ON THE RELATION BETWEEN DEFAULT THEORIES AND AUTOEPISTEMIC LOGIC*

Kurt Konolige
Artificial Intelligence Center and CSLI, SRI International
333 Ravenswood, Menlo Park, Ca. 94025
(415)859-2788 KONOLIGE@AI.SRI.COM

Abstract

Default theories are a formal means of reasoning about defaults: what *normally* is the case, in the absence of contradicting information. Autoepistemic theories, on the other hand, are meant to describe the consequences of reasoning about ignorance: what must be true if a certain fact is *not* known. Although the motivation and formal character of these systems are different, a closer analysis shows that they bear a common trait, which is the indexical nature of certain elements in the theory. In this paper we compare the expressive power of the two systems. First, we give an effective translation of default theories into autoepistemic logic; default theories can thus be embedded into autoepistemic logic. A more suprising result is that the reverse translation is also possible: every set of sentences in autoepistemic logic can be effectively rewritten as a default theory. The formal equivalence of these two differing systems is thus established. Some benefits of this analysis are that it gives an interpretive semantics to default theories, and yields insight into the nature of defaults in autoepistemic reasoning.

1 Introduction

Default reasoning can be informally described as jumping to conclusions based on what is normally the case. To say that ''power corrupts," for example, is to say that for typical x, in typical situations, *x* will be corrupted by the exercise of authority.

Default logic [9] is a formalization of default reasoning. An agent's knowledge base (KB), its collection of facts about the world, is taken to be a first-order theory. Default reasoning is expressed by *default rules* of the form

which can be read as, roughly, "if a is provable from the KB, and b is consistent with it, then assume w as a default." Unlike ordinary first-order inference rules, default rules are *defeasible:* given KB containing just a, for example, the rule above would allow the inference of a;, but if -b is added to the KB, then the default rule is no longer applicable. Default rules are thus nonmonotonic inference rules.

In default logic, the default rules operate at a metatheoretic level, as they are not expressed in the language of the KB, and are not inference rules within the KB. Rather, they can be thought of as a means of taking a KB and transforming it into another one by the addition of sentences which are not logically derivable from the original. The transformation is defined in terms of a fix-point operator. This formulation of default reasoning leads us to ask several questions, which do not have readily apparent answers. The first concerns the expressiveness of the logic. Certain simple types of defaults can be readily stated; for example, "power corrupts" could be expressed as

$$\frac{Powerful(x) : MCorrupt(x)}{Corrupt(x)}$$

But it is not clear that more complicated constructs could be accommodated. A case in point is conditional defaults, where a default rule is the conclusion of an implication; or defaults whose consequent is itself a default. Because the default rules are not part of the logical language, there is no obvious, straightforward expression of these concepts.

The second question, related to the first, concerns the semantics of default theories. Because defaults are expressed as inference rules operating in conjunction with a fixed-point construction, it is not clear what the *meaning* of such objects as *MB* is. In some recent work, there have been proposals for a semantics for a restricted class of default theories [6] and for default theories in general [1]. In both cases, the "semantics" is a reformulation of the KB-transformation induced by the defaults in terms of restrictions on the models of the KB. Although such a reformulation can provide an alternative view of the construction of default theories, it does not provide a semantics in the sense of providing an interpretation for default rules in a model structure (an *interpretive semantics).* Indeed, because defaults are expressed as inference rules, they are not amenable to interpretation in this fashion.

Our idea in this paper is to define default reasoning within the theory of the KB itself, rather than as a transformation of the KB. If we take the sentences of a KB to be

'This research was supported in part by contract N00014-80-C-0296 from the Office of Naval Research, and in part by a gift from the System Development Foundation.

the knowledge or beliefs of an agent, then defaults can be expressed by referring to what an agent *doesn't know*. The default that "power corrupts" could be stated informally as:

> If x is powerful, then assume x is corrupt
> *if nothing known contradicts it.* (3)

It is easy to see that such reasoning is defeasible in the presence of additional information about the integrity of x. From a formal point of view, it is clear that to assert this statement, the language of the KB must be augmented by a construction that refers to the KB as a whole.

Let us call a theory containing an operator that refers to the theory itself an *indexical theory*. We will use the expression $L\phi$ within a theory to mean that the sentence ϕ is part of theory itself. Now we can rephrase the default rule (1) in the following manner, using the operator L:

$$L\alpha \wedge \neg L\neg\beta \supset \omega\ . \tag{4}$$

The intent of a rule of this form is something like: "if α is in the KB, and $\neg\beta$ is *not* in the KB, then ω is true." The negation sign in $\neg\beta$ arises because we have chosen to use provability operator L which is the dual of the consistency operator M. Because L is an operator of the KB language, we have been able to express the default within the language of the KB itself, rather than as a metatheoretic construct.

