{H}^{dl}_{(P,S),I}(\delta) = false$ if and only if the following conditions all hold:

1. for all $J \in S$, $J(\varphi) = true$,

2. for every $i \in \{1, \ldots n\}$ there is a $J \in P$ such that $J(\psi_i) = true$,

3. $I(\chi) = false$.

The conditions 1, 2 and 3 above are equivalent to the following conditions *i*, *ii* and *iii* respectively.

i. $\mathcal{H}^2_{(S,P),I}(L\varphi) = true$, which in turn is equivalent to $\mathcal{H}^2_{(P,S),I}(\neg L\varphi) = false$,

ii. for every $i \in \{1, \ldots n\}$ there is a $J \in P$ such that $J(\neg\psi_i) = false$, which in turn is equivalent to $\mathcal{H}^2_{(P,S),I}(L\neg\psi_i) = false$

iii. $\mathcal{H}^2_{(P,S),I}(\chi) = true$

The set of conditions *i-iii* holds if and only if $\mathcal{H}^2_{(P,S),I}(kon(\delta)) = false$, which concludes the claim. □

Proof of theorem 3.31: By lemma 3.32, it follows directly that for any belief pair (P, S)

$$\{I \mid \mathcal{H}^{dl}_{(P,S),I}(\Delta) = true\} = \{I \mid \mathcal{H}^2_{(P,S),I}(kon(\Delta)) = true\}, \tag{11}$$

1. From (11) it follows directly that $\mathcal{E}_\Delta(P, S) = \mathcal{D}_{kon(\Delta)}(P, S)$.
2. Recall that $E_\Delta(P) = Q$ iff $\mathcal{E}_\Delta(P, P) = (Q, Q)$, and thus by the previous equality, iff $\mathcal{D}_{kon(\Delta)}(P, P) = (Q, Q)$, which, by proposition 3.15, happens iff $D_{kon(\Delta)}(P) = Q$.
3. By $\mathcal{E}_\Delta(P, S) = \mathcal{D}_{kon(\Delta)}(P, S)$ it immediately follows that

   $$\mathcal{E}^1_\Delta(\cdot, S) = \mathcal{D}^1_{kon(\Delta)}(\cdot, S)$$

   and thus that $lfp(\mathcal{E}^1_\Delta(\cdot, S)) = lfp(\mathcal{D}^1_{kon(\Delta)}(\cdot, S))$. This proves $E^{st}_\Delta = D^{st}_T$.
4. From the previous equation, it follows directly that

$$\begin{aligned} \mathcal{E}^{st}_\Delta(P, S) &= (E^{st}_\Delta(S), E^{st}_\Delta(P)) \\ &= (D^{st}_T(S), D^{st}_T(P)) \\ &= \mathcal{D}^{st}_T(P, S). \end{aligned}$$

It immediately follows from equations 1-4 that Δ and $kon(\Delta)$ have the same (partial) extensions and (partial) expansions in terms of possible world structures. □

3.5.1 The link with logic programming

A point that should be addressed at least briefly is the link with logic programming with negation, if only by a description. A logic program with negation is a set of formulas called *rules*, which are of the form

$$c \leftarrow a_1, \ldots a_n, \mathbf{not} b_1, \ldots, \mathbf{not} b_m,$$

where all a_i, all b_i and c are literals; c is called the *head of* r and a_1, ..., a_n, $\mathbf{not} b_1$, ..., $\mathbf{not} b_m$ the *body of* r. The symbol **not** denotes negation as failure, i.e. $\mathbf{not}\varphi$ is true if all quests for a proof of the truth of φ end in failure. This should not be confused with $\neg\varphi$, which is true if there is proof that φ is false. In this difference between negation and negation as failure one can see the analogy with the application of a default rule: the consequent of a default is true if the prerequisite is true and *there is no proof* of the negation of the justifications of the default. A logic programming rule can thus be translated to default theory as follows (where we call Δ the *default translation* of the rule):

$$\Delta : c \leftarrow a_1, \ldots, a_n, \mathbf{not} b_1, \ldots, \mathbf{not} b_m \mapsto \frac{a_1 \wedge \ldots \wedge a_n : \neg b_1 \wedge \ldots \wedge \neg b_m}{c}.$$

The default translation $\Delta(P)$ of a logic program P is simply the set of default translations of all rules in P.

Two distinct semantics for logic programming with negation are the so-called supported models and stable models. If P is a set of programming clauses then a model I is a *supported model* of P iff it is exactly the set of heads of clauses in P of which the bodies are made true by I. In other terms, it is a fixpoint of the operator

$$T_P : I \mapsto \{c \mid c \text{ is the head of a rule } r \text{ in } P \text{ such that } I \text{ makes true the body of } r\}.$$

The operator T_P is called the *immediate consequence operator for* P or the *Van Emden-Kowalski operator for* P ([22]).

Later, Gelfond and Lifschitz ([10]) proposed an alternative semantics based on an operator GL_P, the fixpoints of which are called *stable models*. In this semantics an atom is only true if it has a constructive argumentation. We will illustrate what a constructive argumentation means by an example:

Example 3.33. Consider the following program

$$P = \{r_1 = (p \leftarrow p), r_2 = (q \leftarrow \mathbf{not} p)\}.$$

Then P has two supported models $E_1 = \{p\}$ and $E_2 = \{q\}$. Indeed, E_1 proves only the body of r_1, so $T_P(E_1) = E_1$. Also, E_2 does not prove p, hence it proves $\mathbf{not} p$, and so $T_P(E_2) = E_2$. However, in E_1 the formula p has no constructive argumentation. It is simply chosen to be true, after which the set is automatically closed under rule r_1. The formula q in E_2 does have a constructive argumentation: since we have no given facts, we cannot prove the truth of

p and therefore we establish that **not**p is true, after which we can apply rule r_2 to conclude q. By this argumentation the supported model E_2 is also a stable model as defined by Gelfond and Lifschitz whereas E_1 is not. □

The difference between the two semantics explained above is the same as that between default logic and autoepistemic logic in a quote from [8] we cited before: "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". Indeed, in [7], Denecker et al. explain that a so-called 4-valued van Emden-Kowalski operator (denoted by $\mathcal{T}_P$) does for the semantics of supported models and stable models of a logic program P exactly what the operators $\mathcal{D}_T$ and $\mathcal{E}_\Delta$ do for the extensions and expansions of an autoepistemic theory T and a default theory Δ, respectively. Moreover, the default translation connects the semantics of logic programming and default logic in a similar way that the Konolige translation connects those of default and autoepistemic logic. In other terms, the fixpoints of the operators T_P, GL_P and $\mathcal{T}_P$ of a logic program P correspond precisely to the fixpoints of $E_{\Delta(P)}$, $E^{st}_{\Delta(P)}$ and $\mathcal{E}_{\Delta(P)}$ of the default theory $\Delta(P)$, which in their turn correspond precisely to the fixpoints of $D_{kon(\Delta(P))}$, $D^{st}_{kon(\Delta(P))}$ and $\mathcal{D}_{kon(\Delta(P))}$ of the autoepistemic theory $kon(\Delta(P))$. Indeed, the example above illustrates this:

