

Reiter's Default Logic Is a Logic of Autoepistemic Reasoning And a Good One, Too

Marc Denecker

Department of Computer Science
K.U. Leuven
Celestijnenlaan 200A
B-3001 Heverlee, Belgium

Victor W. Marek
Mirosław Truszczyński

Department of Computer Science
University of Kentucky
Lexington, KY 40506-0633, USA

Abstract: A fact apparently not observed earlier in the literature of nonmonotonic reasoning is that Reiter, in his default logic paper, did not directly formalize *informal* defaults. Instead, he translated a default into a certain natural language proposition and provided a formalization of the latter. A few years later, Moore noted that propositions like the one used by Reiter are fundamentally different than defaults and exhibit a certain *autoepistemic* nature. Thus, Reiter had developed his default logic as a formalization of autoepistemic propositions rather than of defaults.

The first goal of this paper is to show that some problems of Reiter's default logic as a formal way to reason about informal defaults are directly attributable to the autoepistemic nature of default logic and to the mismatch between informal defaults and the Reiter's formal defaults, the latter being a formal expression of the autoepistemic propositions Reiter used as a representation of informal defaults.

The second goal of our paper is to compare the work of Reiter and Moore. While each of them attempted to formalize autoepistemic propositions, the modes of reasoning in their respective logics were different. We revisit Moore's and Reiter's intuitions and present them from the perspective of *autotheoremhood*, where theories can include propositions referring to the theory's own theorems. We then discuss the formalization of this perspective in the logics of Moore and Reiter, respectively, using the unifying semantic framework for default and autoepistemic logics that we developed earlier. We argue that Reiter's default logic is a better formalization of Moore's intuitions about autoepistemic propositions than Moore's own autoepistemic logic.

1 Introduction

In this volume we celebrate the publication in 1980 of the special issue of the Artificial Intelligence Journal on Nonmonotonic Reasoning that included three seminal

papers: *Logic for Default Reasoning* by Reiter (1980), *Nonmonotonic Logic I* by McDermott and Doyle (1980), and *Circumscription — a form of nonmonotonic reasoning* by McCarthy (1980). While the roots of the subject go earlier in time, these papers are universally viewed as the main catalysts for the emergence of nonmonotonic reasoning as a distinct field of study. Soon after the papers were published, nonmonotonic reasoning attracted widespread attention of researchers in the area of artificial intelligence, and established itself firmly as an integral sub-area of knowledge representation. Over the years, the appeal of nonmonotonic reasoning went far beyond artificial intelligence, as many of its research challenges raised fundamental questions to philosophers and mathematical logicians, and stirred substantial interest in those communities.

The groundbreaking paper by McCarthy and Hayes (1969) about ten years before had captured the growing concern with the logical representation of *common sense knowledge*. Attention focused on the representation of *defaults*, propositions that are true for most objects, that commonly assume the form “*most A’s are B’s*.”¹ Defaults arise in all applications involving common sense reasoning and require specially tailored forms of reasoning. For instance, a default “*most A’s are B’s*” under suitable circumstances should enable one to infer from the premise “*x is an A*” that “*x is a B*.” This inference is *defeasible*. Its consequent “*x is a B*” may be false even if its premise “*x is an A*” is true. It may have to be withdrawn when new information is obtained. Providing a general, formal, domain independent and elaboration tolerant representation of defaults and an account of what inferences can be *rationally* drawn from them was the artificial intelligence challenge of the time.

The logics proposed by McCarthy, Reiter, and McDermott and Doyle were developed in an attempt to formalize reasoning where defaults are present. They went about it in different ways, however. McCarthy’s *circumscription* extended a set of first-order sentences with a second-order axiom asserting *minimality* of certain predicates, typically of *abnormality predicates* that capture the exceptions to defaults. This reflected the assumption that the world deviates as little as possible from the “normal” state. Circumscription has played a prominent role in nonmonotonic reasoning. In particular, it has been a precursor to preference logics (Shoham, 1987) that provided further important insights into reasoning about defaults.

Reiter (1980) and McDermott and Doyle (1980), on the other hand, focused on the inference pattern “*most A’s are B’s*.” In Reiter’s words (Reiter, 1980, p. 82):

‘We take it [that is, the default “Most birds can fly” — DMT to mean something like “If an x is a bird, then in the absence of any information to the contrary, infer that x can fly.”’

Thus, Reiter (and also McDermott and Doyle) quite literally equated a default “*most A’s are B’s*” with an inference rule that involves, besides the premise “*x is an A*”, an additional premise “there is no information to the contrary” or, more specifically, “*there is no information indicating that “x is not a B.*” The role of this latter premise, a *consistency* condition, is to ensure the rationality of applying the default. In logic, inference rules are meta-logical objects that are not expressed in a logical language.

¹In this paper, we interpret the term “default” as an informal statement “*most A’s are B’s*” (Reiter, 1980). The term is sometimes interpreted more broadly to capture *communication conventions*, *frame axioms* in temporal reasoning, or statements such as “*normally* or *typically*, *A’s are B’s*”.

Reiter, McDermott and Doyle sought to develop a logic in which such meta-logical inference rules could be stated *in* the logic itself. They equipped their logics with a suitable *modal operator* (in the case of Reiter, embedded within “his” default expression) to be able to express the *consistency* condition and, in place of a default “*most A’s are B’s*”, they used the statement “*if x is an A and if it is consistent (with the available information) to assume that x is a B , then x is a B .*” We will call this latter statement the *Reiter-McDermott-Doyle* (RMD, for short) proposition associated with the default.

Moore (1985) was one of the first, if not the first, who realized that defaults and their RMD propositions are of a different nature. This is how Moore (1985, p. 76) formulated RMD propositions in terms of theoremhood and non-theoremhood:

‘[In the approaches of McDermott and Doyle, and of Reiter — DMT] the inference that birds can fly is 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.’

Moore then contended that RMD propositions are *autoepistemic* statements, that is, introspective statements referring to the reasoner’s own belief or the theory’s own theorems. He pointed out fundamental differences between the nature of default propositions and autoepistemic ones and argued that the logics developed by McDermott and Doyle (1980) and, in the follow-up paper, by McDermott (1982), are attempts at a logical formalization of of autoepistemic statements and not of defaults. Not finding the McDermott and Doyle formalisms quite adequate as autoepistemic logics, Moore (1984, 1985) proposed an alternative, the *autoepistemic logic*.

Unfortunately, Moore did not refer to the paper by Reiter (1980) but only to those by McDermott and Doyle (1980) and McDermott (1982), and his comments on this topic were not extrapolated to Reiter’s logic. Neither did Moore explain what could go wrong if a default is replaced by its RMD proposition. Yet, if Moore is right then given the close correspondence between Reiter’s and McDermott and Doyle’s views on defaults, also Reiter’s logic is an attempt at a formalization of autoepistemic rather than of default propositions. Moreover, if defaults are really fundamentally different from autoepistemic propositions, as Moore claimed, it should be possible to find demonstrable defects of Reiter’s default logic for reasoning about defaults that could be attributed to the different nature of a default and of its Reiter’s autoepistemic translation.

Our main objective in Section 2 is to argue that Moore was right. We show there two forms of such defects that (1) the RMD proposition is not always *sound* in the sense that inferences made from it are not always rational with respect to the original defaults, and (2) the RMD proposition is not always *complete*, that is, there are sometimes rational inferences from the original defaults that are not covered by this particular inference rule. In fact, both types of problems can be illustrated with examples long known in the literature.

In the remaining sections, we explain Reiter’s default logic as a formalization of autoepistemic propositions and show that in fact, Reiter’s default logic is a better formalization of Moore’s intuitions than Moore’s own autoepistemic logic. On a formal level, our investigations exploit the results on the unifying semantic framework for default logic and autoepistemic logic that we proposed earlier (Denecker, Marek, and

Truszczyński, 2003). That work was based on an algebraic fixpoint theory for non-monotone operators (Denecker, Marek, and Truszczyński, 2000). We show that the different dialects of autoepistemic reasoning stemming from our informal analysis can be given a principled formalization using these algebraic techniques. In our overview, we will stress the view on autoepistemic logic as a logic of *autotheoremhood*, in which theories can include propositions referring to the theory's own theorems.

Some history. We mentioned that Moore's comments concerning the RMD proposition and the formalisms by McDermott and Doyle (1980) and McDermott (1982) have never been applied to Reiter's logic. For example, Konolige (1988), who was the first to investigate the formal link between autoepistemic reasoning and default logic, wrote that "*the motivation and formal character of these two systems [Reiter's default and Moore's autoepistemic logics – DMT] are different*". This bypasses the fact that Reiter, as we have seen, starts his enterprise of building default logic after translating a default into a proposition which Moore later identified as an autoepistemic proposition.

There may be several reasons why Moore's comments have never been extrapolated to Reiter's logic. As mentioned before, one is that Moore did not refer to the paper by Reiter (1980) but only to the papers by McDermott and Doyle (1980) and McDermott (1982). In addition, the logics of Reiter and, respectively, McDermott and Doyle were quite different; the formal connection was not known at that time (mid 1980s) and was established only about five years later (Truszczyński, 1991). Also autoepistemic and default logics seemed to be quite different (Marek and Truszczyński, 1989), and eventually turned out to be different in a certain precise sense (Gottlob, 1995). Moreover, the intuitions underlying the nonmonotonic logics of the time had not been so clearly articulated, not even in Moore's work as we will see later in the paper, and were not easy to formalize. This was clearly demonstrated about ten years later by Halpern (1997), who reexamined the intuitions presented in the original papers of default logic, autoepistemic logic and Levesque's (1990) related logic of only knowing and showed gaps and ambiguities in these intuitions, and various non-equivalent ways in which they could be formalized.

As a result, the nature of autoepistemic propositions, its relationship to defaults and what may go wrong when the latter are encoded by the first, was never well understood. The relevance of Moore's claims for Reiter's default logic has never become generally acknowledged. Reiter's logic has never been thought of and has never been truly analyzed as a formalization of autoepistemic reasoning. The influence of Reiter's paper has been so large, that even today, the default "*most A's are B's*" and the statement "*if x is an A and if it is consistent to assume that x is a B, then x is a B*"² are still considered synonymous in some parts of the nonmonotonic reasoning community. Yet, in fact, they are quite different and, more importantly, a logical representation of the second is unsatisfactory for reasoning about the first.

²Or its propositional version "*if A and if it is consistent to assume B, then B*".

2 Reiter's Defaults Are Not Defaults But Autoepistemic Statements

Our goal below is to justify the claim in the title of the section. To avoid confusion, we emphasize that by a *default* we mean an informal expression of the type *most A's are B's*. In Reiter's approach (similarly in that of McDermott and Doyle), the default is first translated into an *RMD proposition* *if x is an A and if it is consistent with the available information to assume that x is a B, then x is a B*, which is then expressed by a *Reiter's default expression* in default logic:

$$\frac{A(x) : MB(x)}{B(x)}.$$

To explain the section title, let us assume a setting in which a human expert has knowledge about a domain that consists of propositions and defaults. In the approach of Reiter (the same applies to McDermott and Doyle), the expert builds a knowledge base T by including in T formal representations of the propositions (given as formulas in the language of classical logic) and of RMD propositions of the defaults (given by the corresponding Reiter's default expressions). The presence of Reiter's default expressions in T means that T contains propositions referring to its own information content, i.e., to what is consistent with T , or dually to what T entails or does not entail. Moore (1985) called such reflexive propositions *autoepistemic* and argued that they statements could be phrased in terms of theorems and non-theorems of T .

Reiter developed a default expression as a formal expression of the RMD proposition rather than of the default itself (the same holds for McDermott and Doyle). This is why this logic expression does not capture the full informal content of the default. When considered more closely, it indeed becomes apparent that a default and its RMD proposition are not equivalent or even related in a strict logical sense. A straightforward possible-world analysis reveals this. The default might be true in the actual world (say 95% of the A 's are B 's) but if there is just one x that is an A and not a B , and for which T has no evidence that it is not a B , the RMD proposition is false in this world and x is a witness of this. Thus, it is obvious that in many applications where a default holds, its RMD proposition does not. Conversely, the default might not hold in the actual world (few A 's are in fact B 's) yet the expert knows all x 's that are not B 's, in which case the RMD proposition is true.

