Ideal introspective belief

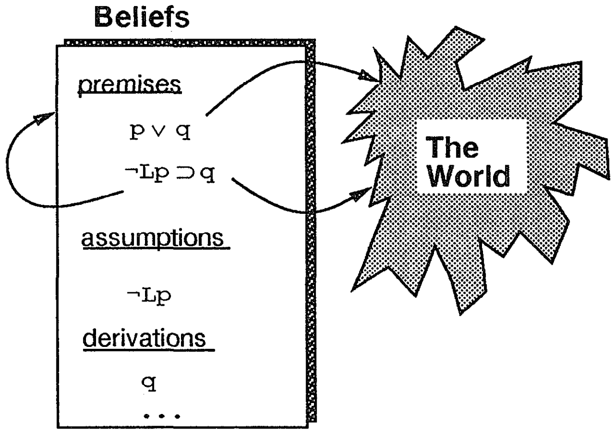
Kurt Konolige*
Artificial Intelligence Center
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
konolige@ai.sri.com

Abstract

Autoepistemic (AE) logic is a formal system characterizing agents that have complete introspective access to their own beliefs. AE logic relies on a fixed point definition that has two significant parts. The first part is a set of assumptions or hypotheses about the contents of the fixed point. The second part is a set of reflection principles that link sentences with statements about their provability. We characterize a family of ideal AE reasoners in terms of the minimal hypotheses that they can make, and the weakest and strongest reflection principles that they can have, while still maintaining the interpretation of AE logic as self-belief. These results can help in analyzing metatheoretic systems in logic programming.

Introduction

What kind of introspective capability can we expect an ideal agent to have? This question is not easily answered, since it depends on what kind of model we take for the agent's representation of his own beliefs. Autoepistemic logic (Moore [10]) uses a sentential or list semantics, which looks like this:



The beliefs of the agent are represented by sentences in a formal language. For simplicity, we consider just a propositional language $\mathcal{L}_0$, and a modal extension $\mathcal{L}_1$ which has modal atoms of the form $L\phi$, where ϕ is a sentence of $\mathcal{L}_0$.

The arrow indicates that the intended semantics of the beliefs from $\mathcal{L}_0$ is given by the real world, e.g., the belief q is the agent's judgment that q is true in the real world. Of course an agent's beliefs may be false, so that in fact q may not hold in the world. On the other hand, beliefs of the form $L\phi$ refer to the agent's knowledge of his own beliefs, so the semantics is just the belief set itself.

An agent starts with an initial set of beliefs, the *premises*. Through assumptions and derivations, he accumulates further beliefs, arriving finally at a belief set that is based on the premises. In order for an agent to be ideally introspective, the belief set Γ must satisfy the following equations:

$$\begin{array}{ll} \text{The premises are in } \Gamma. & \\ \phi \in \Gamma \text{ and } \phi \in \mathcal{L}_0 \rightarrow & L\phi \in \Gamma \\ \phi \notin \Gamma \text{ and } \phi \in \mathcal{L}_0 \rightarrow & \neg L\phi \in \Gamma \end{array} \qquad (1)$$

Any set Γ from $\mathcal{L}_1$ that satisfies these conditions, and is closed under tautological consequence, will be called $\mathcal{L}_1$-stable (or simply stable) for the premises Γ. The definition and term "stable set" are from Stalnaker [13]. The beliefs are stable in the sense that an agent has perfect knowledge of his own beliefs according to the intended semantics of L, and cannot infer any more atoms of the form $L\phi$ or $\neg L\phi$.

Although an ideal agent's beliefs will be a stable set containing his beliefs, not just any such set will do. For example, if the premises are $\{p \vee q\}$, one stable set is $\{p \vee q, p, Lp, L(p \vee q), \cdots\}$. This set contains the belief p, which is unwarranted by the premises. The constraint of making the belief set stable guarantees that the beliefs will be introspectively complete, but it does not constrain them to be soundly based on the premises. Moore recognized this situation in formulated autoepistemic logic; his solution was to ground the belief set by making every element derivable from the premises and some assumptions about beliefs. The reason he needed a set of assumptions is that negative

*The research reported in this paper was supported by the Office of Naval Research under Contract No. N00014-89-C-0095.

introspective atoms (of the form $\neg L\phi$) are not soundly derivable from the premises alone. For example, consider the premise set $\{\neg Lp \supset q, p \vee q\}$. We would like to conclude $\neg Lp$, since there is no reasonable way of coming to believe p. But an inference rule that would allow us to conclude $\neg Lp$ would have to take into account all possible derivations, including the results of its own conclusion. This type of circular reasoning can be dealt with by adding a set of assumptions about what we expect *not* to believe, and checking at the end of all derivations that these assumptions are still valid.