Example 3.33 (continued). When looking at the example above, we get that there is only one extension of the default translation of P:

$$\Delta(P) = (\{\frac{p : true}{p}, \frac{true : \neg p}{q}\}, \emptyset),$$

being $Th(\{q\})$, which corresponds to the stable model E_2 of P. Also, we get that there are two expansions of the Konolige translation of $\Delta(P)$:

$$kon(\Delta(P)) = \{Lp \rightarrow p, \neg Lp \rightarrow q\},$$

being the stable sets with kernels $Th(\{p\})$ and $Th(\{q\})$, which correspond to the supported models of P. □

4 Prioritizing default logic and autoepistemic logic

As we saw many times now, default theories and autoepistemic theories often have multiple extensions and/or expansions. In terms of possible world structures this manifests itself in multiple fixpoints of the operators in the previous chapter. Suppose for example that P and Q are two expansions of an ae-theory $T \subseteq \mathcal{L}_{ae}$. This means that (P, P) and (Q, Q) are both partial expansions, i.e. fixpoints of $\mathcal{D}_T$. Since $\mathcal{D}_T$ is $\preceq_{pr}$-monotone, an application of the operator to a belief pair will result in a revision of the belief pair that is at least as precise. But since (P, P) and (Q, Q) are both complete belief pairs neither of them can be made more precise via operator $\mathcal{D}_T$; that also means that neither (P, P) nor (Q, Q) is more precise than the other and thus that the relevance of them as belief pairs is incomparable. Now, this isn't always a problem. In some situations there is no expansion or extension that is preferred over the others. For instance this is the case in the well-known example of the Nixon diamond:

quakers are pacifists
Nixon is a quaker

Nixon is a pacifist

republicans are not pacifists
Nixon is a republican

Nixon is not a pacifist

Obviously there are two extensions: one in which Nixon is a pacifist and one in which he is not. With no extra information none of the two is preferred over the other. In our Tweety example however, it is clear that we prefer the extension in which Tweety cannot fly. This is because the information that Tweety is a penguin is more *specific* than the information that Tweety is a bird: the general rule is that birds fly - penguins are a specific type of bird that is an exception to this rule.

In general, if there are multiple extensions or expansions we want to be able to somehow express preference of some extensions over other extensions. A way of attacking this problem is through prioritization, i.e. to describe how to prefer one expansion over the other, given a preference relation on some of the formulas involved. In the Tweety example, we could for instance express the fact that we would prefer non-flying birds over flying penguins by the preference relation $bird \wedge \neg flies > penguin \wedge flies$, and from this deciding that the expansion containing $bird, penguin$ and $\neg flies$ is preferred to the expansions containing $bird, penguin$ and $flies$.

Rintanen described in [19] several ways of lexicographically prioritizing default logic. In this section we will explain this method of prioritizing default logic and extend it to autoepistemic logic using the Konolige translation. Furthermore, we will investigate a preference relation derived from these methods.

We will finish by stating a problem involving the works of Denecker et al. as explained in the previous section: how can the prioritization methods of extensions and expansions in default logic and autoepistemic logic be connected.

4.1 How to handle preferences

A considerable number of classifying approaches to preference handling in nonmonotonic logic is described in [5]. To distinguish the properties of lexicographical prioritization from others, we mention especially the following classifying approaches:

Meta-level vs. object-level preferences. Whether the preferences are imposed "externally" on the rules of the system or whether the preferences themselves become objects within the system. The first approach could for example concern the finding of the most preferred expansion of a theory, given some externally imposed preference relation on the formulas in the theory, and thus is a way of formalizing preferences about a theory. The second approach allows the use of preferences on the object-level; one would be able to reason about these preferences within the theory, for example by including formulas like $\varphi \rightarrow \delta_1 < \delta_2$ or defaults like $\frac{\varphi : \delta_1 < \delta_2}{\delta_1 < \delta_2}$. The first is easier to implement, the second is more flexible. As we will later see, lexicographical prioritization is a meta-level approach.

Properties of the preference structure. We will mostly be dealing with irreflexive partial orders as preference orders, properties which are most widely used and accepted when dealing with preference structures. Indeed,

- nothing should be more preferred than itself (irreflexivity),
- if A is more preferred than B then B is not more preferred than A (asymmetry),
- if A is more preferred than B and B is more preferred than C then A is more preferred than C (transitivity), and
- one need not have a preference of A over B or vice versa for any two entities A and B (partiality).

Sometimes we will impose some restrictions such as totality or absence of infinite ascending chains to ensure some results.

Prescriptive vs. descriptive preferences. In ***de***scriptive prioritization one defines why some expansions of a theory are more preferred than others. This gives a ranking on desired outcomes: the preferred outcome is the one where (for example) the most preferred rules are applied. In ***pre***scriptive prioritization the definition of the semantics is altered, for example by requiring that higher priority rules are always applied (if possible), before the lower priority ones are considered. A consequence could be that

the prioritized expansions are not expansions in the non-prioritized sense, since the definition is altered.

The distinction between these two methods and their terminology is illustrated in the following example from [4]: consider the default theory with $W = \emptyset$ and $D = \{\frac{\varphi:\psi}{\psi}, \frac{true:\neg\psi}{\neg\psi}, \frac{true:\varphi}{\varphi}\}$, where the first default is more preferred than the second, and the second is more preferred than the third. A *pre*scriptive method would fail to apply the most preferred default since the prerequisite φ is not provable. However, one might expect to apply the two lesser-preferred defaults, giving an extension containing $\{\varphi, \neg\psi\}$. In a descriptive interpretation, one might observe that by applying the least-preferred default, the most preferred default can be applied; this yields an extension containing $\{\varphi, \psi\}$.

As we will see, lexicographic prioritization is a *de*scriptive method.

4.2 Lexicographic prioritization for default logic

Rintanen defined four so-called lexicographic ways of defining the most preferred extension(s) among the extensions of a default theory, given a strict partial order on the set of defaults. From now on, if two defaults δ and δ' are related by a given order $>$ (i.e. $\delta > \delta'$), then δ is considered to have higher priority than δ'; application of δ is more desirable than application of δ'.

Rintanen used a definition of applied defaults and defeated defaults to define the distinct ways of lexicographic prioritization.