A fundamental difference pointed out by Moore between defaults and autoepistemic propositions, is that the latter are naturally *nonmonotonic* but inference rules used for reasoning with them are not *defeasible*. For example, extending the knowledge base T containing an RMD proposition with new information, e.g., that some x is not a B , may indeed have a nonmonotonic effect and delete some previous inferences, e.g., that x is a B . The initial inference of x is a B , resulted in a fact that was false. However, that inference was not defeasible. The essential property of a defeasible inference is that it may derive a false conclusion from premises that are true in the actual world. For instance, the inference from *most A's are B's* and x is an A that x is a B is defeasible as its consequent may be false while the premises are true. In the context of our example above the theory, say T , entailed the false fact that x is a B from the premises (i) the RMD proposition, (ii) x is an A and (iii) T contained no evidence that x is not a B . It was not defeasible since one of its premises was false. Indeed, the

RMD proposition was false and x was a witness. The inference rules applied are not defeasible (they are, essentially, the introduction of conjunction and modus ponens). To sum up, an inference from a knowledge base involving an RMD proposition may be false but only if the RMD proposition itself is false.

To emphasize further consequences of equating defaults and RMD propositions we will look at well-known examples from the literature. First, we turn our attention to the question whether there are cases when applying the RMD proposition leads to inferences that do not seem rational (lack of “soundness” with respect to understood informally “rationality”). The Nixon Diamond example by Reiter and Criscuolo (1981) and reasoning problems with related inheritance networks illustrate the problems that arise.

Example 1 Richard M. Nixon, the 37th president of the United States, was a Republican and a Quaker. Most Republicans are hawks while most Quakers are doves (pacifists). Nobody is a hawk and a dove. Some people are neither hawks nor doves. Encoding the Reiter-McDermott-Doyle proposition of these defaults in default logic, we obtain the following theory:

$$\begin{array}{c} \text{Republican}(\text{Nixon}) \wedge \text{Quaker}(\text{Nixon}) \\ \forall x(\neg \text{Dove}(x) \vee \neg \text{Quaker}(x)) \\ \hline \frac{\text{Republican}(x) : M \text{Hawk}(x)}{\text{Hawk}(x)} \quad \frac{\text{Quaker}(x) : M \text{Dove}(x)}{\text{Dove}(x)} \end{array} .$$

In default logic, this theory gives rise to two extensions. In one of them Nixon is believed to be a hawk and not a dove, in the other one, a dove and not a hawk. But is this rational? As we mentioned above, the use of an RMD-proposition is rational when it is expected to hold for most x , and hence, in absence of information, it is likely to hold for some specific x . But in the case of Nixon, we know in advance that at least one of the two “Nixon” instances of the RMD propositions has to be wrong. As to which one is wrong, without further information one could as well throw a coin. Moreover, it is not unlikely that they are both wrong and that in fact, Nixon is neither dove nor hawk. And in fact, it seems more rational not to apply any of the defaults, leading to a situation where it is not known whether Nixon is a dove, a hawk or neither. The rationale of using the RMD proposition as a substitute for the default does not hold for Nixon or any other republican quaker for that matter. \square

Example 2 Let us assume now that all quakers are republicans. In this case, the default that most quakers (say 95%) are doves is more specific than and overrules the default that most republicans (say 95%) are hawks. It is rational here to give priority to the quaker default, leading to the defeasible conclusion that Nixon is a dove. However, this conclusion cannot be derived from the RMD propositions because their consistency premise “*it is consistent to assume that x is a dove (respectively a hawk)*” is too general to take such information into account. \square

Such scenarios were studied in the context of inheritance hierarchies (Touretzky, 1986). To reason correctly on this sort of applications using Reiter’s logic, the consistency condition of the RMD propositions has to be tweaked to take the hierarchy into account and give priority to the quaker default. For example, we can reformulate the

RMD proposition of the default “most republicans are hawks” as “if x is known to be a republican and it is consistent to assume that he is a hawk and it is consistent to assume that he is not a quaker, then x is a hawk”, which takes additional information into account. Such modified rules can of course be represented in default logic. After all, the logic was developed for representing (defeasible) inference rules. But, as in the examples above, they cannot be *inferred* from the RMD propositions. And the inferences that can be drawn from the RMD propositions are not always the rational ones.

The next problem that arises is of a complementary nature and concerns (lack of) completeness with respect to “rational” inferences. Are there cases where rational albeit defeasible inferences can be drawn from defaults that cannot be inferred from RMD propositions? As suggested above by our general discussion, the answer is indeed positive. After all, the RMD proposition expresses only a single and quite specific type of inference that might be associated with a default.

Example 3 As an illustration, let us consider the defaults *most Swedes are blond* and *most Japanese have black hair*. Nobody is both Swede and Japanese, or has both blond and black hair. If we learn know that Boris is a Swede or a Japanese then, given that he cannot be both Swede and Japanese, it seems rational to conclude defeasibly that Boris’s hair is blond or black. In other words, defaults can (sometimes) be combined and together give rise to defeasible inference rules like:

$$\frac{\text{Boris is Swede or Japanese} : M \text{ Boris's hair is blond or black}}{\text{Boris's hair is blond or black}} .$$

If all we know is that Boris is Swede or Japanese, the conclusion of this rule cannot be drawn from the two original RMD propositions for the simple reason that for each, one of their premises is not satisfied: it is not known that Boris is a Swede, and neither is it known that he is Japanese. For instance, in the logic of Reiter, the two defaults would be encoded as

$$\frac{\text{Swede}(x) : M \text{ Blond}(x)}{\text{Blond}(x)} \quad \text{and} \quad \frac{\text{Japanese}(x) : M \text{ Black}(x)}{\text{Black}(x)} .$$

If we only know $\text{Swede}(\text{Boris}) \vee \text{Japanese}(\text{Boris})$, then neither $\text{Swede}(\text{Boris})$ nor $\text{Japanese}(\text{Boris})$ can be established. Therefore, the premises of neither rule are established and no inference can be made. Even more, if we accept Reiter’s logic as a logic of autoepistemic propositions, these conclusions *should not be drawn* from these expressions. \square

This example shows a clear case of a desired defeasible inference that cannot be drawn from the rules expressed in the two RMD propositions. A default expression in Reiter’s logic that would do the job has to encode explicitly the combined inference rule:

$$\frac{\text{Swede}(x) \vee \text{Japanese}(x) : M(\text{Blond}(x) \vee \text{Black}(x))}{\text{Blond}(x) \vee \text{Black}(x)} .$$

This expresses an inference rule which is not derivable from the original RMD propositions in the logics of Reiter, McDermott, Doyle, or Moore. Default logic does not support such reasoning unless the combined inference rule is explicitly encoded as well.

Example 4 Assume that we now find out that Boris has black hair. Given that he is Japanese or Swede, and given the defaults for both, it seems rational to assume that he is Japanese. Can we infer this from the combined inference rules expressed above and given that nobody can be blond and black, or Swede and Japanese? The answer is no and, consequently, yet another inference rule should be added to obtain this inference. \square

Problems of these kind were reported many times in the NMR literature and prompted attempts to “improve” Reiter’s default logic so as to capture additional defeasible inferences of the informal default. This is, however, a difficult enterprise, as it starts from a logic whose semantical apparatus is developed for a very specific form of reasoning, namely autoepistemic reasoning. And while at the formal level the resulting logics (Brewka, 1991; Schaub, 1992; Lukaszewicz, 1988; Mikitiuk and Truszczyński, 1995) capture some aspects of defaults that Reiter’s logic does not, also they formalize a small fragment only of what a default represents and, certainly, none has evolved into a method of reasoning about defaults. In the same time, theories in these logics entail formulas that cannot be justified from the point of view of default logic as an autoepistemic logic.

To summarize, an RMD proposition expresses one defeasible inference rule associated with a default. It often derives rational assumptions from the default but not always, and it may easily miss some useful and natural defeasible inferences. The RMD proposition is autoepistemic in nature; Reiter’s original default logic is therefore a formalism for autoepistemic reasoning. As a logic in which inference rules can be expressed, default logic is quite useful for reasoning on defaults. The price to be paid is that the human expert is responsible for expressing the desired defeasible inference rules stemming from the defaults and for fine-tuning the consistency conditions of the inference rules in case of conflicting defaults. This may require substantial effort and leads to a methodology that is not elaboration tolerant.

While our discussion shows that in general, RMD propositions and Reiter’s defaults do not align well with the informal concept of a default of the form *most A’s are B’s*, there are other nonmonotonic reasoning patterns that are correctly expressed through Reiter’s defaults. In particular, patterns such as communication conventions, database or information storage conventions and policy rules in the typology of McCarthy (1986), can be expressed well by *true* autoepistemic propositions and, consequently, are correctly formalized in Reiter’s logic. E.g., the convention that an airport customs database explicitly contains the nationality of only non-American passengers, is correctly specified by the Reiter default

$$\frac{: MNationality(x) = USA}{Nationality(x) = USA}.$$

Similarly, the policy rule that the departmental meetings are normally held on Wednesdays at noon, is correctly formalized by

$$\frac{: MTime(meeting) = "Wed, noon"}{Time(meeting) = "Wed, noon"}.$$

In spite of such examples, the fact remains that default logic is not a logic of defaults. Are there other logics that could be regarded as such? There have been several interesting attempts at formalizing defaults *most A’s are B’s*. Most important of

them focused on defaults as *conditional assertions* and on abstract nonmonotonic consequence relations (Makinson, 1989; Lehmann, 1989; Pearl, 1990; Kraus, Lehmann, and Magidor, 1990; Lehmann and Magidor, 1992). This research direction resulted in elegant mathematical theories and deep insights into the nature of some forms of nonmonotonic reasoning. However, it is not directly related to our effort here. Thus, rather than to discuss it we refer to the papers we cited.

Instead, in the remainder of the paper, we focus on the second objective identified in the introduction. That is, we provide an informal basis to autoepistemic reasoning, we place Reiter's default logic firmly among dialects of autoepistemic reasoning, and show that Reiter's logic was a watershed point that pinpointed one of the most fundamental and most important forms of autoepistemic reasoning.

3 Studies of Relationships Between Default Logic and Autoepistemic Logic

Konolige (1988) was the first to investigate a formal link between default and autoepistemic logic. He proposed the following translation *Kon* from default logic to autoepistemic logic:

$$\frac{\alpha : M\beta_1, \dots, \beta_n}{\gamma} \quad \mapsto \quad K\alpha \wedge \neg K\neg\beta_1 \wedge \dots \wedge \neg K\neg\beta_n \rightarrow \gamma$$

and argued that *Kon* was equivalence preserving in the sense that default extensions of the default theory were exactly the autoepistemic expansions of its translation. This translation is intuitively appealing, essentially expressing formally the RMD proposition of the default in modal logic, and it indeed plays an important role in the story. Nevertheless, it turned out that this translation was only partially correct (Konolige, 1989). Later, Gottlob (1995) presented a correct translation from default logic to autoepistemic logic but also proved that no *modular* translation exists. The latter result showed that these two logics are essentially different in some important aspect. As a result, the autoepistemic nature of default logic, which Moore had implicitly pointed at, and his implicit criticism on default logic as a logic of defaults were never widely acknowledged.

But Reiter's logic is just that — a logic of autoepistemic reasoning. Moreover, in many respects it is a better logic of autoepistemic reasoning than the one by Moore. Our goal now is to reconsider the intuitions of autoepistemic reasoning, to distinguish between different dialects of it and to develop principled formalizations for these dialects. In particular, we relate Reiter's and Moore's logics, and explain in what sense Reiter's logic is better than Moore's. Our discussion uses the formal results we developed in an earlier paper (Denecker et al., 2003). There we used the algebraic fixpoint theory for arbitrary lattice operators (Denecker et al., 2000) to define four different semantics of default logic and of autoepistemic logic. This theory can be summarized as follows.

