Autoepistemic Logic and Introspective Circumscription

Michael Gelfond
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968
mgelfond@cs.ep.utexas.edu

Vladimir Lifschitz
Department of Computer Sciences
and Department of Philosophy
University of Texas at Austin
Austin, TX 78712
vl@cs.utexas.edu

Halina Przymusinska
Computer Science Department
California State Polytechnic University
Pomona, CA 91768
halina@ucrengr.ucr.edu

Grigori Schwarz
Computer Science Department
Stanford University
Stanford, CA 94305
schwarz@flamingo.stanford.edu

Abstract

We investigate the relationship between two epistemic nonmonotonic formalisms: autoepistemic logic and introspective circumscription. Finitely axiomatized autoepistemic theories are shown to be equivalent to the propositional case of introspective circumscription. This theorem is applied to the problem of relating the usual "minimizing" circumscription to autoepistemic logic.

1 Introduction

The aim of this paper is to investigate the relationship between two epistemic nonmonotonic formalisms: autoepistemic logic [Moore, 1985] and introspective circumscription [Lifschitz, 1989].

In recent years, we have seen renewed interest in autoepistemic logic, due primarily to its interaction with the concerns of the logic programming community.

Defining a declarative semantics of *negation as failure* has been long considered an important problem in the theory of logic programming. The mathematical apparatus used earlier for this purpose includes predicate completion [Clark, 1978] and iterated fixed points [Apt *et al.*, 1988]. A simple alternative proposed in [Gelfond, 1987] was to view negation as failure as the combination $\neg\mathsf{B}$, where B is the belief operator of autoepistemic logic. From this perspective, logic programs are merely a special kind of autoepistemic theories.

The epistemic approach to logic programming has led to the notion of a "stable model" [Gelfond and Lifschitz, 1988] and to the "answer set semantics" for logic programs [Gelfond and Lifschitz, 1991]. A modification of the method of [Gelfond, 1987] was shown to be applicable to logic programs of a very general form—to disjunctive programs that may include both negation as failure and classical negation [Lifschitz and Schwarz, 1993].

Introspective circumscription, like autoepistemic logic, formalizes the idea of introspection, but in a very different way. Just as McCarthy's original "minimizing" form of circumscription, it is not really a nonmonotonic logic, but rather a syntactic transformation of classical formulas. Introspective circumscription is not restricted to the propositional case—its definition applies to formulas with quantifiers and equality, and even higher-order quantification does not require any special treatment. For comparison, note that the available approaches to the problem of incorporating quantifiers in autoepistemic logic—[Konolige, 1989] and [Levesque, 1990]—tackle the difficult issue of "quantifying-in" in different ways, and there are varying opinions on what the "correct" definition of predicate autoepistemic logic should look like.

Thus, technically, the two systems appear to be quite different, and introspective circumscription, the younger and less known of the two, may have important advantages. The ease with which it handles quantification and equality is, in particular, of interest to logic programming. Since autoepistemic logic is propositional, the replacement of negation as failure by $\neg\mathsf{B}$ has to be combined with replacing each rule in the program by its ground instances. This step, sometimes undesirable, is unnecessary if introspective circumscription is used instead of autoepistemic logic.

The semantics of each of the two formalisms can be characterized by a fixpoint construction, but the operators involved in the constructions are of different kinds. In the case of introspective circumscription, the fixpoints are interpretations—Boolean vectors, if the circumscribed formula is propositional. In autoepistemic logic, the fixpoints are *sets* of interpretations (or, in another approach, sets of formulas).

The results of this paper show, however, that finitely axiomatized autoepistemic theories are essentially identical to the propositional case of introspective circumscription. The main theorem describes a relationship between the models of a theory in the sense of autoepistemic logic and the models of the corresponding introspective circumscription. As a corollary, we obtain a simple correspondence between the consequence relations of the two nonmonotonic formalisms. This correspondence was first described in the preliminary version of this note (authored by Gelfond and Przymusinska) and proved by a different method, not based on the use of models. A special case was proved in [Lifschitz, 1989], Section 5.3; the possibility of a generalization is mentioned there, but the exact statement has remained unpublished until now.

Since minimizing circumscription is a special case of the introspective version, our theorem can be applied to the problem of relating minimizing circumscription to autoepistemic logic. In this way, we prove results similar to those established by Konolige [1989].