Definition 4.1. *Let $\Delta = (D, W)$ be a default theory and $E \subseteq \mathcal{L}$ a set of formulas. A default $\delta = \frac{\varphi:\psi_1,\ldots,\psi_n}{\chi} \in D$ is said to be* applied *in E (notation $app(\delta, E)$) if $\varphi \in E$ and $E \nvdash \neg\psi_i$ for $i \in \{1, \ldots, n\}$. The default δ is said to be* defeated *in E (notation $def(\delta, E)$) if $\varphi \in E$ and $E \vdash \neg\psi_i$ for some $i \in \{1, \ldots, n\}$.*

When E is an extension and $app(\delta, E)$ then of course $\chi \in E$ as well, in accordance with **D3** of the definition of extensions (definition 2.2). Note however that $def(\delta, E)$ need not imply that $\chi \notin E$. Also, note that $app(\delta, E)$ implies $\neg def(\delta, E)$ and $def(\delta, E)$ implies $\neg app(\delta, E)$, but the converses do not hold in general. Indeed, $\neg app(\delta, E)$ and $\neg def(\delta, E)$ are both true if and only if $pre(\delta) \notin E$, in which case the default cannot even be considered for application. When δ is prerequisite-free (i.e. $pre(\delta) = true$) the converses obviously do hold and thus $app(\delta, E)$ if and only if $\neg def(\delta, E)$. For convenience, we write $app(\delta, E, E')$ for $app(\delta, E) \wedge \neg app(\delta, E')$ and $def(\delta, E, E')$ for $def(\delta, E) \wedge \neg def(\delta, E')$.

Now that we defined the ways of measuring a defaults success (application in an extension) or failure (defeat in an extension) we can define when an extension is a preferred extension, relative to a preference order on the set of defaults. One way to do this is to prefer an extension that has the more-preferred defaults applied in it. Another way is to prefer an extension with the lesser-preferred defaults defeated in it. The following definition by Rintanen elaborates.

Definition 4.2. *Let $\Delta = (D, W)$ be a default theory and $>$ a strict partial order on the set of defaults D. A set E is a* $D_>$-preferred extension *of Δ iff it is an extension of Δ and there is a strict total order $>'$ on D extending the partial order $>$ (i.e. $> \subseteq >'$), such that for all extensions E' of D:*

$$\forall \delta \in D\ (app(\delta, E', E) \rightarrow \exists \delta' \in D\ (app(\delta', E, E') \wedge \delta' >' \delta)). \qquad (12)$$

Also, E is a $D_>$-preferred$_2$ extension *of Δ iff it is an extension of Δ and for all extensions E' of Δ there is a strict total order $>'$ on D extending $>$ such that (12) holds. By replacing (12) by*

$$\forall \delta \in D\ (def(\delta, E, E') \rightarrow \exists \delta' \in D\ (def(\delta', E', E) \wedge \delta' >' \delta)). \qquad (13)$$

we obtain the definitions of $D_>$-preferred$_d$ extension of Δ *and* $D_>$-preferred$_{2d}$ extension of Δ*, respectively.*

The four ways of prioritization in the definition are thus obtained by switching the order of the quantifiers "for all extensions" and "there is a strict total order" and by choosing either (12) or (13) as a means of preference. Note that for prerequisite-free defaults (12) and (13) are the same, since then

$$\begin{array}{lll} app(\delta, E, E') & \text{iff} & app(\delta, E) \wedge \neg app(\delta, E') \\ & \text{iff} & \neg def(\delta, E) \wedge def(\delta, E') \\ & \text{iff} & def(\delta, E', E). \end{array}$$

Example 4.3. Consider the following example from [19] as an illustration of the definitions:

$$\Delta = (\emptyset, (\delta_1 = \frac{true : \neg q \wedge r}{\neg q \wedge r}, \delta_2 = \frac{true : p}{p}, \delta_3 = \frac{true : \neg p \wedge \neg r}{\neg p \wedge \neg r}, \delta_4 = \frac{true : q}{q})).$$

The default theory Δ has four extensions:

$$E_{\delta_2\delta_4} = Th(\{p, q\}), E_{\delta_1\delta_2} = Th(\{\neg q, r, p\}) and E_{\delta_3\delta_4} = Th(\{\neg p, \neg r, q\})$$

(the indices indicate which defaults are applied in the extensions). Consider the preference structure

$$D_> = \{\delta_1 > \delta_2, \delta_3 > \delta_4\}.$$

There are six possible total relations containing the partial preference relation, namely

$$\begin{array}{ll} \{\delta_1 >_1 \delta_2 >_1 \delta_3 >_1 \delta_4\} & \{\delta_3 >_4 \delta_4 >_4 \delta_1 >_4 \delta_2\} \\ \{\delta_1 >_2 \delta_3 >_2 \delta_2 >_2 \delta_4\} & \{\delta_3 >_5 \delta_1 >_5 \delta_4 >_5 \delta_2\} \\ \{\delta_1 >_3 \delta_3 >_3 \delta_4 >_3 \delta_2\} & \{\delta_3 >_6 \delta_1 >_6 \delta_2 >_6 \delta_4\} \end{array}$$

Then $>_1$ and $>_4$ illustrate both the $D_>$-preferredness and $D_>$-preferredness$_2$ of $E_{\delta_1\delta_2}$ and $E_{\delta_3\delta_4}$ respectively. Also, $E_{\delta_2\delta_4}$ is $D_>$-preferred$_2$, since $>_1$ satisfies (13)

for $E = E_{\delta_2\delta_4}$ and $E' = E_{\delta_3\delta_4}$ and $>_4$ satisfies (13) for $E' = E_{\delta_1\delta_2}$. However, $E_{\delta_2\delta_4}$ is not $D_>$-preferred, since $>_1, >_2$ and $>_3$ refute (12) with $E' = E_{\delta_1\delta 2}$ and $>_4, >_5$ and $>_6$ refute (12) with $E' = E_{\delta_3\delta_4}$. In this example, all the above also holds for $D_>$-preferredness$_d$ in stead of $D_>$-preferredness and $D_>$-preferredness$_{2d}$ in stead of $D_>$-preferredness$_2$. □

Proposition 4.4. *If E is a $D_>$-preferred ($D_>$-preferred$_d$) extension of Δ then it is also a $D_>$-preferred$_2$ ($D_>$-preferred$_{2d}$) extension of Δ. The converses do not hold.*

Proof: The implications are immediate from definition 4.2; the converses are refuted by the example above. □

As we said in section 4.1, lexicographical prioritization is a descriptive method of preference handling. This is quite literally illustrated by the phrase "A set E is a $D_>$-preferred extension of Δ iff *it is an extension of* Δ *and* there is ..." in definition 4.2. We thus get a partition of the set of extensions of Δ in a set of $D_>$-preferred extensions and not $D_>$-preferred extensions not by altering the definition of extensions, but rather by describing why some extensions are more preferred than others. This is essentially why we can easily extend the definition to include *weak* extensions, too. Application of a default is namely defined in any set E; whether E is an extension, a weak extension, or any other set of formulas. So, by replacing "extension" by "weak extension" in definition 4.2 we obtain the definitions of $D_>$-preferred$_{(2)(d)}$ weak extensions of Δ. (By "$D_>$-preferred$_{(2)(d)}$" we mean "$D_>$-preferred, $D_>$-preferred$_2$, $D_>$-preferred$_d$ and $D_>$-preferred$_{2d}$, respectively".)