A complete lattice $\langle L, \leq \rangle$ induces a complete bilattice $\langle L^2, \leq_p \rangle$, where \leq_p is the precision order on L^2 defined as follows: $(x, y) \leq_p (u, v)$ if $x \leq u$ and $v \leq y$. Tuples (x, x) are called exact. For any \leq_p -monotone operator $A : L^2 \rightarrow L^2$ that is *symmetric*, that is, $A(x, y) = (u, v)$ if and only if $A(y, x) = (v, u)$, we can define

$A : L^2 \rightarrow L^2$		Kripke-Kleene least fixpoint
$O_A : L \rightarrow L$	$O_A(x) = A_1(x, x)$	Supported fixpoints
$S_A : L \rightarrow L$	$S_A(x) = \text{lfp}(A_1(\cdot, x))$	Stable fixpoints
$\mathcal{S}_A : L^2 \rightarrow L^2$	$\mathcal{S}_A(x, y) = (S_A(y), S_A(x))$	Well-founded least fixpoint

Table 1: Lattice operators and the corresponding semantics

three derived operators. These four operators identify four different types of fixpoints or least fixpoints (when the derived operator is monotone). They are summarized in Table 1 (where the operator $A_1(\cdot, \cdot)$ used to define O_A is the projection of A on the first coordinate).

By assumption, A is a \leq_p -monotone operator on L^2 and its \leq_p -least fixpoint is called the *Kripke-Kleene* fixpoint of A . Fixpoints of the operator O_A correspond to exact fixpoints of A (x is a fixpoint of O_A if and only if (x, x) is a fixpoint of A) and are called *supported* fixpoints of A . The operator S_A is an anti-monotone operator on L . Its fixpoints yield exact fixpoints of A (if x is a fixpoint of S_A then (x, x) is a fixpoint of A). They are called *stable* fixpoints of the operator A . It is clear that stable fixpoints are supported. The operator \mathcal{S}_A is a \leq_p -monotone operator on L^2 and its \leq_p -least fixpoint is called the *well-founded* fixpoint of A (fixpoints of \mathcal{S}_A are also fixpoints of A). The names of these fixpoints reflect the well-known semantics of logic programming, where they were first studied by means of operators on lattices. Taking Fitting's four-valued immediate consequence operator (Fitting, 1985) for A , we proved (Denecker et al., 2000) that the four different types of fixpoint correspond to four well-known semantics of logic programming: Kripke-Kleene semantics (Fitting, 1985), supported model semantics (Clark, 1978), stable semantics Gelfond and Lifschitz (1988) and well-founded semantics (Van Gelder et al., 1991).

This elegant picture extends to default logic and autoepistemic logic Denecker et al. (2003). In that paper, we identified the semantic operator \mathcal{E}_Δ for a default theory Δ , and the semantic operator \mathcal{D}_T for an autoepistemic theory T . Both operators were defined on the bilattice of possible-world sets, which we introduce formally in the following section. Just as for logic programming, each operator determines three derived operators and so, for each logic we obtain four types of fixpoints, each inducing a semantics. Some of these semantics turned out to correspond to semantics proposed earlier; other semantics were new. Importantly, it turned out that the operators \mathcal{E}_Δ and $\mathcal{D}_{Kon(\Delta)}$ are identical. Hence, Konolige's mapping turned out to be equivalence preserving for *each* of the four types of semantics! Table 2 summarizes the results. The first two lines align the theories and the corresponding operators. The last four lines describe the matching semantics (the new semantics for autoepistemic and default logics obtained from this operator-based approach Denecker et al. (2000) are in bold font).

From this purely mathematical point of view Konolige's intuition seems basically right. His mapping failed to establish a correspondence between Reiter extensions and Moore expansions *only* because they are on different levels in the hierarchy of the semantics. Once we correctly align the dialects, his transformation works perfectly. Conversely, we also proved that the standard method to eliminate nested modalities in the modal logic S5 can be used to translate any autoepistemic logic theory T into a

default theory Δ	\xrightarrow{Kon}	autoepistemic theory T
semantic operator \mathcal{E}_Δ	\xrightarrow{Kon}	semantic operator \mathcal{D}_T
KK-extension	\xrightarrow{Kon}	KK-extension (Denecker et al., 1998)
Weak extensions (Marek and Truszczyński, 1989)	\xrightarrow{Kon}	Moore expansions (Moore, 1984)
Reiter extensions (Reiter, 1980)	\xrightarrow{Kon}	Stable extensions
Well-founded extension (Baral and Subrahmanian, 1991)	\xrightarrow{Kon}	Well-founded extension

Table 2: The alignment of default and autoepistemic logics

default theory that is equivalent to T under each of the four semantics.

While the non-modularity result by Gottlob (1995) had shown that default logic and autoepistemic logic are essentially different logics, our results summarized above unmistakably point out that default and autoepistemic logics are tightly connected logical systems. They suggest that the four semantics formalize different *dialects* of autoepistemic reasoning and that Reiter and Moore formalized different dialects. Therefore, in the rest of the paper, we will view Reiter's logic simply as a fragment of modal logic, as identified by Konolige's mapping.

4 Formalizing Autoepistemic Reasoning — an Informal Perspective

In our paper (Denecker et al., 2003) we developed a purely algebraic, abstract study of semantics. The study identified the (nonmonotone) operators of autoepistemic and default logic theories, and applied the different notions of fixpoints to them. What that paper was missing was an account of what these fixpoint constructions mean at the informal level and how the different dialects in the framework differ. Being as clear as possible about the informal semantics of autoepistemic theories is essential, as it is there where problems with formal accounts start.

This is the gap that we close in the rest of this paper. To this end we first return to the original concern of Reiter, and of McDermott and Doyle. Let us suppose that we have incomplete knowledge about the actual world, represented in, say, a first order theory T , and that we know that most A 's are B 's. Following the Reiter, McDermott and Doyle approach, we would like to assert the following proposition:

If for some x , $T \models A(x)$ and $B(x)$ is consistent with T (that is, $T \not\models \neg B(x)$), then $B(x)$.

In fact, we would like to express this statement *in* the logic and, moreover, to add this proposition, with its references to what T entails or does not entail, to T itself. What we obtain is a theory T that refers to its own theorems. In this view then,

modal literals $K\varphi$ in an autoepistemic theory $T = \{\dots F[K\varphi] \dots\}$ are to be interpreted *informally* as statements $T \models \varphi$, and the theory T itself as having the form $T = \{\dots F[T \models \varphi] \dots\}$, emphasizing the intuition of the self-referential nature of autoepistemic theories.

This view reflects what seems to us the most precise intuition that Moore proposed: to view autoepistemic propositions as inference rules. Specializing the discussion above to the autoepistemic formula

$$K\alpha_1 \wedge \dots \wedge K\alpha_n \wedge \neg K\beta_1 \wedge \dots \wedge \neg K\beta_m \rightarrow \gamma \quad (1)$$

we can write it (informally) as:

$$T \models \alpha_1 \wedge \dots \wedge T \models \alpha_n \wedge T \not\models \beta_1 \wedge \dots \wedge T \not\models \beta_m \rightarrow \gamma,$$

and understand it (informally) as an inference rule:

$$\begin{array}{l} \text{if } \alpha_1, \dots, \alpha_n \text{ are theorems and } \beta_1, \dots, \beta_m \text{ are not theorems} \\ \text{then } \gamma \text{ holds.} \end{array} \quad (2)$$

which is consistent with Moore's (1985, p. 76) position we cited earlier. Alternatively, $K\varphi$ can be read as " φ can be derived, or proven" (again, from the theory itself), which amounts at the informal level just to a different wording. We will refer to this notion of theorem and derivation as *autotheorem* and *autoderivation*, respectively. Accordingly, we will call the basic Moore's perspective as that of *autotheoremhood*.

The autotheoremhood view can be seen as a special case of a more generic view, also proposed by Moore, based on *autoepistemic agents*. In this view which, incidentally, is the reason behind the name *autoepistemic logic*, an autoepistemic theory is seen as a set of introspective propositions, believed by the agent, about the actual world and his own beliefs about it. The crucial assumption is the one which Levesque (1990) dubbed later the *All I Know* Assumption: the assumption that all that is known by the agent is *grounded* in his theory, in the sense that it belongs to it or can be derived from it. In the case of the autotheoremhood view, the agent is nothing else than a personification of the theory itself, and what it knows is what it entails. We discuss alternative instances of this agent-based view in the next section.

But let us now focus on developing the autotheoremhood perspective. We regard it as a more precise intuition that is more amenable to formalization despite the fact that self-reference, which is evidently present in the notion of autotheoremhood, is a notoriously complex phenomenon. It plagued, albeit in a different form, the *theory of truth* in philosophical logic with millennia-old paradoxes (Tarski, 1983; Kripke, 1975; Barwise and Etchemendy, 1987). The best known example is the famous *liar* paradox:

“This sentence is false.”

An autoepistemic theory that is clearly reminiscent of this paradox is:

$$T_{liar} = \{\neg KP \rightarrow P\}.$$

In the autotheoremhood view, this theory states that if *it* does not entail P then P holds. However, if P is not entailed, then we *have* an argument for P , and if P is entailed, the unique proposition of the theory is trivially satisfied; no argument for

P can be constructed. This is *mutatis mutandis* the argument for the inconsistency of the liar sentence. In view of the difficulties that self-reference has posed to the development of the theory of truth, it would be naive to hope that a crisp, unequivocal formalization of autoepistemic logic existed.

Moore (1985, p. 82) explained the difficulty of defining the semantics for autoepistemic inference rules (2) as follows. When the inference rules are monotonic, that is, when $m = 0$,

‘once a formula has been generated at a given stage, it remains in the generated set of formulas at every subsequent stage. [...] The problem with attempting to follow this pattern with nonmonotonic inference rules [that is, when $m > 0$ (note of the authors)] is that we cannot draw nonmonotonic inferences reliably at any particular stage, since something inferred at a later stage may invalidate them.’

To put it differently, the problem is that when a rule (2) is applied to derive γ at some stage when all α_i 's have been inferred to be theorems and none of the β_j 's has been derived, later inferences may derive some β_j and hence, invalidate the derivation of γ . In such case, Moore argues, all we can do is to characterize the desired result as the solution of a *fixpoint equation* instead of computing it by a *fixpoint construction*:

‘Lacking such an iterative structure, nonmonotonic systems often use nonconstructive “fixed point” definitions, which do not directly yield algorithms for enumerating the “derivable” formulas, but do define sets of formulas that respect the intent of the nonmonotonic inference rules.’

This was an extremely clear and compelling representation of intuitions behind not only the Moore's own autoepistemic logic, but also behind the formalisms of McDermott and Doyle, and of Reiter, too, for that matter.

It is useful now to look at these ideas from a more formal point of view. Let us consider a modal theory T over some vocabulary Σ . Let T consist of “inference rules” of the form (2), where for simplicity we assume that all formulas $\alpha_i, \beta_j, \gamma$ are objective (that is, contain no modal operator).³ The inference processes that Moore had in mind are syntactic in nature and are derivations of formulas. Yet, it is straightforward to cast these inference processes in semantical terms.

Let \mathcal{W} be the set of all Σ -interpretations. A state of belief is represented as a set $B \subseteq \mathcal{W}$ of possible worlds.⁴ Intuitively, each element $w \in B$ represents a *possible world*, a state of affairs that satisfies the agent's beliefs. A world $w \notin B$ represents an *impossible world*, a state of affairs that violates at least one proposition of the agent. Given a set B representing the worlds held possible by an agent, the following, standard, definition formalizes which (modal) formulas the agent believes (or knows — we do not distinguish between these two modalities in our discussion).

Definition 1 We define the satisfiability relation $B, w \models \varphi$ as in the modal logic *S5* by the standard recursive rules of propositional satisfaction augmented with one additional rule:

$$B, w \models K\varphi \text{ if for every } v \in B, B, v \models \varphi.$$

³Our approach works equally well for arbitrary modal theories.