2 Review of Autoepistemic Logic

The treatment of autoepistemic logic in this note follows [Lifschitz and Schwarz, 1993], and is based on the ideas of [Moore, 1984], [Levesque, 1990] and [Schwarz, 1992].

Autoepistemic formulas are built from (propositional) atoms using propositional connectives and the modal operator B. An *interpretation* is a set of atoms. A *structure* is a pair (I, S), where I is an interpretation, and S a set of interpretations. The satisfaction relation between a structure and an autoepistemic formula is defined inductively, as follows. For an atom F, $(I, S) \models F$ iff $F \in I$. For any formula F, $(I, S) \models \mathsf{B}F$ iff, for every $J \in S$, $(J, S) \models F$. The propositional connectives are handled in the usual way.

If F is nonmodal, the relation $(I, S) \models F$ obviously turns into the usual satisfaction relation of classical logic, $I \models F$. If F is atomic, then $(I, S) \models \mathsf{B}F$ iff F belongs to $\bigcap_{J \in S} J$. We will denote this set by $\cap S$.

A *theory* is a set of autoepistemic formulas, *axioms*. A set S of interpretations is an *autoepistemic model* of a theory T if it satisfies the equation

$$S = \{I : \text{for each } F \in T,\ (I, S) \models F\}. \tag{1}$$

An autoepistemic formula F is *entailed* by T if, for every autoepistemic model S of T and for every $I \in S$, $(I, S) \models F$.

Example 1. Let T be $\{\mathsf{B}P_1 \vee P_2\}$, where P_1 and P_2 are atoms. Since this axiom can be written also as $\neg \mathsf{B}P_1 \supset P_2$, T is the counterpart of the logic program $\{P_2 \leftarrow \mathit{not}\ P_1\}$ under the correspondence defined in [Gelfond, 1987]. By the definition of satisfaction, a structure (I, S) satisfies $\mathsf{B}P_1 \vee P_2$ when $P_1 \in \cap S$ or $P_2 \in I$. Thus (1) can be written in this case as

$$S = \{I : P_1 \in \cap S \text{ or } P_2 \in I\}.$$

In order to find all solutions, note that, for any S, the right-hand side of this equation is the set of all interpretations if $P_1 \in \cap S$, and $\{I : P_2 \in I\}$ otherwise.

Hence these two sets—let's call them S_1 and S_2—are the only possible solutions. Moreover, $\cap S_1 = \emptyset$ and $\cap S_2 = \{P_2\}$; neither set includes P_1. It follows that S_2 is an autoepistemic model of T, and S_1 is not. Consequently, an autoepistemic formula F is entailed by T iff, for every $I \in S_2$, (I, S_2) satisfies F. For instance, T entails P_2, $\neg \mathsf{B} P_1$ and $\mathsf{B} P_2$. It is clear that a nonmodal formula is entailed by T iff it is entailed by P_2 in classical logic.

Example 2. Let T be $\{P_1 \vee P_2, P_1 \supset \mathsf{B} P_1, P_2 \equiv \mathsf{B} P_2\}$. As will be seen later, this theory corresponds to circumscribing P_1, with P_2 fixed, in the formula $P_1 \vee P_2$. For any S, the right-hand side of (1) is one of 4 sets, depending on which of the following cases tales place. If $P_1, P_2 \in \cap S$, then the right-hand side of (1) is $\{I : P_2 \in I\}$. If $P_1 \in \cap S$ and $P_2 \notin \cap S$, then it is $\{I : P_1 \in I, P_2 \notin I\}$. If $P_1 \notin \cap S$ and $P_2 \in \cap S$, then it is $\{I : P_1 \notin I, P_2 \in I\}$. Finally, if $P_1, P_2 \notin \cap S$, then it is $\emptyset$. Two of these sets, $\{I : P_1 \in I, P_2 \notin I\}$ and $\{I : P_1 \notin I, P_2 \in I\}$, are the models of T. Some of the formulas that are satisfied by both models are $P_1 \vee P_2$, $\neg P_1 \vee \neg P_2$, $BP_1 \equiv \neg P_2$, $BP_2 \equiv P_2$. It follows that these formulas are entailed by T.