The introduction of an indexical operator is an added complexity, for now we allow our initial KB to contain not only statements about the world, but also about its own contents. Indeed, even interpreting the modal operators of (4) is problematical. Fortunately, the mathematical properties of indexical theories have recently been studied by Moore [8] as a formalization for a another type of nonmonotonic reasoning, called *autoepistemic reasoning*, in which an agent reasons about the relationship of her knowledge to the world. He has derived an elegant and natural interpretive semantics for indexical theories incorporating the self-referential operator L. This semantics gives an interpretation to the operator L based on individual model structures.

We are naturally led to ask what relationship exists between default theories and their corresponding expression in AE logic. Are they essentially different in the sense that agents using each one would have widely differing sets of beliefs? The answer, which is the main result of this paper, is *no*: default logic and AE logic sanction the same inferences on corresponding initial inputs. This has several important consequences. Since default rules are expressible in AE logic, both default and autoepistemic reasoning can be combined within this single formalism. Also, the formal expression of defaults gains the benefits of an interpretive semantics.

A second and more surprising consequence is that AE logic is no more expressive than default logic, even though the L operator is part of the language: there exists a translation from every set of AE logic premises into a corresponding default theory. As we shall see, by translating the appropriate AE logic statements, it is possible to construct default theories with the effect of conditional defaults, defaults whose conclusion is a default, and so on. The expression of these concepts is still much more natural in terms of the L-operator, but the mathematical properties of the corresponding default theories are the same.

Of independent interest are some results in the theory of AE logic, especially the characterization and equivalence of *strongly grounded* and *minimal* extensions.

2 Autoepistemic logic

Autoepistemic (AE) logic was defined by Moore [8] as a formal account of an agent reasoning about her own beliefs. The agent's beliefs are assumed to be a set of sentences in some logical language augmented by a modal operator L. The intended meaning of $L\phi$ is that ϕ is one of the the agent's beliefs; thus the agent could have beliefs about her own beliefs. For example, consider a space shuttle flight director who believes that it is safe to launch not because of any positive information, but by reasoning that if something were wrong, she would know about it from her engineers. This belief can be expressed using sentences of the augmented language. If P stands for "it is safe to launch the shuttle," then

$$\neg L\neg P \supset P \tag{5}$$

expresses the flight director's self-knowledge. Equation (5) is a logical constraint between a belief state ($L\neg P$) and a condition on the world (P).

The primary focus of AE logic is a normative one: given an initial (or *base*) set of beliefs A about the world, what final set T should an ideal *reflective* agent settle on? If we restrict ourselves for the moment to languages without the self-belief operator, then clearly an ideal agent should believe all of the logical consequences of her base beliefs, a condition sometimes referred to as *logical omniscience* [2]. More formally, let the expression $\Gamma \models \phi$ mean that the sentence ϕ is logically implied by the set of sentences Γ. Then, if the base set is A, the belief set T of an ideal agent is given by:

$$T = \{\phi \mid A \models \phi\} \tag{6}$$

The presence of a self-belief operator complicates matters. Because the intended meaning of $L\phi$ depends on the belief set of the agent, the definition of the belief set itself becomes circular, which necessitates the use of a fixed-point equation to define T. In this section we will present this definition and give several alternative formulations that will prove useful.

2.1 Logical preliminaries

We begin with a language $\mathcal{L}$ for expressing self-belief, and introduce valuations of $\mathcal{L}$. The treatment generally follows and extends Moore [8], but differs in two ways. First, the base language is first-order rather than propositional; but this is a minor change, because no quantifying-in is permitted. Second, ideal belief sets are defined with a fixed-point equation over valuations of the language. This definition is equivalent to Moore's original one, but leads to different insights on the nature of the ideal belief set, simpler proofs of many results, and several natural extensions. Because of space limitations, no proofs are included; they may be found in the fuller version of the paper (Konolige [4]).