⁴A *possible-world set* is a special Kripke structure in which the accessibility relation is total.

We then define $B \models K\varphi$ (φ is believed or known in state B) if for every $w \in B$, $B, w \models \varphi$.

This definition extends the standard definition of truth in the sense that if φ is an objective formula then $B, w \models \varphi$ if and only if $w \models \varphi$. We define $Th(B) = \{\varphi \mid B \models K\varphi\}$ and $Th_{obj}(B)$ the restriction of $Th(B)$ to objective formulas. These sets represent all modal formulas and all objective formulas, respectively, known in the state of belief B .

It is natural to order belief states according to “how much” they believe or know. For two belief states B_1 and B_2 , we define $B_1 \leq_k B_2$ if $Th_{obj}(B_1) \subseteq Th_{obj}(B_2)$ or, equivalently, if $B_2 \subseteq B_1$. The ordering \leq_k is often called the *knowledge ordering*. We observe that $B_1 \leq_k B_2$ does not entail $Th(B_1) \subseteq Th(B_2)$, due to the nonmonotonicity of modal literals $\neg K\varphi$ expressing ignorance, some of which may be true in B_1 and false in B_2 .

We can see Moore’s inference processes as sequences $(B_i)_{i=0}^\lambda$ of possible-world sets such that $B_0 = \mathcal{W}$, the possible-world set of maximum ignorance in which only tautologies are known. In each derivation step $B_i \rightarrow B_{i+1}$, some worlds $w \in B_i$ might be found to be impossible and eliminated in B_{i+1} ; other worlds $w \notin B_i$ might be established to be possible and added to B_{i+1} . This process is described through Moore’s semantic operator D_T , which maps a possible-world set B to the possible-world set $\{w \mid B, w \models T\}$. For theories consisting of formulas (1), $D_T(B_i)$ is exactly the set of all possible worlds that satisfy the conclusions γ of all inference rules that are “active” in B_i , that is, for which $B_i \models K\alpha_j$, $1 \leq j \leq n$, and $B_i \not\models K\beta_j$, $1 \leq j \leq m$.

Let us come back to Moore’s claims. The nonmonotonicity of the inference rules (2), or more precisely, formulas (1) is due to the negative conditions $\neg K\beta_j$ (β_j not known, not proved, not a theorem). So let us assume that $m = 0$ for all inference rules in T .⁵ One can show that under this assumption D_T is a monotone operator with respect to \subseteq : if $B_1 \subseteq B_2$, then $D_T(B_1) \subseteq D_T(B_2)$. This can be rephrased in terms of knowledge ordering: if $B_1 \leq_k B_2$, then $D_T(B_1) \leq_k D_T(B_2)$. In other words, the operator D_T is also monotone in terms of the knowledge ordering \leq_k . Moore’s inference process $(B_i)_{i=0}^\lambda$ is now an *increasing* sequence in the knowledge order \leq_k . It yields a least fixpoint B_T in the knowledge order (equivalently, the greatest fixpoint of D_T in the subset order \subseteq). Every other fixpoint of D_T contains more knowledge than B_T . The fixpoint B_T is the intended belief state associated with the theory T of monotonic inference rules.

In the general case of nonmonotonic inference rules ($m > 0$, for some rules), the operator D_T may not be monotone. The inference process constructed with D_T may oscillate and never reach a fixpoint, or may reach an unintended fixpoint due to the fact that it may derive that a world is impossible on the basis of an assumption $\neg K\beta_i$ which is later withdrawn. In such case, stated Moore, all we can do is to focus on possible-world sets that “respect the intent of the nonmonotonic inference rules” as expressed by a *fixpoint equation* associated to T , rather than being the result of a *fixpoint construction*. In this way Moore arrived at his semantics of autoepistemic logic, summarized in the following definition.

⁵For arbitrary theories T , the corresponding assumption is that there are no modal literals $K\varphi$ occurring positively in T .

Definition 2 An autoepistemic expansion of a modal theory T over Σ is a possible-world set $B \subseteq \mathcal{W}$ such that $B = D_T(B)$.

We agree with Moore that the condition of being a fixpoint of D_T is a necessary condition for a belief state to be a possible-world model of T . However, it is obviously not a sufficient one, at least not in the autotheoremhood view on T . This is obvious, as this semantics does not coincide with Moore's own ideas on the semantics of monotonic inference rules. A counterexample is the following theory:

$$T = \{KP \rightarrow P\}.$$

This theory consists of a unique monotonic inference rule, albeit a rather useless one as it says "if P is a theorem then P holds". According to Moore's account of monotonic inference rules, the intended possible-world model of this theory is $\mathcal{W} = \{\emptyset, \{P\}\}$ (we assume that $\Sigma = \{P\}$). Yet, T has two autoepistemic expansions, the second being the self-supported possible-world set $\{\{P\}\}$.

It is worth noting that this theory is related to yet another famous problematic statement in the theory of truth, namely the *truth sayers*:

"This sentence is true."

The truth value of this statement can be consistently assumed to be true, or equally well, to be false. Therefore, in Kripke's (1975) three-valued truth theory, the truth value of the truth sayers is *undetermined* **u**. In case of the related autoepistemic theory $\{KP \rightarrow P\}$, also Moore's semantics does not determine whether P is known or not. But in the autotheoremhood view, it is clear that P should not be known and this transpires from Moore's own explanations on monotonic inference rules.⁶ We come back to the issue of self-supported expansions in Section 5, where we explore alternative perspectives on autoepistemic propositions, in which such self-supported expansions might be acceptable.

The main question then is: Can we improve Moore's method to build inference processes in the presence of nonmonotonic inference rules in T ? In this respect, the situation has changed since 1984. The algebraic fixpoint theory for nonmonotone lattice operators (Denecker et al., 2000), which we developed and then used to build the unifying semantic framework for default and autoepistemic logics (Denecker et al., 2003), gives us new tools for defining fixpoint constructions and fixpoint equations which can be applied to Moore's problem.

We illustrate now these tools in an informal way and refer to these intuitions later when we introduce major concepts for a formal treatment. Let us consider the theory:

$$T = \{P, \neg KP \rightarrow Q, KQ \rightarrow Q\}.$$

Informally, the theory expresses that P holds, that if P is not a theorem then Q holds, and that if Q is a theorem, then Q holds. Intuitively, it is clear what the model of this theory should be: P is a theorem, hence the second formula cannot be used to derive

⁶There does not seem to be an analogous strong argument why the truth sayers sentence should be false. Yet, Fitting (1997) proposed a refinement of Kripke's theory of truth in which truth is minimized and the truth sayers statement is *false*. For this, he used the same well-founded fixpoint construction that we will use below to obtain a semantics that minimizes knowledge for autotheoremhood theories.

Q and neither can the truth sayer proposition $KQ \rightarrow Q$. Therefore, the intended possible-world set is $B_T = \{\{P\}, \{P, Q\}\}$, that is, P is entailed, Q is unknown.

It is easily verified that B_T is a fixpoint of D_T . Yet D_T has a second, unintended fixpoint $\{\{P, Q\}\}$ which contains more knowledge than B_T . This is a problem as it is this unintended fixpoint that is obtained by iterating D_T starting with \mathcal{W} . The reason for this mistake is that the second, nonmonotonic inference rule applies in the initial stage $B_0 = \mathcal{W}$ when $\neg KP$ holds. Later, when P is derived, the conclusion that Q is a theorem continues to reproduce itself through the third truth sayer rule.

The problem above is that at each step and for each world w an *assumption* is made of whether w is possible or impossible. Each such an assumption might be right or wrong. These assumptions are revised by iterated application of D_T . In the context of monotonic inference rules, the only wrong assumptions that might be made during the monotonic fixpoint construction starting in \mathcal{W} are that some world is possible, while in fact it turns out to be impossible. But these wrong assumptions can never lead to an erroneous application of an inference rule: if a condition $K\varphi$ of an inference rule holds when w is assumed to be possible, then it will still hold when w turns out to be impossible. But in the context of nonmonotonic inference rules, an inference rule may fire due to an erroneous assumption and its conclusion might be maintained through a circular argument in all later iterations. In our scenario, it is the initial assumption that worlds in which P is false are possible that lead to the assumption that worlds in which Q is false are impossible, and this assumption is later reproduced by a circular reasoning using the third truth sayer proposition for Q .

The solution to this problem is very simple: *never make any unjustified assumption about the status of a world*. Start without any assumption about the status of any worlds and only assign a specific status when certain. We will elaborate this idea in two steps. In the first step, we illustrate this idea for a simplification T' of T , in which the third axiom $KQ \rightarrow Q$ has been deleted.

1. Initially, no world is known to be possible or impossible. At this stage, the truth value of the unique modal literal KP in T' cannot be established. Yet, some things are clear. First, all worlds in which P is false, that is, \emptyset and $\{Q\}$, are certainly impossible since they violate the first formula, P , of T' . Second, the world $\{P, Q\}$ is definitely possible since no matter whether P is a theorem or not, this world satisfies the two formulas of T' . All this can be established without making a single unsafe assumption. Thus, the only world about which we are uncertain at this stage is the world $\{P\}$ in which Q is false. Due to the second axiom, this world is possible if P is known and impossible otherwise.

2. In the next pass, we first use the knowledge that we gained in the previous step to re-evaluate the modal literal KP . In particular, it can be seen that P is true in all possible worlds and in the last remaining world of unknown status, $\{P\}$. This suffices to establish that P is a theorem, that is, that KP is true.

With this newly gained information, we can establish the status of the last world and see that $\{P\}$ satisfies the two axioms of T . Hence this world is possible.

The construction stops here. The next pass will not change anything, and we obtain the possible-world set $B_{T'} = \{\{P\}, \{P, Q\}\}$. Now, let us add the third axiom $KQ \rightarrow Q$ back and consider the full theory T .

1. The first step of the construction is identical to the one above and determines the status for all worlds except $\{P\}$: $\{P, Q\}$ is possible, and \emptyset and $\{Q\}$ are impossible.
2. As before, in the second pass, KP can be established to be true. The second modal literal KQ in T cannot be established yet since its truth depends on the status of the world $\{P\}$. The literal would be false if $\{P\}$ is possible, and true otherwise. Thus the truth of the third axiom in $\{P\}$ is still undetermined. We are blocked here.
3. But there is a way out of the deadlock. So far, the methods to determine whether a world is possible or impossible were perfectly symmetrical. The solution lies in breaking this symmetry. In T , we have a truth-sayers axiom: it is consistent to assume that Q is a theorem, and also to assume that Q is not a theorem. In semantical terms, both assumptions on world $\{P\}$ are consistent: if this world is chosen possible, then KQ is false and all axioms are satisfied in $\{P\}$; if it is chosen impossible, then KQ is true. Since we want to interpret the modal operator as a theoremhood modality, it is clear what assumption to make: that Q is not a theorem. We should make the assumption of *ignorance* and take it that the world is possible (and Q is not a theorem). Thus, we obtain again the possible world model $B_T = \{\{P\}, \{P, Q\}\}$.

From these two examples, we can extract the concepts necessary to formalize the above informal reasoning processes. At each step, we have *partial* information about the status of worlds that was gained so far. This naturally formalizes as a 3-valued set of worlds. We call such a set a *partial possible-world set*. Formally, a partial possible-world set \mathcal{B} is a function

$$\mathcal{B} : \mathcal{W} \rightarrow \{\mathbf{t}, \mathbf{f}, \mathbf{u}\},$$

where \mathcal{W} is the collection of all interpretations. Standard, *total* possible-world sets can be viewed as special cases, where the only two values in the range of the function are \mathbf{t} and \mathbf{f} . In the context of a partial possible world \mathcal{B} , we call a world w *certainly possible* if $\mathcal{B}(w) = \mathbf{t}$ and *potentially possible* if $\mathcal{B}(w) = \mathbf{t}$ or \mathbf{u} . Likewise, we call a world w *certainly impossible* if $\mathcal{B}(w) = \mathbf{f}$ and *potentially impossible* if $\mathcal{B}(w) = \mathbf{f}$ or \mathbf{u} . If $\mathcal{B}(w) = \mathbf{u}$, w is potentially possible and potentially impossible. We define $CP(\mathcal{B})$ as the set of certainly possible worlds of \mathcal{B} , $PP(\mathcal{B})$ as the set of potentially possible worlds, and likewise, $CI(\mathcal{B})$ and $PI(\mathcal{B})$ as the sets of certainly impossible, respectively potentially impossible worlds of \mathcal{B} .