This model-theoretic presentation of autoepistemic logic is equivalent to its traditional description in terms of "stable expansions." A set E of autoepistemic formulas is said to be a *stable expansion* of T if the set of formulas derivable in classical propositional logic from

$$T \cup \{\mathsf{B} G : G \in E\} \cup \{\neg \mathsf{B} G : G \notin E\}$$

equals E [Moore, 1985]. The relationship between stable expansions and autoepistemic models is quite simple: E is a stable expansion of T iff

$$E = \{F : \text{for each } I \in S,\ (I, S) \models F\}$$

for some autoepistemic model S of T. A proof can be found in [Schwarz, 1992].

3 The Nonmodal Counterpart of an Autoepistemic Formula

In the rest of the paper, we concentrate on autoepistemic theories with finitely many axioms. Without loss of generality, we can restrict attention to the formulas in which B is applied to atoms only; indeed, other occurrences of B can be eliminated by introducing explicit definitions ([Schwarz and Truszczyński, 1992], Theorem 4.2). For instance, the problem of computing the models of

$$\{P_1 \vee \mathsf{B} \neg P_2\} \tag{2}$$

can be reduced to the same problem for the theory

$$\{P_1 \vee \mathsf{B} P_3, P_3 \equiv \neg P_2\}.$$

In order to determine whether (2) entails $\neg\mathsf{B}(P_1 \wedge P_2)$, we can check whether

$$\{P_1 \vee \mathsf{B}P_3, P_3 \equiv \neg P_2, P_4 \equiv (P_1 \wedge P_2)\}$$

entails $\neg\mathsf{B}P_4$.

Thus, from now on, autoepistemic formulas will be assumed to be propositional combinations of the formulas $P_1, \dots, P_n, \mathsf{B}P_1, \dots, \mathsf{B}P_n$, where $P_1, \dots, P_n$ are atoms. If we abbreviate the first half of this list by P, and the second half by $\mathsf{B}P$, then any autoepistemic formula of this kind can be written as $F(P, \mathsf{B}P)$, with all occurrences of B explicitly shown.

The *nonmodal counterpart* of an autoepistemic formula $F(P, \mathsf{B}P)$ is obtained from it by substituting new propositional atoms $BP_1, \dots, BP_n$ for all occurences of the subformulas $\mathsf{B}P_1, \dots, \mathsf{B}P_n$. For instance, the nonmodal counterpart of the axiom from Example 1 is $BP_1 \vee P_2$. If we denote the list of new propositional atoms by BP, then the nonmodal counterpart of $F(P, \mathsf{B}P)$ can be written as $F(P, BP)$.

4 Review of Introspective Circumscription

Introspective circumscription is a further development of the notion of "autocircumscription" due to Perlis [1988]. As defined in [Lifschitz, 1989], it is a syntactic transformation of second-order formulas. Since we are interested here in the propositional case of introspective circumscription, the only quantifiers that are needed are quantifiers over propositional variables. Such quantifiers can be eliminated in favor of propositional connectives:

$$\forall p_i F(p_i) \equiv F(\mathit{false}) \wedge F(\mathit{true}).$$

Consider a propositional combination $A(P, BP)$ of the atoms P, BP. The *introspective circumscription operator* transforms this formula into the formula

$$A(P, BP) \wedge \bigwedge_i [BP_i \equiv \forall p(A(p, BP) \supset p_i)],$$

where p is a tuple of distinct propositional variables $p_1, \dots, p_n$. We will denote this formula by $A^*(P, BP)$.

Example 3. Consider the nonmodal counterpart of the axiom from Example 1. The introspective circumscription $(BP_1 \vee P_2)^*$ is

$$(BP_1 \vee P_2) \wedge [BP_1 \equiv \forall p_1 p_2((BP_1 \vee p_2) \supset p_1)] \wedge [BP_2 \equiv \forall p_1 p_2((BP_1 \vee p_2) \supset p_2)].$$

Having eliminated the quantifiers as described above, we determine that the part $\forall p_1 p_2((BP_1 \vee p_2) \supset p_1)$ is equivalent to *false*, and the part $\forall p_1 p_2((BP_1 \vee p_2) \supset p_2)$ is equivalent to $\neg BP_1$. Consequently, $(BP_1 \vee P_2)^*$ can be rewritten as

$$(BP_1 \vee P_2) \wedge \neg BP_1 \wedge (BP_2 \equiv \neg BP_1),$$

or

$$P_2 \wedge \neg BP_1 \wedge BP_2.$$

Example 4. Consider the nonmodal counterpart of the conjunction of the axioms from Example 2. The introspective circumscription

$$((P_1 \vee P_2) \wedge (P_1 \supset BP_1) \wedge (P_2 \equiv BP_2))^*$$

can be simplified in a similar way, and the final result can be written as

$$(P_1 \vee P_2) \wedge (\neg P_1 \vee \neg P_2) \wedge (BP_1 \equiv \neg P_2) \wedge (BP_2 \equiv P_2).$$