Let $\mathcal{L}_0$ be a first-order language with functional terms. The normal formation rules for formulas of first-order languages hold. A *sentence* of $\mathcal{L}_0$ is a formula with no free variables; an *atom* is a sentence of the form $P(t_1, \cdots, t_n)$. We extend $\mathcal{L}_0$ by adding a unary modal operator L; the extended language is called $\mathcal{L}$. $\mathcal{L}$ can be defined recursively as containing all the formation rules of $\mathcal{L}_0$, plus the following:

$$\text{If } \phi \text{ is a sentence of } \mathcal{L}, \text{ then so is } L\phi. \tag{7}$$

An expression $L\phi$ is a *modal atom*. Sentences and atoms of $\mathcal{L}_0$ are called *ordinary*. Note that nestings such as $LL\phi$ are modal atoms (and hence sentences) of $\mathcal{L}$. A sentence has a *modal depth* n if its modal operators are nested to a depth of n, *e.g.*, $L(P \vee LP)$ has a modal depth of 2. We use the abbreviation $\mathcal{L}_n$ for the set of all sentences of modal depth n or less. Because the argument of a modal operator never contains free variables, there is no quantifying into the scope of a modal atom, *e.g.*, $\exists x LPx$ is not allowed.

From the point of view of first-order valuations, the modal atoms $L\phi$ are simply nilary predicates. Our intended interpretation of these atoms is that ϕ is an element of the belief set of the agent. So we will consider valuations of $\mathcal{L}$ to be standard first-order valuations, with the addition of a belief set Γ. The atoms $L\phi$ are interpreted as true or false depending on whether ϕ is in Γ.

The interaction of the interpretation of L with first-order valuations is often a delicate matter in this paper, and so a perspicuous terminology for talking about valuations is necessary. In particular, it is often useful to decouple the interpretation of modal and ordinary atoms. First-order valuations are built upon the truthvalues of atoms: for ordinary atoms, this is given by a structure $(U, \varphi, \mathcal{R})$, where φ is a mapping from terms to elements of the universe U, and $\mathcal{R}$ is a set of relations over U, one for each predicate. We will refer to any such structure as an *ordinary index*, and denote it with the symbol I. Modal atoms are given a truthvalue by a belief set Γ, which is called a *modal index*.

The truthvalue of any sentence in $\mathcal{L}$ can be determined by the normal rules for first-order valuations, given an ordinary and modal index. We write $\models_{I,\Gamma} \phi$ if a valuation (I, Γ) satisfies ϕ. The valuation rule for modal atoms can be written as:

$$\models_{I,\Gamma} L\phi \quad \text{if and only if} \quad \phi \in \Gamma \tag{8}$$

A valuation which makes every member of a set of sentences true is called a *model* of the set. A sentence which is true in every member of a class of valuations is called *valid* with respect to the class. The following classes of valuations are useful:

$$\begin{array}{rl} \models_\Gamma & \text{valuations with modal index } \Gamma \\ \models & \text{all valuations} \\ \Sigma \models & \text{models of } \Sigma \end{array} \tag{9}$$

A sentence ϕ is a *first-order consequence* (FOC) of a set of sentences Σ if it is true in all models of Σ. Σ is *closed under first-order consequence* if it contains all sentences that are true in all of its models.

2.2 Autoepistemic extensions

Now we return to the original question of what an ideal reflective agent should believe. Obviously, we want to use equation (6), with an appropriate choice for logical implication. Given that the intended meaning of L is *self*-belief, it becomes obvious that we should consider all models in which the interpretation of $L\phi$ is the belief set of the agent itself, that is, the valuations we consider all have a modal index which is the belief set of the agent. Following Moore, we call such valuations *autoepistemic* (or *AE*), and define the concept of an extension of a base set of beliefs.

Definition 2.1 *Any set of sentences T which satisfies the equation*

$$T = \{\phi \mid A \models_T \phi\}$$

is an autoepistemic extension *of A.*

This is a fixed-point equation for the belief set T of a reflective agent, given premises A.[1] It is similar to the belief set definition for a nonreflective agent (equation 6) in that it contains A and is closed under first-order consequence. AE extensions are candidates for the belief sets of ideal reflective agents; however, as we will show in section 2.4 below, there is an additional restriction which such sets should obey.

Example 2.1 A base set A may give rise to one, no, or several AE extensions. As we show below, any set of ordinary sentences has exactly one extension. The extension for the base set $A = \{P\}$ contains all the first-order consequences of P, but no other ordinary formulas. It contains modal atoms of the form $L\phi$, where ϕ is a FOC of A, and $\neg L\psi$, where ψ is not a FOC of A.

The base set $A = \{LP\}$ has no extensions. For suppose T is such an extension; either $P \in T$ or $P \notin T$. Clearly the latter cannot be the case, for then for *any* sentence ϕ,

[1] Moore [8] originally defined the concepts of soundness and completeness of a belief set relative to a base set, using AE valuations. AE extensions are an equivalent definition; see Konolige [4].