At each inference step $\mathcal{B}_i \rightarrow \mathcal{B}_{i+1}$, we evaluated the propositions of T in one or more unknown worlds w , given the partial information available in \mathcal{B}_i . When all propositions of T turned out to be true in w , w was derived to be possible; if some evaluated to false, w was inferred to be impossible. To capture this formally, we need a three-valued truth function to evaluate theories in the context of a world w , the one we are examining, and a partial possible-world set \mathcal{B} . The value of this truth function on a theory T , denoted as $|T|^{\mathcal{B}, w}$, is selected from $\{\mathbf{t}, \mathbf{f}, \mathbf{u}\}$. There are some obvious properties that this function should satisfy.

1. The three-valued truth function should coincide with the standard (implicit) truth function for modal logic in total possible-world sets. In particular, when \mathcal{B} is a total possible-world set, that is, \mathcal{B} has no unknown worlds, then $|T|^{\mathcal{B}, w}$ should be true precisely when $\mathcal{B}, w \models T$ (and false, otherwise).

2. The three-valued truth function should be monotone with respect to the *precision* of the partial possible-world sets. A more precise partial possible-world set is one with fewer (with respect to inclusion) unknown worlds.

The intuition presented in (2) can be formalized as follows. We define $\mathcal{B} \leq_p \mathcal{B}'$ if $\mathcal{B}(w) \leq_p \mathcal{B}'(w)$, where the latter (partial) order \leq_p on truth values is the one generated by $\mathbf{u} \leq_p \mathbf{t}$ and $\mathbf{u} \leq_p \mathbf{f}$.

A three-valued truth function $|T|^{\mathcal{B},w}$ is monotone in \mathcal{B} if $\mathcal{B}' \leq_p \mathcal{B}''$ implies that $|T|^{\mathcal{B}',w} \leq_p |T|^{\mathcal{B}'',w}$. In particular, if $|T|^{\mathcal{B},w}$ is monotone in \mathcal{B} and B is a total possible-world set such that $\mathcal{B}' \leq_p B$, then $|T|^{\mathcal{B}',w} = \mathbf{t}$ implies that $B, w \models T$, and $|T|^{\mathcal{B}',w} = \mathbf{f}$ implies that $B, w \not\models T$.

Designing such a three-valued truth function is routine, the problem is that there is more than one sensible solution. One approach, originally proposed by Denecker et al. (1998), extends Kleene's (1952) three-valued truth evaluation to modal logic.

Definition 3 For a formula φ , world $w \in \mathcal{W}$ and partial possible-world set \mathcal{B} , we define $|\varphi|^{\mathcal{B},w}$ using the standard Kleene truth evaluation rules of three-valued logic augmented with one additional rule:

$$|K\varphi|^{\mathcal{B},w} = \begin{cases} \mathbf{f} & \text{if } |\varphi|^{\mathcal{B},w'} = \mathbf{f}, \text{ for some } w' \text{ such that } \mathcal{B}(w') = \mathbf{t} \\ \mathbf{t} & \text{if } |\varphi|^{\mathcal{B},w'} = \mathbf{t}, \text{ for all } w' \text{ such that } \mathcal{B}(w') = \mathbf{t} \text{ or } \mathbf{u} \\ \mathbf{u} & \text{otherwise.} \end{cases}$$

For a theory T , we define $|T|^{\mathcal{B},w}$ in the standard way of three-valued logic:

$$|T|^{\mathcal{B},w} = \begin{cases} \mathbf{f} & \text{if } |\varphi|^{\mathcal{B},w} = \mathbf{f}, \text{ for some } \varphi \in T \\ \mathbf{t} & \text{if } |\varphi|^{\mathcal{B},w} = \mathbf{t}, \text{ for all } \varphi \in T \\ \mathbf{u} & \text{otherwise.} \end{cases}$$

To illustrate the use of this truth function, let us evaluate the formula $K\varphi$, where φ is objective, in the context of a partial possible-world set \mathcal{B} and an arbitrary world w . We have $|K\varphi|^{\mathcal{B},w} = \mathbf{t}$ if $PP(\mathcal{B}) \models K\varphi$, that is, if all potentially possible worlds satisfy φ . Likewise, we have $|K\varphi|^{\mathcal{B},w} = \mathbf{f}$ if $CP(\mathcal{B}) \not\models K\varphi$, that is, at least one certainly possible world violates φ . Let B be a more precise total possible world set; that is, $\mathcal{B} \leq_p B$ or equivalently, $PP(\mathcal{B}) \supseteq B \supseteq CP(\mathcal{B})$. Then, obviously, if $K\varphi$ holds true in \mathcal{B} , the formula is true in B , and if $K\varphi$ is false in \mathcal{B} then it is false in B as well. In general this truth function is *conservative* (that is, \leq_p -monotone) in the sense that if a formula evaluates to true or false in some partial possible-world set, then it has the same truth value in every more precise possible-world set thus, in particular, in every total possible-world set B such that $\mathcal{B} \leq_p B$.

It is easy to see (and it was proven formally by Denecker et al. (2003)) that this truth function satisfies the two desiderata listed above. We also note that this is not the only reasonable way in which the three-valued truth function can be defined. We will come back on this topic in Section 4.5.

We now review the framework of semantics of autoepistemic reasoning we introduced in our study of the relationship between the default logic of Reiter and the autoepistemic logic of Moore (Denecker et al., 2003). We listed these semantics in

the previous section. All semantics in the framework require that a (partial) possible-world model \mathcal{B} of an autoepistemic theory be justified by some type of an inference process:

$$\mathcal{B}_0 \rightarrow \mathcal{B}_1 \rightarrow \dots \rightarrow \mathcal{B}_n = \mathcal{B}.$$

At each step i , modal literals $K\varphi$ appearing in T are evaluated in \mathcal{B}_i . When such literals are derived to be true or false, this might lead to further inferences in \mathcal{B}_{i+1} . Taking the semantic point of view, we understand an inference here as a step in which some worlds of undetermined status are derived to be possible and some others are derived impossible.

Dialects of autoepistemic logic, and so of default logic, too, differ from each other in the nature of the derivation step $\mathcal{B}_i \rightarrow \mathcal{B}_{i+1}$, and in initial assumptions \mathcal{B}_0 they make. Some dialects make no initial assumptions at all; in some others making certain initial “guesses” is allowed. In this way, we obtain autoepistemic logics of different degrees of *groundedness*. In the following sections, we describe inference processes underlying each of the four semantics in the framework described in Section 3.

Finally, we link the above concepts with the algebraic lattice theoretic concepts sketched in the previous section and used in the semantic framework of Denecker et al. (2003). There, the different semantics of an autoepistemic theory T emerged as different types of fixpoints of a \leq_p -monotone operator \mathcal{D}_T on the bilattice consisting of arbitrary pairs (B, B') of possible-world sets. The partial possible-world sets \mathcal{B} correspond to the *consistent* pairs $(PP(\mathcal{B}), CP(\mathcal{B}))$ in this bilattice; a pair (B, B') is *consistent* if $B \supseteq B'$, that is, certainly possible worlds are potentially possible. Inconsistent pairs give rise to possible-world sets that in addition to truth values **t**, **f** and **u** require the fourth one **i** for “inconsistency”. The Kleene truth function defined above can be extended easily to a four-valued truth function on the full bilattice.

The operator \mathcal{D}_T on that bilattice was then defined as follows:

$$\mathcal{D}_T(\mathcal{B}) = \mathcal{B}', \text{ if for every } w \in \mathcal{W}, \mathcal{B}'(w) = |T|^{\mathcal{B}, w}.$$

We observe that this operator maps partial possible-world sets into partial possible-world sets and that it coincides with Moore's derivation operator D_T when applied on total possible-world sets.

In the sequel, we will often represent a partial possible-world set \mathcal{B} in its bilattice representation, as the pair $(PP(\mathcal{B}), CP(\mathcal{B}))$ of respectively potentially possible and certainly possible worlds. For example, the least precise partial possible-world set \perp_p for $\Sigma = \{P, Q\}$ will be written as $(\{\emptyset, \{P\}, \{Q\}, \{P, Q\}\}, \emptyset)$: all worlds are potentially possible; no world is certainly possible.

We will now discuss the four semantics discussed above that define different dialects of autoepistemic reasoning.

4.1 The Kripke-Kleene semantics

This semantics is a direct formalization of the discussion above. We are given a finite modal theory T (we adopt the assumption of finiteness to simplify presentation, but it can be omitted). A *Kripke-Kleene inference process* is a sequence

$$\mathcal{B}_0 \rightarrow \dots \rightarrow \mathcal{B}_n$$

of partial possible-world sets such that:

1. \mathcal{B}_0 is the totally unknown partial possible-world set. That is, for every $w \in \mathcal{W}$, $\mathcal{B}_0(w) = \mathbf{u}$. We denote this partial possible-world set by \perp_p . This choice of the starting point indicates that Kripke-Kleene inference process does not make any initial assumptions.
2. For each $i = 0, \dots, n-1$, there is a set of worlds U such that for every $w \in U$, $\mathcal{B}_i(w) = \mathbf{u}$, $|T|^{\mathcal{B}_i, w} \neq \mathbf{u}$ and $\mathcal{B}_{i+1}(w) = |T|^{\mathcal{B}_i, w}$, and for every $w \notin U$, $\mathcal{B}_i(w) = \mathcal{B}_{i+1}(w)$. Thus, in each step of the derivation the status of the worlds that are certainly possible and certainly impossible does not change. All that can change is the status of some worlds of unknown status (worlds, that are potentially possible and potentially impossible). This set is denoted by U above. It is not necessary that U contains all worlds that are unknown in \mathcal{B}_i . In the derivation, worlds in U become certainly possible or certainly impossible, depending on how the theory T evaluates in them. If for such a potentially possible world $w \in U$, $|T|^{\mathcal{B}_i, w} = \mathbf{t}$, w becomes certainly possible. If $|T|^{\mathcal{B}_i, w} = \mathbf{f}$, w becomes certainly impossible. Otherwise, the status of w does not change. As such a derivation starts from the least precise, hence assumption-free, partial possible-world set \perp_p , all these derivations are assumption-free.
3. The halting condition: no more inferences can be made once we reach the state \mathcal{B}_n . Here this means that for each unknown $w \in \mathcal{W}$, $|T|^{\mathcal{B}_n, w} = \mathbf{u}$. The process terminates.

This precise definition formalizes and generalizes the informal construction we presented in the previous section. When applied to the theory we considered there,

$$T' = \{P, \neg KP \rightarrow Q\},$$

one Kripke-Kleene inference process that might be produced is (we represent here worlds, or interpretations, as sets of atoms they satisfy, and partial possible-world sets \mathcal{B} as pairs $(PP(\mathcal{B}), CP(\mathcal{B}))$):

$$\begin{array}{ll} \perp_p \rightarrow \mathcal{B}_1 = (\{\emptyset, \{P\}, \{P, Q\}\}, \emptyset) & \{Q\} \text{ certainly impossible} \\ \rightarrow \mathcal{B}_2 = (\{\{P\}, \{P, Q\}\}, \emptyset) & \emptyset \text{ certainly impossible} \\ \rightarrow \mathcal{B}_3 = (\{\{P\}, \{P, Q\}\}, \{\{P, Q\}\}) & \{P, Q\} \text{ certainly possible} \\ \rightarrow \mathcal{B}_4 = (\{\{P\}, \{P, Q\}\}, \{\{P\}, \{P, Q\}\}) & \{P\} \text{ certainly possible.} \end{array}$$

The first derivation can be made since $|P \wedge (\neg KP \rightarrow P)|^{\perp_p, w} = \mathbf{f}$, for $w = \{Q\}$ (in fact, for every w , in which P is false). The second derivation is justified similarly as the first one. The third derivation follows as $|P \wedge (\neg KP \rightarrow Q)|^{\mathcal{B}_2, w} = \mathbf{t}$, for $w = \{P, Q\}$, and the fourth one as $|P \wedge (\neg KP \rightarrow Q)|^{\mathcal{B}_3, w} = \mathbf{t}$, for $w = \{P\}$. Let us explain one more detail of the last of these claims. Here, $|P|^{\mathcal{B}_3, w} = \mathbf{t}$ holds because P holds in $w = \{P\}$. Moreover, $|KP|^{\mathcal{B}_3, w} = \mathbf{t}$ as P holds in every world that is potentially possible in \mathcal{B}_3 . Thus, $|\neg KP|^{\mathcal{B}_3, w} = \mathbf{f}$ and so indeed, $|\neg KP \rightarrow Q|^{\mathcal{B}_3, w} = \mathbf{t}$.