5 Main Theorem

Since a finite axiom set can be replaced by the conjunction of the axioms, we can restrict attention to autoepistemic theories with a single axiom. We will identify such a theory, $\{A(P, \mathsf{B}P)\}$, with its axiom $A(P, \mathsf{B}P)$.

The main theorem asserts that the models of an introspective circumscription $A^*(P, BP)$ (that is, the interpretations satisfying this formula) are in a one-one correspondence with the structures (I, S) such that S is an autoepistemic model of $A(P, \mathsf{B}P)$ and $I \in S$.

The definition of this correspondence f is quite simple. For any set S of interpretations, let S^b stand for $\{BP_i \ : \ P_i \in \cap S\}$. For any structure (I, S), we define: $f(I, S) = I \cup S^b$.

There is a close connection between this function f and the notion of the nonmodal counterpart introduced above: For any autoepistemic formula $F(P, \mathsf{B}P)$ and any structure (I, S),

$$(I, S) \models F(P, \mathsf{B}P) \text{ iff } f(I, S) \models F(P, BP). \tag{3}$$

This is easy to verify by structural induction on F.

Main Theorem. *For any autoepistemic formula $A(P, \mathsf{B}P)$, f maps the set of structures (I, S) such that S is an autoepistemic model of $A(P, \mathsf{B}P)$ and $I \in S$ onto the set of models of $A^*(P, BP)$. Moreover, f is one-one on the structures (I, S) such that S is an autoepistemic model of $A(P, \mathsf{B}P)$.*

For instance, according to Example 1, the only autoepistemic model of $\mathsf{B}P_1 \vee P_2$ is $\{I : P_2 \in I\}$, that is, $\{\{P_2\}, \{P_1, P_2\}\}$. We compute:

$$f(\{P_2\}, \{\{P_2\}, \{P_1, P_2\}\}) = \{P_2, BP_2\};$$
$$f(\{P_1, P_2\}, \{\{P_2\}, \{P_1, P_2\}\}) = \{P_1, P_2, BP_2\}.$$

These interpretations are the models of the circumscription $(BP_1 \vee P_2)^*$ that was computed in Example 3.

The proof of the theorem is given in Section 7.

The theorem immediately leads to a correspondence between the autoepistemic models of $A(P, \mathsf{B}P)$ and the models of $\exists p A^*(p, BP)$ (that is, the subsets of BP satisfying this formula).

Corollary 1. *For any autoepistemic formula* $A(P, \mathsf{B}P)$, $S \mapsto S^b$ *is a one-one map of the set of nonempty autoepistemic models of* $A(P, \mathsf{B}P)$ *onto the set of models of* $\exists p A^*(p, BP)$.

Proof. Take a nonempty autoepistemic model S of $A(P, \mathsf{B}P)$, and let $I \in S$. By the theorem, $I \cup S^b$ is a model of $A^*(P, BP)$. Consequently, S^b is a model of $\exists p A^*(p, BP)$. Now take a model U of $\exists p A^*(p, BP)$, and let I be a subset of P such that $I \cup U$ is a model of $A^*(P, BP)$. By the theorem, there exist an autoepistemic model S of $A(P, \mathsf{B}P)$ and an interpretation $J \in S$ such that $J \cup S^b = I \cup U$. By intersecting both sides of this equality with BP, we conclude that $S^b = U$. Finally, $S \mapsto S^b$ is one-one on the autoepistemic models of $A(P, \mathsf{B}P)$ because, for any I, S^b can be obtained by intersecting $f(I, S)$ with BP.

Corollary 2. *An autoepistemic formula* $F(P, \mathsf{B}P)$ *is entailed by an autoepistemic formula* $A(P, \mathsf{B}P)$ *iff* $F(P, BP)$ *is entailed by* $A^*(P, BP)$ *in classical logic.*

Proof. According to the theorem, $F(P, BP)$ is entailed by $A^*(P, \mathsf{B}P)$ iff, for every autoepistemic model S of $A(P, \mathsf{B}P)$ and every $I \in S$, $f(I, S) \models F(P, BP)$; by (3), the last formula is equivalent to $(I, S) \models F(P, \mathsf{B}P)$.