$A \models_T \phi$ (because $\models_{I,T} A$ is false for any I). Now suppose $P \in T$. In this case, we can construct an interpretation which satisfies A but falsifies P, namely one in which I makes P false. Therefore it cannot be that $A \models_T P$, and so P is not in T, a contradiction.

The base set $\{LP \supset P\}$ has two extensions, one of which contains P, and the other of which does not.

The base set $\{\neg LP \supset Q,\ \neg LQ \supset P\}$ has two extensions; in one of them, LP is true and LQ is not, and in the other the reverse.

Suppose an agent has only ordinary sentences in her base set A. This base set can be used to construct a belief set in an iterative fashion, starting with ordinary formulas and continually adding sentences with deeper nestings of modal operators.

Proposition 2.2 (Marek) *If A is a set of ordinary sentences, then it has exactly one AE extension T. T_0 is the first-order closure of A.*

We now turn our attention to an alternative characterization of AE extensions. Essentially, we seek to remove the self-referential index T in the implication operator of Definition 2.1. To do this, we introduce and analyze a special type of belief set, the *stable set*.

2.3 Stable sets

Following Stalnaker [10], we call a belief set Γ *stable* if it satisfies the following three properties:

1. Γ is closed under first-order consequence.[2]
2. If $\phi \in \Gamma$, then $L\phi \in \Gamma$.
3. If $\phi \notin \Gamma$, then $\neg L\phi \in \Gamma$.

The connection between stable sets and AE extensions is the following:

Proposition 2.3 (Moore) *Every AE extension of A is a stable set containing A.*

The strict converse of this proposition is not true, since there can be stable sets containing A which are not AE extensions of A. The simplest example is $A = \{LP\}$, which has no AE extension (see example 2.1). Yet there are many stable sets which contain LP.

A partial converse is available if we consider stable sets as AE extensions of their own ordinary sentences.

Proposition 2.4 *Every stable set Γ is an AE extension of Γ_0.*

[2] Stalnaker considered propositional languages and so used tautological consequence.

Stable sets are thus AE extensions of their ordinary sentences. From proposition 2.2, we know that every such AE extension is unique; hence every stable set is uniquely determined by its ordinary sentences.

Proposition 2.5 (Moore) *If two stable sets agree on ordinary formulas, they are equal.*

The set of ordinary formulas contained in a stable set is closed under first-order consequence. Different stable sets thus have different sets of FO-closed ordinary formulas. We now show that stable sets cover the sets of FO-closed ordinary formulas, that is, every such FO-closed set is the ordinary part of some stable set.

Proposition 2.6 *Let W be a set of ordinary formulas closed under first-order consequence. There is a unique stable set T such that $T_0 = W$. W is called the* kernel *of the stable set.*

We are now ready to give a second semantic characterization of AE extensions. Since AE extensions are stable, let us consider restricting the range of modal indices on the logical implication operator to just stable sets; we indicate this by $\models_{ss}$. From proposition 2.5, we know that the ordinary formulas of a stable set uniquely determine it. As usual, let T_0 be the set of ordinary formulas of T, and $\overline{T}_0$ the set of ordinary formulas *not* in T. Then, if T is stable, it must be the case that $\models_T$ is equivalent to $LT_0 \cup \neg L\overline{T}_0 \models_{ss}$, because LT_0 and $\neg L\overline{T}_0$ specify only those models in which the modal index is the unique stable set containing exactly the ordinary formulas T_0. This suggests how we can replace $\models_T$ in the definition of AE extensions.

Proposition 2.7 *T is an AE extension of A if and only if it satisfies the equation*

$$T = \{\phi \mid A \cup LT_0 \cup \neg L\overline{T}_0 \models_{ss} \phi\}\ .$$

By using a stronger type of implication ($=_{ss}$ over stable sets), we have been able to eliminate all self-referential assumptions except for those involving the ordinary formulas of T. This proposition also hints that the nesting of L-operators gives no extra expressive power to the language, since only ordinary formulas are important in characterizing the fixed point. This is indeed so, and we have the following proposition.