The shortest derivation sequence that corresponds exactly to the informal construction of the previous section is:

$$\perp_p \rightarrow (\{\{P\}, \{P, Q\}\}, \{\{P, Q\}\}) \rightarrow (\{\{P\}, \{P, Q\}\}, \{\{P\}, \{P, Q\}\}).$$

The fact that there may be multiple Kripke-Kleene inferences processes is not a problem as all of them end in the same partial possible-world.

Proposition 1 *For every modal theory T , all Kripke-Kleene inference processes converge to the same partial possible-world set, which is the \leq_p -least fixpoint of the operator \mathcal{D}_T .*

We call this special partial possible-world set the *Kripke-Kleene extension* of the modal theory T .

While the Kripke-Kleene construction is an intuitively sound construction, it has an obvious disadvantage: in general, its terminating partial belief state may not match the intended belief state even if T consists of “monotonic” inference rules (no negated modal atoms in the antecedents of formulas of the form (1)). An example where this happens is the truth sayer theory:

$$T = \{KP \rightarrow P\}.$$

It consists of a single monotonic inference rule, and its intended total possible-world set is $\{\{\emptyset, \{P\}\}$, which in the current (PP, CP) notation corresponds to

$$(\{\{\emptyset, \{P\}\}, \{\{\emptyset, \{P\}\}\}).$$

However, the one and only Kripke-Kleene construction is

$$\perp_p \rightarrow (\{\emptyset, \{P\}\}, \{\{P\}\}).$$

Then the construction halts. No more Kripke-Kleene inferences on the status of worlds can be made and the intended possible-world set is not reached.

We conclude with a historical note. The name Kripke-Kleene semantics was used for the first time in the context of the semantics of logic programs by Fitting (1985). Fitting built on ideas in an earlier work by Kleene (1952), and on Kripke's (1975) theory of truth, where Kripke discussed how to handle the liar paradox.

4.2 Moore's autoepistemic logic

Moore's autoepistemic logic has a simple formalization in our framework. A possible-world set B is an autoepistemic expansion of T if there is a *one-step* derivation for it:

$$\mathcal{B}_0 \rightarrow \mathcal{B}_1,$$

where $\mathcal{B}_0 = \mathcal{B}_1 = B$. Clearly, here we allow the inference process to make initial assumptions. Moreover, in the derivation step $\mathcal{B}_0 \rightarrow \mathcal{B}_1$ we simply verify that we made no incorrect assumptions and that no additional inferences can be drawn. The inference (more accurately here, the verification) process works as follows:

1. A world w is derived to be possible if $\mathcal{B}_0, w \models T$.
2. A world w is derived to be impossible if $\mathcal{B}_0, w \not\models T$.

Thus, formally, $\mathcal{B}_1 = \{w \mid \mathcal{B}_0, w \models T\} = D_T(\mathcal{B}_0)$. Consequently, the limits of this derivation process are indeed precisely the fixpoints of the Moore's operator D_T (we stress that we talk here only about total possible-world sets).

Since \mathcal{D}_T coincides with D_T on total possible-world sets, all autoepistemic expansions are fixpoints of \mathcal{D}_T . Thus, we have the following result.

Proposition 2 *The Kripke-Kleene extension is less precise than any other autoepistemic expansion of T . If the Kripke-Kleene extension is total, then it is the unique autoepistemic expansion of T .*

The weakness of Moore's logic from the point of view of modeling the autoepistemic view has been argued above. In Section 5, we will discuss another interpretation of autoepistemic logic in which his semantics may be more adequate.

4.3 The well-founded knowledge derivation

The problem with the Kripke-Kleene derivation is that it treats ignorance and knowledge in the same way. Ignorance is reflected by the presence of possible worlds. Knowledge is reflected by the presence of impossible worlds. In the Kripke-Kleene derivation, both possible and impossible worlds are derived in a symmetric way, by evaluating the theory T in the context of a world w , given the partial knowledge \mathcal{B} .

What we would like to do is to impose ignorance as a default. That a world is possible should not have to be derived. A world should be possible *unless* we can show that it is impossible. In other words, we need to impose a principle of *maximizing ignorance*, or equivalently, *minimizing knowledge*. Under such a principle, it is obvious that the possible-world set $\{\{P\}\}$ cannot be a model of the truth sayer theory $T = \{KP \rightarrow P\}$. It does not minimize knowledge while the other candidate for a model, the possible-world set $\{\emptyset, \{P\}\}$, does.

To refine the Kripke-Kleene construction of knowledge, we need an additional derivation step that allows us to introduce the assumption of ignorance. Intuitively, in such a derivation step, we consider a set U of unknown worlds, which are turned into certainly possible worlds to maximize ignorance.

Formally, a *well-founded inference process* is a derivation process $\mathcal{B}_0 \rightarrow \dots \rightarrow \mathcal{B}_n$ that satisfies the same conditions as a Kripke-Kleene inference process except that some derivation steps $\mathcal{B}_i \rightarrow \mathcal{B}_{i+1}$ may also be justified as follows (by the *maximize-ignorance* principle):

MI: There is a set U of worlds such that $\mathcal{B}_{i+1}(w) = \mathcal{B}_i(w)$ for all $w \notin U$ and for all $w \in U$, $\mathcal{B}_i(w) = \mathbf{u}$, $\mathcal{B}_{i+1}(w) = \mathbf{t}$ and $|T|^{\mathcal{B}_{i+1}, w} = \mathbf{t}$.

In other words, in such a step we pick a set U of unknown words, assume that they are certainly possible, and verify that this assumption was justified, that is, under the increased level of ignorance, all of them turn out to be certainly possible. To put it yet differently, we select a set U of unknown worlds, for which it is consistent to assume that they are certainly possible, and we turn them into certainly possible worlds (increasing our ignorance). By analogy with the notion of an unfounded set of atoms (Van Gelder et al., 1991), we call the set of worlds U , with respect to which the maximize-ignorance principle applies at the partial belief state \mathcal{B}_i , an *unfounded* set for \mathcal{B}_i .

We also note that the halting condition of a well-founded inference process is stronger than that for a Kripke-Kleene process. This means that for each unknown world w of \mathcal{B}_n , $|T|^{\mathcal{B}_n, w} = \mathbf{u}$ and in addition, \mathcal{B}_n does not allow a MI inference step, that is, it has no non-empty unfounded set.

There are two properties of well-founded inference processes that are worth noting.

Proposition 3 *All well-founded inference processes converge to the same (partial) possible-world set.*

This property gives rise to the *well-founded extension* of the modal theory T defined as the limit of *any* well-founded inference process. This limit can be shown to coincide with the well-founded fixpoint of \mathcal{D}_T , that is, the \leq_p -least fixpoint of the operator $\mathcal{S}_{\mathcal{D}_T}$ defined in the previous section.

Another important property concerns theories with no positive occurrences of the modal operator (for instance, theories consisting of formulas (1) with no modal literals $\neg K\beta_j$ in the antecedent).

Proposition 4 *If T contains only negative occurrences of the modal operator, then the well-founded extension is the \leq_k -least fixpoint of D_T .*

This property shows that the well-founded extension semantics has all key properties of the desired semantics of sets of “monotonic inference rules.” Let us revisit the truth sayer theory:

$$T = \{KP \rightarrow P\}.$$

The Kripke-Kleene construction is

$$\perp_p \rightarrow (\{\emptyset, \{P\}\}, \{\{P\}\}).$$

The inference that $\{P\}$ is possible is also sanctioned under the rules of the well-founded inference process. However, while there is no Kripke-Kleene derivation that applies now, the maximize-ignorance principle does apply and the well-founded inference process can continue. Namely, in the belief state given by $(\{\emptyset, \{P\}\}, \{\{P\}\})$, there is one world of unknown status (neither certainly impossible, nor certainly possible): \emptyset . Taking $U = \{\emptyset\}$ and applying the maximize-ignorance principle to U , we see that the well-founded inference process extends and yields $(\{\emptyset, \{P\}\}, \{\emptyset, \{P\}\})$. This possible-world set is total and so, necessarily, the limit of the process. Thus, this (total) possible-world set $\{\emptyset, \{P\}\}$ is the well-founded extension of the theory $\{KP \rightarrow P\}$.

The well-founded extension is total not only for monotonic theories. For instance, let us consider the theory:

$$T = \{KP \leftrightarrow Q\} \text{ or equivalently, } \{KP \rightarrow Q, \neg KP \rightarrow \neg Q\}.$$

Intuitively, there is nothing known about P , hence Q should be false. The unique Kripke-Kleene inference process ends where it starts, that is, with \perp_p . Indeed, when KP is unknown, no certainly possible or certainly impossible worlds can be derived. However, the possible-world set $U = \{\emptyset, \{P\}\}$ is unfounded with respect to \perp_p . Indeed, if both worlds are assumed possible, KP evaluates to false, and both worlds satisfy T . Thus, in the well-founded derivation we can establish that and then, in the next two steps, we can derive the impossibility of the two remaining unknown worlds, first of $\{Q\}$ and then of $\{P, Q\}$. This yields the following well-founded inference process:

$$\begin{aligned} \perp_p &\rightarrow \mathcal{B}_1 = (\{\emptyset, \{P\}, \{Q\}, \{P, Q\}\}, \{\emptyset, \{P\}\}) \\ &\rightarrow \mathcal{B}_2 = (\{\emptyset, \{P\}, \{P, Q\}\}, \{\emptyset, \{P\}\}) \\ &\rightarrow \mathcal{B}_3 = (\{\emptyset, \{P\}\}, \{\emptyset, \{P\}\}). \end{aligned}$$

In other cases, the well-founded extension is a partial possible-world set. An example is the theory:

$$\{\neg KP \rightarrow Q, \neg KQ \rightarrow P\}.$$

In this case, there is only one well-founded inference process, which derives that $\{P, Q\}$ is a certainly possible world and derives no certainly impossible worlds. That is, the well-founded extension is: $(\{\emptyset, \{P\}, \{Q\}, \{P, Q\}\}, \{\{P, Q\}\})$.

4.4 Stable possible-world sets

We recall that a partial possible-world set \mathcal{B} corresponds to the pair of total possible-worlds sets: $(PP(\mathcal{B}), CP(\mathcal{B}))$, where $PP(\mathcal{B})$ is the set of potentially possible worlds and $CP(\mathcal{B})$ is the set of certainly possible worlds.

We now define a *stable derivation* for a possible-world set B as a sequence of partial belief states of the form:

$$(\mathcal{W}, B) \rightarrow (PP_1, B) \rightarrow \dots \rightarrow (PP_{n-1}, B) \rightarrow (PP_n, B),$$

where:

1. $PP_n = B$
2. For every $i = 0, \dots, n-1$, and for every $w \in PP_i \setminus PP_{i+1}$, $|T|^{(PP_i, B), w} = \mathbf{f}$. That is, some worlds w in which T is false with respect to $\mathcal{B}_i = (PP_i, B)$ become certainly impossible and are removed from PP_i to form PP_{i+1} .
3. Halting condition: for every $w \in PP_n$, $|T|^{(PP_n, B), w} = \mathbf{t}$ or \mathbf{u} .