In autoepistemic logic, a belief set T is called *grounded in premises A* if all of its members are tautological consequences of $A \cup LT_0 \cup \neg L\overline{T}_0$, where $LT_0 = \{L\phi \mid \phi \in T \cap \mathcal{L}_0\}$, and $\neg L\overline{T}_0 = \{\neg L\phi \mid \phi \in \mathcal{L}_0 \text{ and } \phi \notin T\}$. This concept of groundedness is fairly weak, since it relies not only on assumptions about what isn't believed ($\neg L\overline{T}_0$), but also about what is (LT_0). In this paper we consider belief sets that use only assumptions $\neg L\overline{T}_0$ in forming the belief set T. Everything else in the belief set will follow deductively (and monotonically) from the premises A and the assumptions $\neg L\overline{T}_0$. In some sense $\neg L\overline{T}_0$ is the minimal set of assumptions that we can use in this manner; for every smaller set, we have to resort to nonmonotonic rules, such as negation-as-failure [6], in order to form a stable set. For this reason we call a belief set grounded in A and $\neg L\overline{T}_0$ *ideally grounded.*

Ideally grounded logics are similar to the modal nonmonotonic logics defined in [8, 12, 7], but allow an agent to make fewer assumptions about his own beliefs. The main difference is that ideally grounded logics are more grounded in the premises than modal nonmonotonic logics, and in general will have fewer unmotivated extensions (see Section).

In the rest of this paper we explore ideally grounded belief sets from the perspective of introspective reflection principles. We are able to characterize the minimal set of principles that will yield a stable set of beliefs, and also (once nested belief operators are introduced) the maximal ones. The resultant family of introspective logics fill in a hierarchy between strongly and moderately grounded autoepistemic logic [5], and suggest that the moderately grounded fixed-point is the best system for an ideal agent with perfect awareness of his beliefs.

Minimal ideal introspection

In this and the following section we restrict the language to $\mathcal{L}_1$, containing no nesting of the belief operator. This presents a simple system to explore the consequences of ideal introspection. In Section we relax this restriction and consider the fully nested modal language $\mathcal{L}$.

An ideally grounded introspective agent determines his belief set using the following fixed-point equation:

$$T = \{\phi \mid A \cup \neg L\overline{T}_0 \vdash_S \phi\}\,, \qquad (2)$$

where S is some system of inference rules. Any set T that satisfies this equation will be called an *ideally grounded extension* of A. The set $T_0 = T \cap \mathcal{L}_0$ is the *kernel* of T.

In the remainder of this section we consider the minimal set of rules S that guarantees a stable belief set for T. Because a stable set is closed under tautological consequence, the rules S must contain a complete set of propositional rules. In addition, whenever ϕ is in the belief set, we want to infer $L\phi$. The following two rules fulfill these conditions.

Rule Taut. From the finite set of sentences X infer ϕ, if ϕ is a tautological consequence of X.

Rule Reflective Up. From ϕ infer $L\phi$, if $\phi \in \mathcal{L}_0$.

Proposition 1 *Let RN be the rules Taut and Reflective Up. Every RN-extension of A is a $\mathcal{L}0$ stable set containing A.*

Proof. Every extension is closed under tautological consequence by rule Taut, and the premises must be in it, by the properties of $\vdash$. The condition $\phi \in \Gamma$ and $\phi \in \mathcal{L}_0 \rightarrow L\phi \in \Gamma$ holds because of rule Reflective Up. The condition $\phi \notin \Gamma$ and $\phi \in \mathcal{L}_0 \rightarrow \neg L\phi \in \Gamma$ holds since any proposition ϕ not in T will be part of the assumptions $\neg L\overline{T}_0$. ∎