Proposition 2.8 *Every sentence in $\mathcal{L}$ is equivalent (under $\models_{ss}$) to a sentence whose modal atoms are of the form $L\phi$, with $\phi \in \mathcal{L}_0$.*

2.4 Strongly grounded extensions

One way of looking at the equation of proposition (2.7) is to see what type of reasoning it sanctions for reflective agents. An agent is justified in believing at least the consequences (under $\models_{ss}$) of her base set A, together with the assumptions LTo and $\neg L\overline{T}_0$. Moore has called belief sets defined in this way grounded in A, because they are derived from A and assumptions about self-belief.[3] However, this notion of groundedness is a fairly weak one, and we may wish to strengthen it. Consider, for example, the base set $A = \{LP \supset P\}$. A has two AE extensions, which we call T and T' (see example 2.1). T contains P and LP, while V does not contain P, but has $\neg LP$. The difference between these is precisely whether LP is introduced as an assumption in the fixed-point equation (2.7). For the belief set T, the agent's belief in P is grounded in her assumption that she believes P. If she chooses to believe P, she is justified in believing it precisely because she made it one of her beliefs. This certainly seems to be an anomolous situation, since the agent can, simply by choosing to assume that a fact about the world is true, be justified in that assumption without any objective information.

We would like to define a stronger notion of groundedness to eliminate this circularity of justifications. Now consider the belief set definition'given in proposition 2.7:

$$T = \{\phi \mid A \cup LT_0 \cup \neg L\overline{T}_0 \models_{ss} \phi\} .$$

The set of ordinary sentences in the belief set is To. LTO is the assumption that the agent believes all of these sentences. There would be no circular justifications if we replace LTo by LA in the fixed-point definition: we are assured that the derivation of facts about the world does not depend on the assumption of belief in those facts. The inclusion of LA is necessary because an ideally reflective agent should at least believe that her base beliefs are beliefs.

From this discussion, we define the following notion of strongly grounded.

Definition 2.2 A set of sentences T is strongly grounded in A if it obeys the constraint:

$$\Gamma \subseteq \{\phi \mid A \cup LA \cup \neg L\overline{\Gamma}_0 \models_{ss} \phi\} .$$

A certain natural class of AE extensions is strongly grounded, as we will shorty show. But not every AE extension is strongly grounded.

Example 2.9 The base set $A = \{LP \supset P\}$ has two extensions, only one of which is strongly grounded. The extension containing P cannot be strongly grounded, because P cannot be derived without the assumption of LP.

[3]Moore actually used a different but equivalent definition of groundedness; in his version, a set T is grounded in A if it satisfies:

$$\Gamma \subseteq \{\phi | A \cup L\Gamma \cup \neg L\overline{\Gamma} \models \phi .$$

A more complicated case is the base set $A = \{LP \supset Q, LQ \supset P\}$. Again there are two extensions, one containing the ordinary formulas P and Q, and one without them. For the former, LP and LQ must be assumed together in order to justify P and Q. Because they cannot be derived without this assumption, this extension is not strongly grounded.

The extension of a set of ordinary formulas A is strongly grounded, because every $\phi \in T_0$ is in the first-order closure of A, and so in the stable set containing LA.

Strongly grounded extensions are conservative in what they assume about the world, given the base beliefs. As shown in example 2.9, the base set $\{LP \supset P\}$ has only one strongly grounded extension, for which P is not a belief. In fact, strong groundedness is closely related to another concept, the minimality of ordinary sentences in an extension.

Definition 2.3 An AE extension T of A is minimal for A if there is no other extension T' of A such that $T'_0 \subset T_0$.

Minimal extensions always exist for a base set A that has extensions. Note that there can be more than one minimal extension for a given base set, e.g., $A = \{\neg LP \supset Q, \neg LQ \supset P\}$ has two extensions, both of which are minimal for A. The base set $A = \{LP \supset P\}$ has a single minimal extension, the one which doesn't contain P. Minimal extensions have a natural appeal as candidates for ideal reflective belief sets, because they limit the assumptions an agent makes about the world.

We now prove that, in fact, the minimal AE extensions of A are exactly the extensions strongly grounded in A. Thus we have two independent motivations for choosing these extensions as ideal belief sets.

Proposition 2.10 An AE extension of A is strongly grounded in A if and only if it is minimal. Strongly-grounded extensions obey the equation:

$$T = \{\phi \mid A \cup LA \cup \neg L\overline{T}_0 \models_{ss} \phi\} .$$

2.5 Normal form

The base sentences A of an AE extension can be put into a normal form that will be useful in the next section. We will use the following two facts about sentences of C in establishing a normal form.

1. Every AE sentence is equivalent to a sentence containing modal atoms only of the form Lo> or -Lo, where ϕ is an ordinary sentence.
2. $L\phi \wedge L\psi$ is equivalent to $L(\phi \wedge \psi)$.

These equivalences hold when considering interpretations whose modal indices are stable sets; see Konolige [4].

The first of these facts enables us to consider only base sets A with no nesting of modal operators. As we hinted in the last section, the nesting of L-operators lends no extra expressive power to the language.