We have seen, for instance, that the theory of Example 1 entails P_2, $\neg \mathsf{B}P_1$ and $\mathsf{B}P_2$, and that the corresponding introspective circumscription (Example 3) is the conjunction of these formulas. We have seen also that the theory of Example 2 entails, among others, the formulas $P_1 \vee P_2$, $\neg P_1 \vee \neg P_2$, $BP_1 \equiv \neg P_2$ and $BP_2 \equiv P_2$; those are the conjunctive terms of the formula obtained in Example 4.

6 Minimizing Circumscription

As shown in [Lifschitz, 1989], Section 4, introspective circumscription includes many forms of the traditional minimizing circumscription as special cases. Consider for instance, a propositional formula $A(P)$, and assume that we wish to circumscribe some of the propositional symbols P—for instance, $P_1, \ldots, P_k$—in parallel, with some of the other propositional symbols—for instance, $P_{k+1}, \ldots, P_l$—fixed, and the remaining propositional symbols $P_{l+1}, \ldots, P_n$ allowed to vary. The result of this circumscription will be denoted by $A^c(P)$.

The minimizing circumscription $A^c(P)$ is closely related to the introspective circumscription of the conjunction

$$A(P) \wedge (P_1 \supset BP_1) \wedge \ldots \wedge (P_k \supset BP_k) \wedge (P_{k+1} \equiv BP_{k+1}) \wedge \ldots \wedge (P_l \equiv BP_l).$$

Specifically, if this formula is denoted by $\tilde{A}(P, BP)$, then, according to Proposition 4 from [Lifschitz, 1989], $\tilde{A}^*(P, BP)$ can be written as the conjunction of $A^c(P)$ with explicit definitions of the propositional symbols BP in terms of the remaining propositional symbols P.

For instance, circumscribing P_1, with P_2 fixed, in $P_1 \vee P_2$, corresponds to the case when $k = 1$, $l = n = 2$, and $A(P_1, P_2)$ is $P_1 \vee P_2$. $A^c(P_1, P_2)$ can be written as $(P_1 \vee P_2) \wedge (\neg P_1 \vee \neg P_2)$. Accordingly, the formula obtained in Example 4 is the result of adding to this conjunction two more terms—explicit definitions of BP_1 and BP_2.

Using this fact, we conclude from Corollary 2:

Corollary 3. *A nonmodal formula is entailed by*

$$\{A(P), P_1 \supset \mathsf{B}P_1, \ldots, P_k \supset \mathsf{B}P_k, P_{k+1} \equiv \mathsf{B}P_{k+1}, \ldots, P_l \equiv \mathsf{B}P_l\}$$

iff it is entailed by $A^c(P)$ in classical logic.

For instance, a nonmodal formula is entailed by the theory from Example 4 iff it is entailed by the circumscription of P_1, with P_2 fixed, in $P_1 \vee P_2$.

This corollary is similar to Theorem 3.5 from [Konolige, 1989]. Our representation of the fixed symbols is somewhat different from his. Also, Konolige's theorem is about a first-order extension of autoepistemic logic and thus is not restricted to the propositional case.

Similar results can be established for prioritized circumscription. Consider, for instance, a propositional formula $A(P_1, P_2)$ with just 2 atoms. We would like to minimize P_1 and P_2, with P_1 given a higher priority; let the result of this circumscription be $A^{pc}(P_1, P_2)$. According to Proposition 5 from [Lifschitz, 1989], the introspective circumscription of

$$A(P_1, P_2) \wedge (P_1 \supset BP_1) \wedge ((P_1 \equiv BP_1) \supset (P_2 \supset BP_2))$$

is equivalent to

$$A^{pc}(P_1, P_2) \wedge (BP_1 \equiv P_1) \wedge (BP_2 \equiv P_2).$$

It follows that a nonmodal formula is entailed by the theory

$$\{A(P_1, P_2), P_1 \supset \mathsf{B}P_1, (P_1 \equiv \mathsf{B}P_1) \supset (P_2 \supset \mathsf{B}P_2)\}$$

iff it is entailed by $A^{pc}(P_1, P_2)$ in classical logic.