Definition 4.2 directly determines subsets of the set of extensions of a default theory, namely the $D_>$-preferred ones and the $D_>$-preferred$_2$ ones etcetera. It can also be used to create the following preference relation on extensions, allowing us to compare the priorities of extensions pairwise: given a strict total order $>$ on the set of defaults D and two expansions E and E', let

$$E >_D E' \text{ iff } E \neq E' \wedge \forall\delta(app(\delta, E', E) \rightarrow \exists\delta'(app(\delta', E, E') \wedge \delta' > \delta)), \quad (14)$$

$$E >_D^d E' \text{ iff } E \neq E' \wedge \forall\delta(def(\delta, E, E') \rightarrow \exists\delta'(def(\delta', E', E) \wedge \delta' > \delta)). \quad (15)$$

It is evident that (14) is derived from (12) and (15) from (13). In the first pair it is more important to include preferred defaults than to exclude less-preferred ones and in the second pair vice versa. Rintanen described this as the difference between "choosing for presence" and "choosing for absence". We will elaborate on these preference relations in the following sections.

4.3 Lexicographic prioritization for autoepistemic logic

In default logic, conflicting defaults are the cause of multiple extensions. When prioritizing default logic, the set of defaults is therefore the obvious choice for imposing a preference order on; we can decide which defaults to apply by stating

which ones are preferred (for instance lexicographically). In autoepistemic logic there are no special default rules. Therefore the freedom to choose a preference order is much greater, since we can impose these preferences on any set of ae-formulas. The formulas that are involved in an agents preferences together with the preference relation are gathered in a tuple called a *prioritization*.

Definition 4.5. *Let $P \subseteq \mathcal{L}_{ae}$ be a set of autoepistemic formulas and $>\subseteq P \times P$ a strict partial order on P. Then we call the tuple $\langle P, > \rangle$ a* prioritization *which we will denote by $P_>$.*

The prioritization contains all formulas that are involved in the preference structure of the agent, so those are the only formulas we need to consider when trying to prioritize the set of expansions lexicographically. This means that if E and E' are two expansions and $P_>$ the prioritization, we will consider the formulas in the sets $E \cap P$ and $E' \cap P$ to establish which of the two expansions is more preferred.

4.3.1 Translating prioritizing

Our objective is to prioritize autoepistemic logic in a similar way as we did default logic. What does that mean exactly? We saw in chapter 3 what the exact relation is between default logic and autoepistemic logic and their semantics, namely: if Δ is a default theory and $T = kon(\Delta)$ then E is an extension or a weak extension of Δ if and only if E is the kernel of an extension or an expansion of T, respectively. (Let us call the extension or expansion of which E is the kernel E_L.) Therefore, we want to lexicographically prioritize autoepistemic logic in such a way that a $D_>$-preferred extension or weak extension of Δ is the kernel of a $kon(D)_>$-preferred extension or expansion of T, respectively (and the same for preferred$_2$, preferred$_d$ and preferred$_{2d}$). To do this, we turn to definition 4.2 and "translate" this to autoepistemic logic and subsequently generalize it to all formulas. Recall that for a default $\delta = \frac{\varphi : \psi_1, \ldots, \psi_n}{\chi}$:

$$\begin{aligned} app(\delta, E, E') \quad &\text{iff} \quad app(\delta, E) \text{ and } \neg app(\delta, E'), \\ &\text{iff} \quad \varphi \in E \wedge E \nvdash \neg\psi_i \vee \ldots \vee \neg\psi_n \\ &\qquad \text{and } \neg(\varphi \in E' \wedge E' \nvdash \neg\psi_i \vee \ldots \vee \neg\psi_n). \end{aligned}$$

The Konolige translation of δ is $L\varphi \wedge \neg L\neg\psi_1 \wedge \ldots \wedge \neg L\neg\psi_n \rightarrow \chi$, so application of δ in a set E can best be translated as including

$$L\varphi \wedge \neg L\neg\psi_1 \wedge \ldots \wedge \neg L\neg\psi_n \tag{16}$$

in the smallest stable set containing E; non-application of δ in E' can best be translated as the absence of (16) in the smallest stable set containing E'. Indeed, if E is an extension or expansion, because it is deductively closed, $E \nvdash \psi$ will hold if and only if $\psi \in E$ if and only if $L\psi \in E$, so $app(\delta, E, E')$ holds for extensions or expansions E and E' of Δ if and only if (16) $\in F \backslash F'$ for the corresponding extensions or expansions F and F' of $kon(\Delta)$. For convenience,

we will denote (16) by $kon_{app}(\delta)$. As a direct translation from default logic to autoepistemic logic, we get that (according to definition 4.2) an (autoepistemic!) extension E of $kon(\Delta)$ is $kon(D)_>$-preferred iff there is a strict total order $>'$ extending $>$ such that for every extension E' and $\delta \in D$, if $kon_{app}(\delta) \in (E'\backslash E)$ then there is a $\delta' \in D$ such that $kon_{app}(\delta') \in (E\backslash E')$ and $\delta' >' \delta$. By a simple generalization of formulas of the form $kon_{app}(\delta)$ to arbitrary formulas and of prioritizations of the form $kon(D)_>$ to arbitrary prioritizations we come to the definition of $P_>$-preferredness in autoepistemic logic below. We get the definition of $P_>$-preferredness$_2$ again by switching the quantifiers "there is a strict total order" and "for all extensions E'".

Definition 4.6. *Let $P_>$ be a prioritization and T an ae-theory. Then E is a $P_>$-preferred extension of T iff it is an extension of T and there is a total order $>'$ on P extending the partial order $>$ (i.e. $>\subseteq>'$) such that for all extensions E' of T*

$$\forall\varphi \in (E'\backslash E) \cap P \; \exists\psi \in (E\backslash E') \cap P \; (\psi >' \varphi). \tag{17}$$

Also, E is a $P_>$-preferred$_2$ extension of T iff for all extensions E' of T there is a total order $>'$ on P extending the partial order $>$ such that (17) holds.

By replacing "extension" by "expansion" in the definition above we get the notions *$P_>$-preferred* and *$P_>$-preferred$_2$ expansions of T*. For translating the notions $D_>$-preferred$_d$ and $D_>$-preferred$_{2d}$ from default logic to autoepistemic logic, recall that

$$\begin{aligned} def(\delta, E, E') \quad &\text{iff} \quad def(\delta, E) \text{ and } \neg def(\delta, E'), \\ &\text{iff} \quad pre(\delta) \in E \wedge E \vdash \neg\psi_i \vee \ldots \vee \neg\psi_n \\ &\qquad \text{and } \neg(\varphi \in E' \wedge E \vdash \neg\psi_i \vee \ldots \vee \neg\psi_n). \end{aligned}$$

Similar to the application of δ, the defeat of δ in a set E can best be translated as including

$$L\varphi \wedge (L\neg\psi_1 \vee \ldots \vee L\neg\psi_n) \tag{18}$$

in the smallest stable set containing E; non-defeat of δ in E' can be translated as the absence of (18) in the smallest stable set containing E'. For convenience we denote (18) by $kon_{def}(\delta)$. As before, as a direct translation from default logic to autoepistemic logic, we get that (according to definition 4.2) an (autoepistemic) extension E of $kon(\Delta)$ is $kon(D)_>$-preferred$_d$ iff there is a strict total order $>'$ extending $>$ such that for every extension E' and $\delta \in D$, if $kon_{def}(\delta) \in (E\backslash E')$ then there is a $\delta' \in D$ such that $kon_{def}(\delta') \in (E'\backslash E)$ and $\delta' >' \delta$.