In deriving a normal form for a set of sentences A, we first convert A to an equivalent set without nesting of modal operators, and then, using first-order valid operations, extract all modal atoms from the scope of quantifiers.

Proposition 2.11 Every set of L-scntcnces is equivalent (under $\models_{SS}$) to a set in which each sentence is of the form:

$$\neg L\alpha \vee L\beta_1 \vee \cdots \vee L\beta_n \vee \omega\,, \tag{10}$$

with α, β_i, and w all being ordinary sentences. Any of the disjunct*, except for ω, may be absent.

3 Default and AE extensions

In this section we briefly review default theories, and then present an effective syntactic translation of an arbitrary default theory W into a set of sentences W of AE logic. The main results of this paper are: (1) every default theory has a corresponding AE logic base set A whose minimal extensions are exactly the extensions of the default theory; and (2) every AE logic base set A has a corresponding default theory whose extensions are the minimal AE extensions of A. The translation between the two systems is effective and local, that is, each sentence or default rule is translated in isolation from the others.[4]

3.1 Default extensions

As defined by Reiter [9], a default theory is a pair (TV, D), where TV is a set of first-order sentences, and D is a set of defaults, each of which has the form:

$$\frac{\alpha : M\beta_1, M\beta_2, \ldots M\beta_n}{\omega}$$

A default d is satisfied by a set of sentences T if either (1) a is not in T or some $\neg\beta_i$ is in T (the premisses of the rule are not satisfied), or (2) ω is in T (the conclusion is satisfied). A default extension of (TV, .D), informally, is a minimal set of sentences containing TV, closed under first-order consequence, and satisfying all the defaults D.

If none of α, β_i, or ω contain free variables, then the default is called closed. An open default is treated as a schema for the set of closed defaults that are its substitution instances. We thus need only consider closed defaults, as long as we allow default theories to contain a denumberably infinite set of them.

Default extensions have many of the same properties as AE extensions. There may be one, no, or many extensions of a default theory. The following examples are analogous to the AE extensions in example 2.1.

[4] Imielinski [3] defines the weaker notion of a modular translation: the defaults and first-order parts must be translated independently. Obviously, any local translation is modular.

Example 3.1 The default extension for the theory $(\{P\}, \emptyset)$ (no defaults) is exactly the first-order consequences of P.

The theory $\langle \emptyset, P : /P \rangle$ has one extension, the set of all first-order valid sentences. P is not an element of this extension. This differs from the case of AE extensions for $\{LP \supset P\}$; there is an extension which contains P.

The theory $\langle \emptyset, \{M\neg P/Q,\ M\neg Q/P\} \rangle$ has two extensions; in one of them, P is true and Q is not, and in the other the reverse.

These examples are instructive by comparison to AE extensions. If the theory (W,D) contains no defaults (D empty), then there is exactly one extension, which is the first-order part of the AE extension of W. In general, a default of the form $\alpha : M\beta/\omega$ corresponds to the AE sentence $L\alpha \wedge \neg L\neg\beta \supset \omega$; thus, in the third default theory of the example, there are two default extensions, corresponding to the first-order parts of the two AE extensions of $\{\neg LP \supset Q,\ \neg LQ \supset P\}$. However, note the difference in the case of the second default theory of this example. The default P : $/P$ has only one extension, in which P does not appear. The AE set $\{LP \supset P\}$ has two extensions; the one in which P appears arises from the ability of AE extensions to support circular justifications (assuming LP, the sentence LP $\supset$ P gives a derivation of P). So although it appears that default extensions have corresponding AE extensions for a suitable transformation of the defaults, not all AE extensions will have corresponding default extensions. In fact, as we show below, default extensions correspond to minimal AE extensions.

3.2 Defaults as self-belief

We now define a simple transformation from a default theory (TV, D) to a set of AE sentences A, such that the default extensions of (TV, D) are exactly the kernels (the first-order part) of the minimal AE extensions of A. Thus (as we prove), there is an exact correspondence between default extensions for (TV, D) and minimal AE extensions for A.

The transformation is:

$$\frac{\alpha : M\beta_1 \ldots M\beta_n}{\omega} \quad \mapsto \quad (L\alpha \wedge \neg L\neg\beta_1 \wedge \cdots \wedge \neg L\neg\beta_n) \supset \omega \tag{11}$$

As we mentioned in the introduction, this is the natural interpretation of defaults in terms of introspective knowledge. A paraphrase of the AE sentence for agent would be something like the following: "If I know that a is true, and I have no knowledge that any of the β_i are false, then ω must be true." The key phrase has been emphasized; it is in reasoning about what is not known that the nonmonotonic character of AE logic appears. However, the role of the other parts of the sentence ($L\alpha$ and ω) also deserves closer scrutiny; for example, why does w appear as the consequent, and not $L\omega$? As it stands, this is the transformation that yields the correspondence between default

and AE extensions. We will comment more extensively on the form transformation later, after the basic results are presented.