If a total belief set B has a stable derivation then we call B a *stable extension*. This concept captures the idea of the Reiter's extension of a default theory.

We recall that an inference rule (1) evaluates to false in world w with respect to (PP_i, B) if $w \not\models \gamma$, $PP_i \models K\alpha_i$, for all i , $0 \leq i \leq n$, and $B \not\models K\beta_j$, for all j , $0 \leq j \leq m$. We see here an asymmetric treatment of prerequisites α_i and justifications β_j which are evaluated in two different possible world sets. The same feature shows up, not coincidentally, in Reiter's definition of extension of a default theory.

The intuition underlying a stable derivation comes from a different implementation of the idea that ignorance does not need to be justified and that only knowledge must be justified. In a partial possible-world set \mathcal{B} , the component sets $PP(\mathcal{B})$ and $CP(\mathcal{B})$ have different roles. Since $PP(\mathcal{B})$ determines the certainly impossible worlds, this is the possible-world set that determines what is definitely known. On the other hand the set $CP(\mathcal{B})$ of certainly possible worlds determines what is definitely not known by \mathcal{B} .

A stable derivation for B is a justification for each impossible world of B (each world is initially potentially possible but eventually determined not to be in B , that is, determined impossible in B). The key point is that this justification may use the assumption of the ignorance in B . By fixing $CP(\mathcal{B}_i)$ to be B , it takes the ignorance in B for granted. What is justified in a stable inference process is the impossible worlds of B , not the possible worlds.

We saw above that the theory

$$\{\neg KP \rightarrow Q, \neg KQ \rightarrow P\}$$

has a partial well-founded extension. It turns out that it has two stable extensions $\{\{P\}, \{P, Q\}\}$ and $\{\{Q\}, \{P, Q\}\}$. For instance, the following stable derivation reconstructs $B = \{\{P\}, \{P, Q\}\}$. Note that in any partial possible-world set (\cdot, B) (that is, where the worlds of B are certainly possible), KQ evaluates to false. In all such cases, T evaluates to false in any world in which P is false. Hence we have the following very short stable derivation:

$$(\mathcal{W}, B) \rightarrow (B, B).$$

We now have two key results. The first one links up well-founded and stable extensions.

Proposition 5 *If the well-founded extension is a total possible-world set, it is the unique stable extension.*

The second result shows that indeed, the Konolige's translation works if the semantics of default logic of Reiter and the autoepistemic logic of Moore are correctly aligned. Here we state the result for the most important case of default extensions and stable extensions, but it extends, as we noted earlier, to all semantics we considered.

Proposition 6 *For every default theory Δ , B is an extension of Δ if and only if B is a stable extension of $Kon(\Delta)$.*

4.5 Discussion

We have obtained a framework with four different semantics. This framework is parameterized by the truth function. We have concentrated on the Kleene truth function but other viable choices exist. One is super-valuation (van Fraassen, 1966) which defines $|T|^{\mathcal{B},w}$ in terms of the evaluation of T in all possible world sets $B \geq_P \mathcal{B}$ approximated by \mathcal{B} . In particular,

$$|T|^{\mathcal{B},w} = \text{Min}_{\leq_P} \{|T|^{\mathcal{B},w} \mid \mathcal{B} \leq_P B\}.$$

In this way we obtain another instance of the framework, the family of *ultimate* semantics (Denecker et al., 2004). For many theories, the corresponding semantics of the two families coincide but ultimate semantics are sometimes more precise. An example is the theory $\{KP \vee \neg KP \rightarrow P\}$. It's Kripke-Kleene and well-founded extension is the partial possible world set $(\{\emptyset, \{P\}\}, \{\{P\}\})$ and there are no stable extensions. But the premise $KP \vee \neg KP$ is a propositional tautology, making $|T|^{\perp_P, w}$ true if $w \models P$ and false otherwise. As a consequence, the ultimate Kripke-Kleene, well-founded and unique stable extension is $\{\{P\}\}$.

For a scientist interested in the formal study of the informal semantics of a certain type of (informal) propositions this diversity is troubling. Indeed, what is then the nature of autoepistemic reasoning, and which of the semantics that we defined and that can be defined by means of other truth functions is the "correct" one? It is necessary to bring some order to this diversity.

In the autotheoremhood view, the formal semantics should capture the information content of an autoepistemic theory T that contains propositions referring to T 's own information content; the semantics should determine whether a world is possible or

impossible, or equivalently, whether a formula is or is not entailed by T . As we saw, Moore’s semantics of expansions and the Kripke-Kleene extension semantics are arguably less suited in the case of monotonic inference rules with cyclic dependencies (cf. the truth sayer theory). This leaves us with four contenders only: the well-founded and the stable extension semantics and their ultimate versions. All employ a technique to maximize ignorance and correctly handle autoepistemic theories with monotonic inference rules. Which of these semantics is to be preferred?

Let us first consider the choice of the truth function. The semantics based on the Kleene truth function and the ones induced by super-valuation make different trade-offs: the higher precision of the ultimate semantics, which is good, comes at the price of higher complexity of reasoning, which is bad (Denecker et al., 2004). When there is a trade-off between different desired characteristics, there is per definition no *best* solution. Yet, when looking closer, the question of the choice between these two truth functions turns out to be largely *academic* and without much practical relevance. There are classes of autoepistemic theories for which the Kleene and the super-valuation truth functions coincide, and hence, so do the semantics they induce. Denecker et al. (2004, Proposition 6.14) provide an example of such a class. Even more importantly, the semantics induced by Kleene’s truth function and by super-valuation differ only when case-based reasoning on modal literals is necessary to make certain inferences. Except for our own artificial examples introduced to illustrate the formal difference between both semantics (Denecker et al., 2004), we are not aware of any reasonable autoepistemic or default theory in the literature where such reasoning would be necessary. They may exist, but if they do, they will constitute an insignificant fringe. The take-home message here is that in all practical applications that we are aware of, the Kleene truth function suffices and there is no need to pay for the increased complexity of super-valuation. This limits the number of semantics still in the running to only two. Of the remaining two, the most faithful formalization of the autotheoremhood view seems to be the well-founded extension semantics. As we view a theory as a set of inference rules, the construction of the well-founded extension formalizes the process of the application of the inference rules more directly than the construction of the stable extension semantics.

Nevertheless, there are some commonsense arguments for not overemphasizing the differences between these semantics. First, we should keep in mind that theories of interest are those that are developed by human experts, and hence, are meaningful to them. What are the meaningful theories in the autotheoremhood? Not every syntactically correct modal theory makes sense in this view. “Paradoxical” theories such as the liar theory T_{liar} can simply not be ascribed an information content in a consistent manner and are not a sensible theory in the autotheoremhood view. For theories T viewed as sets of inference rules, the inference process associated with the theory should be able to determine the possibility of each world and hence, for each proposition, whether it is a theorem or not of T . In particular, this is the case when the well-founded extension is total. We view theories with theorems that are subject to ambiguity and speculation with suspicion. And so, methodologies based on the autotheoremhood view will naturally tend to produce theories with a total well-founded extension. From a practical point of view, the presence of a unique, constructible state of belief for an autoepistemic theory is a great advantage. For instance, unless the polynomial hierarchy collapses, for such theories the task to construct the

well-founded extension and so, also the unique stable expansion, is easier than that of computing a stable expansion of an arbitrary theory or to determine that none exists. Further, for such theories, skeptical and credulous reasoning (with respect to stable extension) coincide and are easier, again assuming that the polynomial hierarchy does not collapse, than they are in the general case.

For all the reasons above, a human expert using autoepistemic logic in the autotheoremmhood view, will be naturally inclined to build an autoepistemic knowledge base with a well-founded extension that is total. When the well-founded semantics induced by the Kleene truth function is total, the four semantics — the two stable semantics and the two well-founded semantics — coincide! It is so, in particular for the class of theories built of formulas (1) with no recursion through negated modal literals (the so-called *stratified* theories (Gelfond, 1987)). Hence, such a methodology could be enforced by imposing syntactical conditions.

All these arguments notwithstanding, the fact is that many default theories discussed in the literature or arising in practical settings do not have a unique well-founded extension and that the stable and well-founded extension semantics do not coincide⁷. We have seen it above in the Nixon Diamond example. More generally, it is the case whenever the theory includes conflicting defaults and no guidance on how to resolve conflicts. Such conflicts may arise inadvertently for the programmer, in which case a good strategy seems to be to analyze the conflicts (potentially by studying the stable extensions) and to refine the theory by building in conflict-resolution in the conditions of default rules. Otherwise, when conflicts are a deliberate decision of the programmer who indeed does not want to offer rules to resolve conflicts, all we can do is to accept each of the multiple stable extensions as a possible model of the theory and also accept that none of them is in any way preferred to others.

In conclusion, rather than pronouncing a strong preference for the well-founded extension over stable extensions or vice versa, what we want to point out is the attractive features of theories for which these two semantics coincide, and advantages of methodologies that lead to such theories.

5 Autoepistemic Logics in a Broader Landscape

In this section, we use the newly gained insights on the nature of autoepistemic reasoning to clarify certain aspects of autoepistemic logic and its position in the spectrum of logics, in particular in the families of logics of nonmonotonic reasoning and classical modal logics.

A good start for this discussion is Moore's "second" view on autoepistemic logic. Later in his paper, when developing the expansion semantics, Moore rephrased his views on autoepistemic reasoning in terms of the background concept of an *autoepistemic agent*. Such an agent is assumed to be ideally rational and have the powers of perfect introspection. An autoepistemic theory T is viewed as a set of propositions

⁷Some researchers believe that multiple extensions are *needed* for reasoning in the context of incomplete knowledge. Our point of view is different. The essence of incomplete knowledge is that different states of affairs are possible. Therefore, the natural — and standard — representation of a belief state with incomplete knowledge is by one possible-world set with multiple possible worlds, and not by multiple possible-world sets, which to us would reflect the state of mind of an agent that does not know what to believe.

that are known by this agent. Modal literals $K\varphi$ in T now mean “*I (that is, the agent) know φ* ”. The most important assumption, the one on which this informal view of autoepistemic logic largely rests, is that the agent’s theory T represents *all the agent knows* (Levesque, 1990) or, in Moore’s terminology, what the agent knows is *grounded* in the theory. We will call this implicit assumption the *All I Know* Assumption.

Without the *All I Know* Assumption, the theory T would be just a list of believed introspective propositions. The state of belief of the agent might then correspond to any possible-world set B such that $B \models K\varphi$, for each $\varphi \in T$ (where $B \models K\varphi$ if for all $w \in B$, $B, w \models \varphi$). But in many such possible-world sets B , the agent would know much more than what can be derived from T . In this setting, nonmonotonic inference rules such as $KA(x) \wedge \neg K\neg B(x) \rightarrow B(x)$ would not be useful for default reasoning since conclusions drawn from them would not be derived from the information given in T . So the problem is to model the *All I Know* Assumption in the semantics. Moore implemented this condition by imposing that for any model B , if $B, w \models T$, then w is possible according to B , i.e., $w \in B$. Combining both conditions, models that satisfy the *All I Know* Assumption are fixpoints of D_T , that is Moore’s expansions.

Moore’s expansion semantics does not violate the assumptions underlying the autoepistemic agent view. Expansions do correspond to belief states of an ideally rational, fully introspective agent that believes all axioms in T and, in a sense, does not believe more than what he can *justify* from T . But the same can be said for the autotheoremhood view as implemented in the well-founded and stable extension semantics. We may identify the theory with what the agent knows, and the theoremhood operator with the agent’s epistemic operator K , and see the well-founded extension (if it is total) or stable extensions as representing belief states of an agent that can be *justified* from T .