Defintion 4.6 (continued). *Again, by generalization, we get that by replacing (17) by*

$$\forall\varphi \in (E\backslash E') \cap P \; \exists\psi \in (E'\backslash E) \cap P \; (\varphi >' \psi). \tag{19}$$

we obtain the definitions of $P_>$-preferred$_d$ expansion of T and $P_>$-preferred$_{2d}$ expansion of T, respectively.

In [18], the notion we call $P_>$-preferred$_d$ is actually called $P_>$-preferred by Rintanen. It is there explained to be a generalization of preferedness of subtheories in [3]. The notion we call $P_>$-preferred$_{2d}$ is called $P_>$-*maximality* in [18] and is explained to be a generalization of maximality of ordered theory presentations in [20]. We changed the names knowingly so to comply with the analogous definitions in the previous section, which seems only natural to do. The notions $P_>$-preferred and $P_>$-preferred$_2$ have not previously been described in autoepistemic logic. The following theorem summarizes the translation of prioritized default logic to autoepistemic logic.

Theorem 4.7. *Let $\Delta = (D, W)$ be a default theory and $>$ a strict partial order on D. Let $>'$ be the strict partial order on $kon(D)$ such that for all $\delta, \epsilon \in D$, $kon(\delta) >' kon(\epsilon)$ iff $\delta > \epsilon$. Then E is a $D_>$-preferred$_{(2)(d)}$ extension (weak extension) of Δ if and only if E is a $kon(D)_{>'}$-preferred$_{(2)(d)}$ extension (expansion) of $kon(\Delta)$.*

Proof: The theorem holds by the construction of T-preferred$_{(2)(d)}$ extension (expansion) of an ae-theory T. □

Definition 4.6 gives rise to the following preference relations on expansions, which we will further study in section 4.3.3.

Definition 4.8. *Given a total prioritization $P_>$ and two expansions E and E', let*

$$E >_P E' \leftrightarrow (E \cap P) \neq (E' \cap P) \wedge \forall \varphi \in P \cap (E' \backslash E)\ \exists \psi \in P \cap (E \backslash E')\ (\psi > \varphi), \tag{20}$$

$$E >_P^d E' \leftrightarrow (E \cap P) \neq (E' \cap P) \wedge \forall \varphi \in P \cap (E \backslash E')\ \exists \psi \in P \cap (E' \backslash E)\ (\varphi > \psi). \tag{21}$$

It is evident that (20) is derived from (17) and (21) from (19). Note that the symbols $>_P$ and $>_P^d$ are chosen deliberately to reflect both the set P and the relation $>$ of the prioritization $P_>$ (i.e. a prioritization $Q_{>'}$ would lead to preference structures denoted by $>'_Q$ and $>'^d_Q$).

4.3.2 On the term "lexicographic"

Usually the expression "lexicographic order" is used to define an order on strings of elements of some strict total order $\langle P, > \rangle$ as it would be in a dictionary: let $p = (p_1, \ldots, p_k)$ and $q = (q_1, \ldots, q_l)$ be strings of elements from strict total order $\langle P, > \rangle$. Then if $p_i = q_i$ for $i = 1, \ldots, \min\{k, l\}$ then whichever string is the shortest is lexicographically preferred (thus "logic" comes before "logical" in a dictionary). Otherwise, let i be the least such that $p_i \neq q_i$. Then p is lexicographically preferred iff $p_i > q_i$ and q is preferred iff $q_i > p_i$ (thus "logical" comes before "logician").

Example 4.9. If for example $P = \{\varphi_1, \varphi_2, \varphi_3\}$ with $>= \{\varphi_1 > \varphi_2 > \varphi_3\}$ is such a strict total order then

$$\emptyset >_{lex} \{\varphi_1\} >_{lex} \{\varphi_1, \varphi_2\} >_{lex} \{\varphi_1, \varphi_2, \varphi_3\} >_{lex} \{\varphi_1, \varphi_3\} >_{lex} \{\varphi_2\} >_{lex} \{\varphi_2, \varphi_3\} >_{lex} \{\varphi_3\},$$

where $>_{lex}$ stands for lexicographically preferred. (Note though that $\emptyset$ actually does not usually show up in dictionaries.) □

We will explain that the term "lexicographic" is somewhat oddly chosen, since Rintanens different ways of prioritizing are not lexicographical as usual. If we take again the total prioritization $P_> = \{\varphi_1 > \varphi_2 > \varphi_3\}$ and we suppose that all subsets of $P = \{\varphi_1, \varphi_2, \varphi_3\}$ are expansions, then P itself is $>_P$-more preferred than any other expansion, since it satisfies (20) vacuously for all E', and no other expansion satisfies (20) for P. Also, the empty set $\emptyset$ is $>^d_P$-more preferred than any other expansion, since it satisfies (21) for all expansions and no expansion satisfies (21) for the empty set. Further we get that:

$$\{\varphi_1, \varphi_2, \varphi_3\} > \{\varphi_1, \varphi_2\} > \{\varphi_1, \varphi_3\} > \{\varphi_1\} > \{\varphi_2, \varphi_3\} > \{\varphi_2\} > \{\varphi_3\} > \emptyset, \tag{22}$$

$$\emptyset >^d \{\varphi_1\} >^d \{\varphi_2\} >^d \{\varphi_1, \varphi_2\} >^d \{\varphi_3\} >^d \{\varphi_1, \varphi_3\} >^d \{\varphi_2, \varphi_3\} >^d \{\varphi_1, \varphi_2, \varphi_3\}. \tag{23}$$

Neither the order $>$, nor the order $>^d$ is quite the usual lexicographic ordering (none-the-less we will remain using this term for these sorts of orderings). In these examples it becomes visible that the first method is described by "choosing for presence" and the second by "choosing for absence"; the first prefers namely the presence of more-preferred formulas while the second prefers the absence of less-preferred formulas. Note though that for $* \in \{>, >^d, >_{lex}\}$, the following parts are all present:

- $\{\varphi_1\} * \{\varphi_2\} * \{\varphi_3\}$,
- $\{\varphi_1, \varphi_2\} * \{\varphi_1, \varphi_3\} * \{\varphi_2, \varphi_3\}$,
- $\{\varphi_1, \varphi_2\} * \{\varphi_3\}$,
- $\{\varphi_1\} * \{\varphi_2, \varphi_3\}$.

These parts are paradigmatic for the parts that these three lexicographic orders have in common for sets of arbitrary magnitude. Namely, if A and B are sets such that neither $A \backslash B$ nor $B \backslash A$ is empty and each element in $A \backslash B$ is more *-preferred than each element in $B \backslash A$ then $A * B$ for $* \in \{>, >^d, >_{lex}\}$.