7 Proof of Main Theorem

For any subset U of BP, by U^A we will denote the set of all subsets I of P that satisfy the condition

$$I \cup U \models A(P, BP).$$

It is easy to see that

$$\cap U^A = \{P_i \ : \ U \models \forall p(A(p, BP) \supset p_i)\}. \tag{4}$$

Indeed, $P_i \in \cap U^A$ iff, for all I such that $I \cup U \models A(P, BP)$, $P_i \in I$; the last inclusion can be written as $I \cup U \models P_i$.

Lemma. *If $U \models \exists p A^*(p, BP)$ then U^A is an autoepistemic model of $A(P, \mathsf{B}P)$, and $(U^A)^b = U$.*

Proof. Assume that $U \models \exists p A^*(p, BP)$. We will first prove that $(U^A)^b = U$. It is clear from the definition of introspective circumscription that $\exists p A^*(p, BP)$ entails $BP_i \equiv \forall p(A(p, BP) \supset p_i)$. Consequently,

$$U \models BP_i \equiv \forall p(A(p, BP) \supset p_i),$$

that is,

$$U = \{BP_i \ : \ U \models \forall p(A(p, BP) \supset p_i)\}.$$

By (4), the right-hand side equals $\{BP_i \ : \ P_i \in \cap U^A\}$, that is, $(U^A)^b$. In order to show that U^A is an autoepistemic model of $A(P, \mathsf{B}P)$, note first that, by (3),

$$\{I \ : \ (I, U^A) \models A(P, \mathsf{B}P)\} = \{I \ : \ f(I, U^A) \models A(P, BP)\}.$$

Since

$$f(I, U^A) = I \cup (U^A)^b = I \cup U,$$

the right-hand side equals U^A.

Main Theorem. *For any autoepistemic formula $A(P, \mathsf{B}P)$, f maps the set of structures (I, S) such that S is an autoepistemic model of $A(P, \mathsf{B}P)$ and $I \in S$ onto the set of models of $A^*(P, BP)$. Moreover, f is one-one on the structures (I, S) such that S is an autoepistemic model of $A(P, \mathsf{B}P)$.*

Proof. Assume that S is an autoepistemic model of $A(P, \mathsf{B}P)$ and $I \in S$. We want to show that $f(I, S)$ satisfies $A^*(P, BP)$. By the definition of an autoepistemic model, $(I, S) \models A(P, \mathsf{B}P)$; by (3), it follows that $f(I, S) \models A(P, BP)$. It remains to prove that

$$f(I, S) \models BP_i \equiv \forall p(A(p, BP) \supset p_i).$$

Since

$$f(I, S) = I \cup S^b = I \cup \{BP_i \ : \ P_i \in \cap S\},$$

this can be expressed as follows: $P_i \in \cap S$ iff $S^b \models \forall p(A(p, BP) \supset p_i)$. Assume that $P_i \in \cap S$, and let J be any subset of P such that $J \cup S^b \models A(P, BP)$. Since $J \cup S^b = f(J, S)$, it follows from (3) that $(J, S) \models A(P, \mathsf{B}P)$. Because S is an autoepistemic model of $A(P, \mathsf{B}P)$, we conclude that $J \in S$ and $P_i \in \cap S \subset J$. Thus $S^b \models \forall p(A(p, BP) \supset p_i)$. Assume now $S^b \models \forall p(A(p, BP) \supset p_i)$. We claim that $P_i \in \cap S$. Consider an interpretation $J \in S$. Because S is an autoepistemic model of $A(P, \mathsf{B}P)$, we have $(J, S) \models A(P, \mathsf{B}P)$. By (3), it follows that $f(J, S) \models A(P, BP)$. Hence $f(J, S) \models P_i$, which means, by the definition of f, that $P_i \in J$. Since J is an arbitrary element of S, we have proved that $P_i \in \cap S$.

The next step is to show that every model of $A^*(P, BP)$ can be represented in the form $f(I, S)$, where S is an autoepistemic model of $A(P, \mathsf{B}P)$ and $I \in S$. Consider any model $I \cup U$ of $A^*(P, BP)$, where I is a subset of P and U a subset of

BP. We claim that U^A can be taken as S. Since $I \cup U$ satisfies $A(P, BP)$, it follows from the definition of U^A that $I \in U^A$. By the lemma, U^A is an autoepistemic model of $A(P, \mathsf{B}P)$, and $(U^A)^b = U$, so that $f(I, U^A) = I \cup (U^A)^b = I \cup U$.