As we stated in the previous section, Moore’s expansion semantics does not formalize the autotheoremhood view, but it formalizes a dialect of autoepistemic reasoning, based on an autoepistemic agent that accepts states of belief with a weaker notion of justification, allowing for *self-supporting* states of belief. While not appropriate for modeling default reasoning, the semantics may work well in other domains. Indeed, humans sometimes do hold self-supporting beliefs. For example, self-confidence, or lack of self-confidence often are to some extent self-supported. Believing in one’s own qualities makes one perform better. And a good performance supports self-confidence (and self-esteem). Applied to a scientist, this loop might be represented by the theory consisting of the following formulas:

$$\begin{aligned} K(ICanSolveHardProblems) &\rightarrow Happy \\ Happy &\rightarrow Relaxed \\ Relaxed &\rightarrow ICanSolveHardProblems. \end{aligned}$$

Along similar lines, the placebo effect is a medically well-researched fact often attributed to self-supporting beliefs. The self-supporting aspect underlying the placebo-effect can be described by the theory consisting of the rules:

$$\begin{aligned} K(IGetBetter) &\rightarrow Optimistic \\ Optimistic &\rightarrow IGetBetter. \end{aligned}$$

Taking a placebo just flips the patients into the belief that they are getting better. In this form of autoepistemic reasoning of an agent, self-supporting beliefs are justified

and Moore's expansion semantics, difficult to reconcile with the notion of derivation and theorem, may be suitable.

There are yet other instances of the *All I Know* Assumption in the autoepistemic agent view. For example, let us consider the theory $T = \{KP\}$. In the autotheoremhood view, this theory is clearly inconsistent, for there is no way this theory can prove P . The situation is not so clear-cut in the agent view. We see no obvious argument why the agent could not be in a state of belief in which he believes P and its consequences and nothing more than that. In fact, the logic of *minimal knowledge* (Halpern and Moses, 1984) introduced as a variant of autoepistemic logic accepts this state of belief for T .

What our discussion shows is that the *All I Know* Assumption in Moore's autoepistemic agent view is a rather vague intuition, which can be worked out in more than one way, yielding different formalizations and different dialects. It may explain why Moore built a semantics that did not satisfy his own first intuitions (inference rules) and why Halpern (1997) could build several formalizations for the intuitions expressed by Reiter and Moore. In contrast, the autotheoremhood view eliminates the agent from the picture and hence, the difficult tasks to specify carefully the key concepts such as ideal rationality, perfect introspection and, most of all, the *All I Know* Assumption. Instead, it builds on more solid concepts of inference rules, theoremhood and entailment which yields a more precise intuition.

6 Conclusions

We presented here an analysis of informal foundations of autoepistemic reasoning. We showed that there is principled way to arrive at all major semantics of logics of autoepistemic reasoning taking as the point of departure the autotheoremhood view of a theory. We see the main contributions of our work as follows.

First, extending Moore's arguments we clarified the different nature of defaults and autoepistemic propositions. Looking back at Reiter's intuitions, we now see that, just as Moore had claimed about McDermott and Doyle, also Reiter built an autoepistemic logic and not a logic of defaults. We showed that some long-standing problems with default logic can be traced back to pitfalls of using the autoepistemic propositions to encode defaults. On the other hand, we also showed that once we focus theories understood as consisting of autoepistemic propositions and adopt the autotheoremhood perspective, we are led naturally to the Kripke-Kleene semantics, the semantics of expansions by Moore, the well-founded semantics and the semantics of extensions by Reiter.

Second, we analyzed what can be seen as the center of autoepistemic logic: the *All I Know* Assumption. We showed that this rather fuzzy notion leads to multiple perspectives on autoepistemic reasoning and to multiple dialects of the autoepistemic language, induced by different notions of what can be derived from (or is *grounded in*) a theory. One particularly useful informal perspective on autoepistemic logic goes back to Moore's truly insightful view of autoepistemic rules as inference rules. This view, which we called the autotheoremhood view, was the main focus of our discussion. In this view, theories "contain" their own entailment operator and "I" in the *All I Know* Assumption is understood as the theory itself. The most faithful formal-

ization of this view is the well-founded extension semantics but the stable-extension semantics, which extends Reiter's semantics to autoepistemic logic, coincides with the well-founded extension semantics wherever the autotheoremhood view seems to make sense. Thus, it was Reiter's default logic that for the first time incorporated into the reasoning process the principle of knowledge minimization, resulting in a better formalization of Moore's intuitions than Moore's own logic.

Fifteen years ago Halpern (1997) analyzed the intuitions of Reiter, McDermott and Doyle, and Moore, and showed that there are alternative ways, in which they could be formalized. Halpern's work suggested that the logics proposed by Reiter, McDermott and Doyle, and Moore are not necessarily "determined" by these intuitions. We argue here that by looking more carefully at the informal semantics of those logics, they do indeed seem "predestined" and can be derived in a systematic and principled way from a few basic informal intuitions.

Acknowledgments

The work of the first author was partially supported by FWO-Vlaanderen under project G.0489.10N. The work of the third author was partially supported by the NSF grant IIS-0913459.

References

- C. Baral and V. S. Subrahmanian. Duality between alternative semantics of logic programs and nonmonotonic formalisms. In A. Nerode, W. Marek, and V. Subrahmanian, editors, *Proceedings of the 1st International Workshop on Logic Programming and Nonmonotonic Reasoning*, pages 69–86. MIT Press, 1991.
- J. Barwise and J. Etchemendy. *The Liar: An Essay on Truth and Circularity*. Oxford University Press, 1987.
- G. Brewka. Cumulative default logic: in defense of nonmonotonic inference rules. *Artificial Intelligence*, 50(2):183–205, 1991.
- K. L. Clark. Negation as failure. In H. Gallaire and J. Minker, editors, *Logic and Data Bases*, pages 293–322. Plenum Press, 1978.
- M. Denecker, V. W. Marek, and M.ław Truszczyński. Approximating operators, stable operators, well-founded fixpoints and applications in non-monotonic reasoning. In J. Minker, editor, *Logic-based Artificial Intelligence*, The Kluwer International Series in Engineering and Computer Science, pages 127–144. Kluwer Academic Publishers, Boston, 2000.
- M. Denecker, V.W. Marek, and M. Truszczyński. Uniform semantic treatment of default and autoepistemic logics. *Artificial Intelligence*, 143(1):79–122, 2003.
- M. Denecker, V. Marek, and M. Truszczyński. Ultimate approximation and its application in nonmonotonic knowledge representation systems. *Information and Computation*, 192:84–121, 2004.
- M. Denecker, V. Marek, and M. Truszczyński. Fixpoint 3-valued semantics for autoepistemic logic. In *Proceedings of AAAI 1998*, pages 840–845. MIT Press, 1998.

- M. Fitting. A Kripke-Kleene Semantics for Logic Programs. *Journal of Logic Programming*, 2(4):295–312, 1985.
- M. Fitting. A theory of truth that prefers falsehood. *Journal of Philosophical Logic*, 26:477–500, 1997.
- M. Gelfond. On stratified autoepistemic theories. In *Proceedings of AAAI 1987*, pages 207–211. Morgan Kaufmann, 1987.
- M. Gelfond and V. Lifschitz. The stable model semantics for logic programming. In R. A. Kowalski and K. A. Bowen, editors, *Proceedings of the International Joint Conference and Symposium on Logic Programming*, pages 1070–1080. MIT Press, 1988.
- G. Gottlob. Translating default logic into standard autoepistemic logic. *Journal of the ACM*, 42(4), 1995.
- J. Y. Halpern. A critical reexamination of default logic, autoepistemic logic, and only knowing. *Computational Intelligence*, 13(1):144–163, 1997.
- J. Y. Halpern and Y. Moses. Towards a theory of knowledge and ignorance: Preliminary report. In R. Reiter, editor, *Proceedings of the Workshop on Non-Monotonic Reasoning*, pages 125–143, 1984.
- S. C. Kleene. *Introduction to Metamathematics*. Van Nostrand, 1952.
- K. Konolige. On the relation between default and autoepistemic logic. *Artificial Intelligence*, 35:343–382, 1988.
- K. Konolige. Errata: On the relation between default and autoepistemic logic. *Artificial Intelligence*, 41(1):115, 1989.
- S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.
- S. Kripke. Outline of a theory of truth. *Journal of Philosophy*, 72:690–712, 1975.
- D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55:1–60, 1992.
- D. J. Lehmann. What does a conditional knowledge base entail? In R. J. Brachman, H. J. Levesque, and R. Reiter, editors, *Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning, KR 1989*, pages 212–222. Morgan Kaufmann, 1989.
- H. J. Levesque. All I know: a study in autoepistemic logic. *Artificial Intelligence*, 42(2-3): 263–309, 1990.
- W. Lukaszewicz. Considerations on default logic: an alternative approach. *Computational Intelligence*, 4:1–16, 1988.
- D. Makinson. General theory of cumulative inference. In M. Reinfrank, J. de Kleer, M.L. Ginsberg, and E. Sandewall, editors, *Nonmonotonic reasoning*, volume 346 of *Lecture Notes in Computer Science*, pages 1–18, Berlin-New York, 1989. Springer.
- W. Marek and M. Truszczyński. Relating autoepistemic and default logics. In R. J. Brachman, H. J. Levesque, and R. Reiter, editors, *Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning, KR 1989*, pages 276–288. Morgan Kaufmann, 1989.

- J. McCarthy. Circumscription - a form of nonmonotonic reasoning. *Artificial Intelligence*, 13: 27–39, 1980.
- J. McCarthy. Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence*, 28(1):89–116, 1986.
- J. McCarthy and P.J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969.
- D. McDermott. Nonmonotonic Logic II: Nonmonotonic Modal Theories. *Journal of the ACM*, 29(1):33–57, 1982.
- D. McDermott and J. Doyle. Nonmonotonic Logic I. *Artificial Intelligence*, 13(1-2):41–72, 1980.
- A. Mikitiuk and M. Truszczyński. Constrained and rational default logics. In C. S. Mellish, editor, *Proceedings of IJCAI 1995*, pages 1509–1515. Morgan Kaufmann, 1995.
- R. C. Moore. Possible-world semantics for autoepistemic logic. In R. Reiter, editor, *Proceedings of the Workshop on Non-Monotonic Reasoning*, pages 344–354, 1984. Reprinted in: M. Ginsberg, editor, *Readings on Nonmonotonic Reasoning*, pages 137–142, Morgan Kaufmann, 1990.
- R. C. Moore. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, 25(1): 75–94, 1985.
- J. Pearl. System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning. In R. Parikh, editor, *Proceedings of the 3rd Conference on Theoretical Aspects of Reasoning about Knowledge, TARK 1990*, pages 121–135. Morgan Kaufmann, 1990.
- R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1-2):81–132, 1980.
- R. Reiter and G. Criscuolo. On interacting defaults. In P. J. Hayes, editor, *Proceedings of IJCAI 1981*, pages 270–276. William Kaufman, 1981.
- T. Schaub. On constrained default theories. In B. Neumann, editor, *Proceedings of the 11th European Conference on Artificial Intelligence*, pages 304–308. Wiley and Sons, 1992.
- Y. Shoham. Nonmonotonic logics: meaning and utility. In J. P. McDermott, editor, *Proceedings of IJCAI 1987*, pages 388–393. Morgan Kaufmann, 1987.
- A. Tarski. The concept of truth in formalized languages. In J. Corcoran, editor, *Logic, Semantics, Metamathematics: Papers from 1923-38*, pages pp. 152–278. Indianapolis: Hackett, 1983. Translated from Tarski 1935 by J. H. Woodger.
- D. Touretzky. *The mathematics of inheritance systems*. Pitman, London, 1986.
- M. Truszczyński. Modal interpretations of default logic. In R. Reiter and J. Mylopoulos, editors, *Proceedings of IJCAI 1991*, pages 393–398. Morgan Kaufmann, 1991.
- B. van Fraassen. Singular terms, truth-value gaps and free logic. *Journal of Philosophy*, 63(17): 481–495, 1966.
- A. Van Gelder, K. A. Ross, and J. S. Schlipf. The well-founded semantics for general logic programs. *Journal of the ACM*, 38(3):620–650, 1991.