4.3.3 Properties of the preference relation

We would like the given relations $>_P$ and $>_P^d$ to in fact describe preference relations, so we would like them to have certain properties. Irreflexivity of the relations are ensured by the clause $(E \cap P) \neq (E' \cap P)$ in definition 4.8. To ensure that the relations are also asymmetric and transitive and consequently represent preference relations, we have to impose some restrictions, for infinite examples can refute these properties. Consider namely the following example.

Example 4.10. Let $P = \{\varphi_n \mid n \in \mathbb{N}\}$ with $\varphi_n < \varphi_{n+1}$ for all n and let

$$\begin{aligned} E_1 &= \{\varphi_3, \varphi_5, \varphi_7, \varphi_9, \varphi_{11}, \ldots\}, \\ E_2 &= \{\varphi_1, \varphi_2, \varphi_4, \varphi_6, \varphi_8, \varphi_{10}, \ldots\}, \\ E_3 &= \{\varphi_1, \varphi_3, \varphi_5, \varphi_7, \varphi_9, \varphi_{11}, \ldots\}. \end{aligned}$$

Then for all $\varphi_n \in E_1 \backslash E_2$ there is a $\varphi_m \in E_2 \backslash E_1$ such that $\varphi_m > \varphi_n$, for example $m = n+1$. Hence $E_2 >_P E_1$. Similarly we get $E_1 >_P E_2$ and so anti-symmetry fails, since $E_1 \neq E_2$. Also, we have $E_1 > E_2$ and $E_2 > E_3$ but $E_1 \not> E_3$, since $\varphi_1 \in E_3 \backslash E_1$, but $E_1 \backslash E_3 = \emptyset$, and thus (20) fails and so transitivity fails. Also (in case of choice for absence in stead of presence), consider the same set P above, but now with $\varphi_n > \varphi_{n+1}$ for all n. Then it's again easy to see that $E_1 >_P^d E_2$ and $E_2 >_P^d E_1$ and so antisymmetry fails. Also, $E_3 >_P^d E_2 >_P^d E_1$ but $E_3 \not>_P^d E_1$ and so transitivity fails. □

We will show that the orders $>_P$ and $>_P^d$ are irreflexive, asymmetric and transitive if there are no infinite ascending or descending chains, respectively. But first we will give an other characterization for the preference relation on sets of formulas that will help in proving the subsequent theorems.

Lemma 4.11. *Let E, E' be sets of formulas and $P_>$ a prioritization such that $E \cap P$ and $E' \cap P$ have no infinite ascending chains. Then $E >_P E'$ if and only if $E \cap P \neq E' \cap P$ and $\max_>(P \cap (E \cup E') \backslash (E \cap E')) \in E$. Also, if $E \cap P$ and $E' \cap P$ have no infinite descending chains, then $E >_P^d E'$ if and only if $E \cap P \neq E' \cap P$ and $\min_>(P \cap (E \cup E') \backslash (E \cap E')) \notin E'$.*

Proof:

"⇒": Assume that $E >_P E'$, i.e. $E \cap P \neq E' \cap P$ and $\forall \varphi \in P \cap (E' \backslash E)\ \exists \psi \in P \cap (E \backslash E')$ such that $\psi > \varphi$. Since $E \cap P \neq E' \cap P$ it follows that $P \cap (E \cup E') \backslash (E \cap E') \neq \emptyset$ and since also $E \cap P$ and $E' \cap P$ contain no infinite ascending chains it follows that $\max_>(P \cap (E \cup E') \backslash (E \cap E'))$ exists and is either contained in $E \backslash E'$ or in $E' \backslash E$. If the maximum is in $E' \backslash E$, by assumption there must be a $\psi \in E \backslash E'$ such that $\psi > \max_>((E \cup E') \backslash (E \cap E'))$, which is a contradiction. It follows that $\max_>((E \cup E') \backslash (E \cap E')) \in E \backslash E'$.

"⇐": Assume that $E \cap P \neq E' \cap P$ and $\max_>(P \cap (E \cup E') \backslash (E \cap E')) \in E$. Then obviously, since $\max_>(P \cap (E \cup E') \backslash (E \cap E')) \notin E'$, for all $\varphi \in P \cap (E' \backslash E)$

the maximum $\max_>(P \cap (E \cup E')\backslash(E \cap E')$ is $>$-larger than φ), thus $E >_P E'$.

The second proof is analogous and will be omitted. □

The following lemma is used in proving proving proposition 4.13.

Lemma 4.12. *For sets of formulas E and E' and prioritization $P_>$,*

$$E >_P E' \text{ iff } E \cap P >_P E' \cap P \text{ and } E >^d_P E' \text{ iff } E \cap P >^d_P E' \cap P.$$

Proof: Immediate from (20) and (21). □

Proposition 4.13. *Let $P_>$ be a prioritization and E_1, E_2 and E_3 sets of formulas such that their intersection with P have no infinite ascending chains. If $E_1 >_P E_2$ then $E_2 \not>_P E_1$. Also, if $E_1 >_P E_2$ and $E_2 >_P E_3$ then $E_1 >_P E_3$. This shows asymmetry and transitivity of $>_P$. Also, $>^d_P$ is asymmetric and transitive.*

Proof: For symmetry, let $E_1 >_P E_2$. Then $E_1 \cap P \neq E_2 \cap P$ and $\max_>(P \cap (E_1 \cup E_2)\backslash(E_1 \cap E_2)) \in E_1$, so obviously $\max_>(P \cap (E_1 \cup E_2)\backslash(E_1 \cap E_2)) \notin E_2$, hence $E_2 \not>_P E_1$. For transitivity, let E_1, E_2 and E_3 be sets of formulas and let $E_1 >_P E_2$ and $E_2 >_P E_3$. Then we need to show that $E_1 >_P E_3$. Given lemma 4.12, we can without loss of generalization assume that $E_i = E_i \cap P$. We denote the seven relevant sets in the proof as follows:

- $E_1\backslash(E_2 \cup E_3) \mapsto (1)$;
- $E_2\backslash(E_1 \cup E_3) \mapsto (2)$;
- $E_3\backslash(E_1 \cup E_2) \mapsto (3)$;
- $(E_1 \cap E_2)\backslash E_3 \mapsto (12)$;
- $(E_1 \cap E_3)\backslash E_2 \mapsto (13)$;
- $(E_2 \cap E_3)\backslash E_1 \mapsto (23)$;
- $E_1 \cap E_2 \cap E_3 \mapsto (123)$.

Suppose for a contradiction that $E_1 \not>_P E_3$, i.e. there is a $\varphi \in E_3\backslash E_1$ such that $\varphi > \psi$ for all $\psi \in E_1\backslash E_3$. Since $E_3\backslash E_1 = (3) \cup (23), \varphi$ is either in (3) or in (23).