It remains to check that f is one-one. Let (I_1, S_1) and (I_2, S_2) be structures such that S_1 and S_2 are autoepistemic models of $A(P, \mathsf{B}P)$ and $f(I_1, S_1) = f(I_2, S_2)$. From the definition of f, $I_1 = I_2$ and $S_1^b = S_2^b$, so that $\cap S_1 = \cap S_2$. It follows that, for every interpretation I, $(I, S_1) \models BP_i$ iff $(I, S_2) \models BP_i$. Consequently, $(I, S_1) \models A(P, BP)$ iff $(I, S_2) \models A(P, BP)$. Because S_1 and S_2 are autoepistemic models of $A(P, \mathsf{B}P)$, we conclude that $I \in S_1$ iff $I \in S_2$, so that $S_1 = S_2$.

Acknowledgements

This research was supported by NSF grants CDA-9015006, IRI-9101078, IRI-9103112 and IRI-9220645.

References

[Apt *et al.*, 1988] Krzysztof Apt, Howard Blair, and Adrian Walker. Towards a theory of declarative knowledge. In Jack Minker, editor, *Foundations of Deductive Databases and Logic Programming*, pages 89–148. Morgan Kaufmann, San Mateo, CA, 1988.

[Clark, 1978] Keith Clark. Negation as failure. In Herve Gallaire and Jack Minker, editors, *Logic and Data Bases*, pages 293–322. Plenum Press, New York, 1978.

[Gelfond, 1987] Michael Gelfond. On stratified autoepistemic theories. In *Proc. AAAI-87*, pages 207–211, 1987.

[Gelfond and Lifschitz, 1988] Michael Gelfond and Vladimir Lifschitz. The stable model semantics for logic programming. In Robert Kowalski and Kenneth Bowen, editors, *Logic Programming: Proc. of the Fifth Int'l Conf. and Symp.*, pages 1070–1080, 1988.

[Gelfond and Lifschitz, 1991] Michael Gelfond and Vladimir Lifschitz. Classical negation in logic programs and disjunctive databases. *New Generation Computing*, 9:365–385, 1991.

[Konolige, 1989] Kurt Konolige. On the relation between autoepistemic logic and circumscription. In *Proc. IJCAI-89*, pages 1213–1218, 1989.

[Levesque, 1990] Hector Levesque. All I know: a study in autoepistemic logic. *Artificial Intelligence*, 42(2,3):263–310, 1990.

[Lifschitz and Schwarz, 1993] Vladimir Lifschitz and Grigori Schwarz. Extended logic programs as autoepistemic theories. In Luis Moniz Pereira and Anil Nerode, editors, *Logic Programming and Non-monotonic Reasoning: Proceedings of the Second International Workshop*, pages 101–114, 1993.

[Lifschitz, 1989] Vladimir Lifschitz. Between circumscription and autoepistemic logic. In Ronald Brachman, Hector Levesque, and Raymond Reiter, editors, *Proc. of the First Int'l Conf. on Principles of Knowledge Representation and Reasoning*, pages 235–244, 1989.

[McCarthy, 1986] John McCarthy. Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence*, 26(3):89–116, 1986. Reproduced in [McCarthy, 1990].

[McCarthy, 1990] John McCarthy. *Formalizing common sense: papers by John McCarthy*. Ablex, Norwood, NJ, 1990.

[Moore, 1984] Robert Moore. Possible-world semantics for autoepistemic logic. In *Proc. of 1984 Non-monotonic Reasoning Workshop*, 1984.

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

[Perlis, 1988] Donald Perlis. Autocircumscription. *Artificial Intelligence*, 36:223–236, 1988.

[Schwarz and Truszczyński, 1992] Grigori Schwarz and Mirosław Truszczyński. Modal logic S4F and the minimal knowledge paradigm. In Yoram Moses, editor, *Theoretical Aspects of Reasoning about Knowledge: Proc. of the Fourth Conf.*, pages 184–198, 1992.

[Schwarz, 1992] Grigori Schwarz. Minimal model semantics for nonmonotonic modal logics. In *Proc. LICS-92*, pages 34–43, 1992.