In a default, we allow either α or any of the $M\beta_i$ to be missing; the corresponding AE sentence just deletes the appropriate conjunct in the antecedent. The conclusion of the default must always be present (defaults with no conclusion are senseless). Let D' be the set of sentences formed by taking the transforms of defaults P; we call the set {W,D'} the AE transform of(W,D).

Default extensions are the fixed points of an operator r(V). This operator is meant to formalize the informal criteria given above for the extensions of (W, D), namely, it should contain W, be closed under first-order consequence, and satisfy all of D. Let V be an arbitrary set of first-order sentences. Then T(V) is the smallest set satisfying the following properties:

D1. $W \subseteq \Gamma(V)$

D2. T(V) is closed under first-order consequence.[5]

D3. If $\alpha : M\beta/\omega \in D$, $\alpha \in \Gamma(V)$, and $\neg\beta \notin V$, then $\omega \in$ T(V).

Extensions are fixed-points of T, i.e., any set E satisfying E — T(E). As a fixed-point definition, it is similar to the fixed-point account of minimal AE extensions (proposition 2.10). The parameter of T(V) essentially fills the role of the assumptions $\neg L\bar{T}_0$, since -B must not be present in order for the default to be satisfied. Minimality is part of the definition of T(V) (the least set satisfying the conditions D1-D3); if it were excluded, then default extensions corresponding to non-minimal AE extensions would be present.

Now consider a particular default theory (W, D} and an associated extension E = T(E). E is closed under first-order consequence, and hence is the kernel of a unique stable set. This stable set is closely related to the AE transform of (W,D): it is a minimal stable set containing the AE transform. We prove this result as the following proposition.

Proposition 3.2 Let (W,D) be a default theory, with A = {W, D'} its AE transform. Suppose E is an extension of the default theory. Then E is the kernel of a minimal stable set containing A and $\neg L\bar{E}$.

Using this result, we can show that a default theory and its AE transform have the same extensions.

Theorem 3.3 Let A be the AE transform of a default theory Δ. A set E is a default extension of Δ if and only if it is the kernel of a minimal AE extension of A.

*In the original definition, this is stated in terms of deduction rather than logical consequence.

3.3 Semantics

The semantics of AE sentences is an interpretive semantics, in the sense that a sentence ϕ is true or false in an interpretation $\models_{I,\Gamma}$. The interpretation of modal atoms is given by the modal index T according to equation 8. The interpretations themselves are straightforward augmentations of standard first-order interpretations. The problematic characteristics of AE logic, from semantical point of view, occur in the fixed-point definition of extensions (2.1), in which only interpretations containing a certain modal index are considered. So, although it is hard to construct and analyze extensions, all of our ordinary intuitions about the meaning of the language £, its semantics with respect to individual interpretations, is still available.

To give an example of this sort: consider the difference between the two default sentences

$$\text{LBird\{Tweety)} \wedge \neg L\neg Fly(Tweety) \supset Fly(Tweety) \quad (12)$$

and

$$\text{Bird(Tweety)} \wedge \neg L\neg Fly(Tweety) \supset Fly(Tweety) \quad (13)$$

The first of these states that in any interpretation in which Bird(Tweety) is a belief, and -Fly(Tweety) is not a belief, Fly(Tweety) will be true. The antecedent of the second default is less strict: it states only that Bird(Tweety) must be true. The second default permits case analysis of a type not sanctioned by the first. For example, suppose it is known that either Tweety is a bird, or that Tweety is housebroken (Houseb(Tweety)). In every interpretation in which $\neg Fly(Tweety)$ is not a belief, and the second default sentence is true, Houseb(Tweety) ∨ Fly (Tweety) is true. On the other hand, nothing can be concluded by assuming the first default sentence is true, because Bird(Tweety) may not be a belief. As Etherington [1, p. 34] has noted, the second sentence seems more in accord with our intuitions about the way defaults should work.

Another example of the utility of interpretive semantics is in the concepts of equivalence and substitution. Two formulas ϕ and ϕ' of $\mathcal{L}$ are equivalent if they have the same truthvalue in all models. Because the definition of AE extensions is framed in terms of the interpretive semantics, ϕ' can be substituted anywhere ϕ occurs in a base set A, without changing the AE extensions of A. We used this fact extensively in arriving at the normal form for AE sentences in section 2.5.