Suppose that $\varphi \in (23)$. Since no set has infinite ascending chains by assumption, we may assume that φ is the maximal formula in (23). Since φ is also contained by $E_2\backslash E_1$ and $E_1 >_P E_2$, there must be some $\psi \in E_1\backslash E_2$ such that $\psi > \varphi$. Since $E_1\backslash E_2 = (1) \cup (13)$, either $\psi \in (1)$ or $\psi \in (13)$. However, if ψ would be in (1), then by our assumption that φ is more preferred than all formulas in (1), we would have that $\varphi > \psi > \varphi$, which contradicts the asymmetry of the order. Thus we conclude that ψ must

be in (13). Again we may assume that ψ is the maximal formula in (13). But then, since $\psi \in E_3 \backslash E_2$ and $E_2 >_P E_3$, there must be some $\chi \in E_2 \backslash E_3 = (2) \cup (12)$ such that $\chi > \psi$. But suppose $\chi \in (12)$. Then by our assumption that φ is more preferred than all formulas in (12), we would have that $\chi > \psi > \varphi > \chi$, which by transitivity of the order on formulas contradicts the asymmetry of the same order. Also, suppose $\chi \in (2)$. Then again by $E_1 >_P E_2$ and similar argumentation there must be a ξ in either (1) or in (13) such that $\xi > \chi$. The first fails, since φ was more preferred than all formulas in (1) and hence we would get $\xi > \chi > \psi > \varphi > \xi$. The second fails, because then $\xi > \chi > \psi$, whereas we assumed that ψ was the maximal formula in (13).

Suppose that $\varphi \in (3)$**.** By $E_2 >_P E_3$ there must be a $\psi \in E_2 \backslash E_3 = (2) \cup (12)$ such that $\psi > \varphi$. By similar argumentation as above, ψ cannot be in (12), since then we would have $\varphi > \psi > \varphi$. So assume that ψ is the maximal formula in (2). Then by $E_1 >_P E_2$ there must be a $\chi \in E_1 \backslash E_2 = (1) \cup (13)$ such that $\chi > \psi$. Again, χ cannot be in (1), for then we would have $\chi > \psi > \varphi > \chi$, so assume that χ is the maximal formula in (13). Then since $E_2 >_P E_3$, there must be a $\xi \in E_2 \backslash E_3 = (2) \cup (12)$ such that $\xi > \chi$. However, if $\xi \in (12)$ then $\xi > \chi > \psi > \varphi > \xi$, since φ was supposed to be more preferred than all formulas in (12). Moreover, if $\xi \in (2)$, then $\xi > \chi > \psi$, whereas we assumed that ψ was the maximal formula in (2).

In every case of supposing that $E_1 \not> E_3$, we come to contradiction, hence $E_1 > E_3$. The proofs of asymmetry and transitivity of $>_P^d$ are similar and will be omitted. □

So far, the prioritizations of expansions, being $>_P$ given by (20) and $>_P^d$ given by (21), are defined only with respect to a *total* prioritization $P_>$ on formulas. There are two natural possible ways of extending these definitions to prioritizations with respect to a *partial* prioritization $P_>$. One contains an existential quantifier where the other has a universal one (the reasoning for $>_P^d$ is analogous to that for $>_P$, so we only describe the latter):

1. if $P_>$ is a partial prioritization, let $E >_P E'$ iff there is a strict total order $>'$ on P extending $>$ such that $E >'_P E'$,
2. if $P_>$ is a partial prioritization, let $E >_P E'$ iff for each strict total orders $>'$ on P extending $>$, $E >'_P E'$.

The trouble with the first option is that the desired properties of preference structures no longer hold. Consider namely the following counterexample for asymmetry.

Example 4.14. Let $P_>$ be a partial prioritization consisting of the set $P = \{a, b, c\}$ and the partial preference relation $\{a > c, b > c\}$ and let $E = \{a, c\}$ and $E' = \{b, c\}$. There are two possible total orders $>^1$ and $>^2$ on P extending the partial order, being $\{a >^1 b >^1 c\}$ and $\{b >^2 a >^2 c\}$. It is easy to

see that $E >_P^1 E'$ and $E' >_P^2 E$, hence $E >_P E'$ and $E' >_P E$. Transitivity also fails by the following counterexample: let $P_>$ a the partial prioritization with $P = \{a, b, c, d\}$ and $\{a > b, c > b, d > a\}$. Let $E = \{a, b\}$, $E' = \{c, b\}$ and $E'' = \{d, a, b\}$. Then $E >^1 E'$ with the total order $\{d >^1 a >^1 c >^1 b\}$, $E' >^2 E''$ with $\{c >^2 d >^1 a >^1 b\}$, but for no total order $>^3$ extending $>$ will $E >^3 E''$ hold, since d, the greatest element in $P \cap E''$, will always be $>^3$-greater than a, the greatest element in $P \cap E$. □

For the second definition, with the universal quantifier, asymmetry and transitivity do hold. This is trivially true by the asymmetry and transitivity of $>_P$ when $P_>$ is a total prioritization (as long as we assume that there are no infinite ascending chains in the prioritization).

Definition 4.15. *Let $P_>$ be a prioritization and E, E' expansions of some ae-theory T. Then $E >_P E'$ iff for every total order $>'$ on P extending $>$,*

$$E \cap P \neq E' \cap P \wedge \forall \varphi \in P \cap E' \backslash E \; \exists \psi \in P \cap E \backslash E' \; (\psi >' \varphi).$$

Corollary 4.16. *If $P_>$ is a prioritization then $>_P$ is a asymmetric and transitive relation.*

Proof: Immediate by proposition 4.13. □

As a last remark before the next section, we note that the definitions of the preference structures above need not be restricted merely to expansions or extensions, but rather can be viewed as preference structures on sets of formulas in general.

5 Topics of future research

In addition to the previous 4 chapters, the following topics were originally meant to be treated in this thesis. Due to restraints in time, knowledge, so many things, they are now only treated as yet unsolved questions, or topics of future research.

5.1 Prioritization in terms of possible world structures

Given a preference relation $>$ on a set of defaults D or on a set of formulas P we have defined several lexicographic preference relations on sets of expansions and extensions of a default theory Δ or an ae-theory T. Given the interesting results of Denecker et al. on the relation between expansions (and extensions) and the possible world structures linked to them, a sensible quest is that for a characterization of the lexicographic preference relations in terms of possible world structures. In particular, we would like to find a way of prioritizing possible world structures given the prioritization $P_>$ in such a way that it corresponds to the prioritization of sets of formulas given that same prioritization $P_>$. In other words, we would like to define some relation $\succ_P$ on possible world structures in such a way that if E_i are expansions and Q_i are such that $E_i = Th_L(Q_i)$ then $E_1 >_P E_2$ if and only if $Q_1 \succ_P Q_2$.

Since possible world structures are simply sets of interpretations, like expansions are simply sets of formulas, one way of prioritizing possible world structures would be somehow in terms interpretations, as the sets of formulas are prioritized in terms of the formulas. But also, since the only information given to us in advance of the prioritization of expansions is the prioritization $P_>$, we need to link the prioritization of the possible world structures also to this entity $P_>$. Thus the following diagram arises, which (ideally) would commute:

$$\begin{array}{ccc} \varphi > \psi & \overset{?}{\rightsquigarrow} & I \succ J \\ \downarrow & & \downarrow? \\ E_1 >_P E_2 & \leftrightarrow & Q_1 \succ_P Q_2 \end{array}$$

However, it is not obvious how to "translate" a preference relation on formulas to a preference relation on interpretations. Our proposal is therefore to define, in stead of a relation on interpretations, a relation on sets of interpretations as follows: if for some first order formulas $\varphi > \psi$, then let $Q_\varphi \succ Q_\psi$, where $Q_\chi = \{I \mid I(\chi) = true\}$. For future research, it would be interesting to fill in the last questionmark in the following picture, such that it would commute:

$$\begin{array}{ccc} \varphi > \psi & \leftrightarrow & Q_\varphi \succ Q_\psi \\ \downarrow & & \downarrow? \\ E_1 >_P E_2 & \leftrightarrow & Q_1 \succ_P Q_2 \end{array}$$

As a start, we can in any case say the following:

Lemma 5.1. *If $E = Th_L(Q)$ then $\varphi \in E_0$ if and only if $Q \subseteq Q_\varphi$*

Proof: By definition $E = Th_L(Q) = \{\varphi \mid H_Q(L\varphi) = true\}$, so

$$\begin{aligned} E_0 &= \{\varphi \in \mathcal{L} \mid H_Q(L\varphi) = true\} \\ &= \{\varphi \in \mathcal{L} \mid H_{Q,I}(\varphi) = true \text{ for all } I \in Q\} \\ &= \{\varphi \in \mathcal{L} \mid I(\varphi) = true \text{ for all } I \in Q\} \\ &= \bigcap_{I \in Q} \{\varphi \in \mathcal{L} \mid I(\varphi) = true\} \end{aligned}$$

From the above follows that $\varphi \in E_0$ if and only if for all I in Q $I(\varphi) = true$, i.e. $Q \subseteq \{I \mid I(\varphi) = true\} = Q_\varphi$. □

Given this lemma, we conjecture that the way to define $Q_1 \succ_P Q_2$ in terms of sets of the form Q_φ, given a prioritization $P_>$ on formulas will not look very different from the way $E_1 >_P E_2$ was defined in, for example, definition 4.8.

5.2 The link between preferred extensions and preferred expansions

Another subject to address is the link between the different semantics of prioritized default logic and prioritized autoepistemic logic. From chapter 3 we have a very clear connection between the semantics of default logic and autoepistemic and it would be interesting to investigate this connection in the prioritized versions. What do the analogies to the semantics of partial extensions and partial expansions in terms of formulas and sets of formulas look like and do they relate in the same way to extensions and expansions as in terms of possible world structures? That is, can we define operators on pairs of sets of formulas in such a way that their fixpoints correspond to the partial extensions and partial expansions of default theories or ae-theories? Moreover, do these connections preserve the preferredness of extensions and expansions in some way? In other words, how are the preferred extensions of Δ or T related to the preferred weak extensions of Δ or the preferred expansions of T, respectively.

5.3 The connection with other forms of prioritization

Next to lexicographic prioritization, there are a considerable number of other ways of prioritization, as section 4.1 implies. Notable examples are [4] and [6]. Now, there are always disputes about which method is better than another. One way of doing this is naming examples that are treated "unintuitively" by this method or that, as if there is one intuition that is the best. We chose not to go there in this thesis and investigate just one method, without mentioning the "pros and cons". This is largely due to lack of time, because it is an important issue that should be investigated.

References

[1] Grigoris Antoniou. *Nonmonotonic reasoning.* MIT Press, 1997.

[2] Bronisław Knaster. Un théorème sur les fonctions d'ensembles. *Annales de la Société polonaise de mathématiques*, 6:133–134, 1928.

[3] Gerhard Brewka. Preferred subtheories: an extended logical framework for default reasoning. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, pages 1043–1048, Detroit, 1989.

[4] Gerhard Brewka and Thomas Eiter. Prioritizing default logic. In *Intellectics and Computational Logic*, pages 27–45, 2000.

[5] J. Delgrande, T. Schaub, H. Tompits, and K. Wang. A classification and survey of preference handling approaches in nonmonotonic reasoning. *Computational Intelligence*, 20(2):308–334, 2004.

[6] James P. Delgrande and Torsten Schaub. Expressing preferences in default logic. *Artificial Intelligence*, 123(1-2):41–87, 2000.

[7] Marc Denecker, Victor W. Marek, and Mirosław Truszczyński. Approximations, stable operators, well-founded fixpoints and applications in nonmonotonic reasoning. In Jack Minker, editor, *Logic-Based Artificial Intelligence*, pages 127–142. Kluwer Academic Publishers, 2000.

[8] Marc Denecker, Victor W. Marek, and Mirosław Truszczyński. Uniform semantic treatment of default and autoepistemic logics. In Anthony G. Cohn, Fausto Giunchiglia, and Bart Selman, editors, *KR2000: Principles of Knowledge Representation and Reasoning*, pages 74–84, San Francisco, 2000. Morgan Kaufmann.

[9] Marc Denecker, Victor W. Marek, and Mirosław Truszczyński. Ultimate approximation and its application in nonmonotonic knowledge representation systems. *Information and Computation*, 192(1):84–121, 2004.

[10] M. Gelfond and V. Lifschitz. The stable semantics for logic programs. In R. Kowalski and K. Bowen, editors, *Proceedings of the 5th International Symposium on Logic Programming*, pages 1070–1080. MIT Press, 1988.

[11] Georg Gottlob. Translating default logic into standard autoepistemic logic. *Journal of the Association for Computing Machinery*, 42:711–740, 1995.

[12] Kurt Konolige. On the relation between default and autoepistemic logic. *Artificial Intelligence*, 35:343–382, 1988.

[13] Viktor W. Marek and Mirosław Truszczyński. Relating autoepistemic and default logics. In *Proceedings of the first international conference on Principles of knowledge representation and reasoning*, pages 276–288, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[14] Marvin Minsky. A framework for representing knowledge. 1974.

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

[16] Robert Moore. Possible-world semantics for autoepistemic logic. pages 137–142, San Francisco, CA, USA, 1987. Morgan Kaufmann Publishers Inc.

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

[18] Jussi Rintanen. Prioritized autoepistemic logic. In *JELIA '94: Proceedings of the European Workshop on Logics in Artificial Intelligence*, pages 232–246, London, UK, 1994. Springer-Verlag.

[19] Jussi Rintanen. Lexicographic priorities in default logic. *Artificial Intelligence*, 106(2):221–265, 1998.

[20] Mark Ryan. Representing defaults as sentences with reduced priority. In Bernhard Nebel, Charles Rich, and William Swartout, editors, *KR'92. Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference*, pages 649–660. Morgan Kaufmann, San Mateo, California, 1992.

[21] Alfred Tarski. Lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5(2):285–309, 1955.

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