3.4 Expressiveness

The question of expressiveness can be phrased as follows: Is it the case that default sentences of the type (13), or perhaps other AE sentences involving complicated constructions such as embedded L-operators, have no counterpart in default theories? On the face of it this would seem a plausible conjecture, since the L-operator is part of the language, while default rules are not. However, it turns

out that AE logic is no more expressive than default logic: there is an effective transformation of any base set of AE sentences into a default theory, such that the default extensions are exactly the kernels of the minimal AE extensions. To show this, we rely on the fact (see proposition 2.11) that every set of sentences of *C* has an equivalent normal form in which every sentence looks like:

$$\neg L\alpha \vee L\beta_1 \vee \cdots \vee L\beta_n \vee \omega\ , \tag{14}$$

where all of α, β_i, and w are ordinary sentences, ω is always present, and any of the modal atoms may be missing,.

Given any set of L-sentences *A* in normal form, it is possible to effectively construct a corresponding default theory (*W,* D), in the following way. Any ω that appears without other disjuncts is put into *W.* All other sentences are transformed into defaults, in the manner indicated by equation 11. It is easy to see that *A* is the AE transform of *{W, D);* by theorem 3.3, these two have essentially the same extensions. More precisely, we have proven the following theorem:

Theorem 3.4 *For any set of sentences A of C, there is an effectively constructable default theory (W, D) such that E is a default extension of (W, D) if and only if it is the kernel of a minimal extension of A.*

So, suprisingly, default theories have precisely the same expressiveness as AE logic over the modal language *C.* However, two caveats should be noted.

The first is that the expression of various statements about defaults or autoepistemic reasoning may be much more natural in L, because the form of sentences is much less constrained than that of the default inference rules. For example, the second type of default (equation 13) is translated into the default rule:

$$\frac{M\,Fly(Tweety)}{Bird(Tweety) \supset Fly(Tweety)}\ . \tag{15}$$

The atom *Bird(Tweety)* does not appear in the antecedent of the default, but somewhat unnaturally in the consequent.

The second caveat is that, if we extend *C* by allowing quantifying-in (i.e., expressions such as $\exists x.L\phi(x)$), in all likelihood theorem 3.4 will no longer hold. There are a number of reasons to think this; perhaps the most compelling is Levesque's observation [5] that in the presence of quantifying-in, there are sentences with nested belief operators that cannot be reduced to sentences without them.

4 Conclusion

Given the current proliferation of nonmonotonic formalisms, it seems wise to establish comparisons among them, especially regarding expressiveness. The results presented here show that there is an exact correspondence between AE logic over *C* and default theories. There is an effective, local translation between the two that preserves theoremhood, in that the default extensions are the first-order part of the minimal AE extensions.

5 Acknowledgements

Many thanks to David Etherington, Joseph Halpern, Hector Levesque, Karen Myers, and Raymond Reiter for reading and commenting on drafts of this paper.

References

[1] Etherington, D. W. *Reasoning with Incomplete Information: Investigations of Non-Monotonic Reasoning.* PhD thesis, University of British Columbia, Vancouver, British Columbia, 1986.

[2] Hintikka, J. Impossible possible worlds vindicated. *Journal of Philosophical Logic,* 4:475-484, 1975.

[3] Imielinski, T. Results on translating defaults to circumscription. In *Proceedings of the International Joint Conference on Artificial Intelligence,* pages 114-120, Los Angeles, 1985.

[4] Konolige, K. *On the Relation between Default Logic and Autoepistemic Theories,* forthcoming Technical Note, SRI Artificial Intelligence Center, Menlo Park, California, 1987.

[5] Levesque, H. J. *A Formal Treatment of Incomplete Knowledge Bases.* Technical Report 614, Fairchild Artificial Intelligence Laboratory, Palo Alto, California, 1982.

[6] Lukaszewicz, W. Two results on default logic. In *Proceedings of the American Association of Artificial Intelligence,* pages 459-461, University of California at Los Angeles, 1985.

[7] McCarthy, J. Circumscription — a form of nonmonotonic reasoning. *Artificial Intelligence,* 13(1-2), 1980.

[8] Moore, R. C. Semantical considerations on nonmonotonic logic. *Artificial Intelligence,* 25(1), 1985.

[9] Reiter, R. A logic for default reasoning. *Artificial Intelligence,* 13(1-2), 1980.

[10] Stalnaker, R. C. A note on nonmonotonic modal logic. 1980. Department of Philosophy, Cornell University.