Proposition 2 *If for every set $A \subseteq \mathcal{L}_1$, the S-extension of A is an $\mathcal{L}_1$ stable set containing A, then Taut and Reflective Up are admissible rules of S.*

Proof. If Taut is not an admissible rule for some extension T, then it cannot be closed under tautological consequence, and is not a stable set. Similarly, if Reflective Up is not admissible, T will contain ϕ and will not contain $L\phi$ for some proposition ϕ. ∎

These two propositions show that the rules RN form the minimal logic for ideally grounded agents, in the sense that RN extensions produce stable belief sets, and they must be included in any system that produces such sets. Further, every RN extension of A is *minimal for A*: there is no stable set S containing A such that $S_0 \subset T_0$.

Proposition 3 *Every RN extension of A is a minimal stable set for A.*

Proof. Suppose there is a stable set U for A whose kernel is a proper subset of T's. Then U must also satisfy the fixed-point condition, since the rules Reflective Up and Taut are admissible for stable sets (Proposition 2). By hypothesis the set $\neg L\overline{U}_0$ contains $\neg L\overline{T}_0$, and so U_0 must contain every element of T_0, a contradiction. ∎

The proof of this proposition points to a more general result for any class of rules that are sound with respect to the stable set conditions. An inference rule is sound with respect to stable sets if, whenever its antecedents are contained in a stable set, its consequent also must be (e.g., Reflective Up is sound because if ϕ is in a stable set, $L\phi$ must be also).

Proposition 4 *If the rules S are sound, then any S-extension of A is a minimal stable set for A.*

Proof. Suppose there is a stable set U for A whose kernel is a proper subset of T's. Then U must also satisfy the fixed-point condition, since the rules S are admissible for stable sets. By hypothesis the set $\neg L\overline{U}_0$ contains $\neg L\overline{T}_0$, and so U_0 must contain every element of T_0, a contradiction. ■

Groundedness, autoepistemic and default logic

In this section we relate ideally grounded extensions to their close relatives, default logic and AE extensions. Ideal groundedness is somewhat weaker than default logic and strongly grounded AE extensions, but stronger than moderately grounded ones.

Simple as it is, the system RN is almost equivalent to default logic [11]. It is not quite as strongly grounded as the latter; for while there exists a translation from DL to RN that preserves extensions, the inverse translation fails in a few cases.

We will assume that the reader is familiar with DL. A default theory $\langle W, D\rangle$ consists of a set of first-order sentences W and a set of defaults D of the form

$$\alpha : \beta_1, \cdots \beta_n / \gamma\,.$$

Here only the propositional case will be considered, but extending the results to first-order languages is straightforward (as long as no quantifying-in is allowed, e.g., sentences of the form $Qx.L\phi(x)$).

To get a translation to RN, simply take W and add a translation of each default rule, as follows:

$$A = W \cup \{L(\alpha\wedge\alpha)\wedge\neg L\neg\beta_1 \cdots \supset \gamma \mid \alpha : \beta_1, \cdots /\gamma \in D\}\,. \qquad (3)$$

Note the form of the first modal atom: $L(\alpha\wedge\alpha)$, rather than $L\alpha$. Since the beliefs of an agent are closed under tautological consequence, this amounts to the same constraint on beliefs; however, the difference is important for finding extensions, as will be made clear shortly.

Proposition 5 *U is the kernel of an RN extension of A iff it is a DL extension of $\langle W, D\rangle$.*

Proof. Let $A = W \cup \{L(\alpha \wedge \alpha) \wedge \neg L\neg\beta_1 \supset \gamma \mid \alpha : \beta_1, \cdots /\gamma \in D\}$. We will show that the set

$$\Gamma(U) = \{\phi \in \mathcal{L}_0 \mid A \cup \neg L\overline{U} \vdash_{\mathrm{RN}} \phi\}$$

is the least set satisfying the properties:

$W \subseteq \Gamma(U)$.

2 $\Gamma(U)$ is closed under tautological consequence.

3 For $\alpha : \beta_1, \cdots/\gamma \in D$, if $\alpha \in \Gamma(U)$ and $\neg\beta \notin U$, then $\gamma \in \Gamma(U)$.

The first two properties follow directly from the definition of $\Gamma(U)$. The third property follows by simple propositional inference, given the form of A.

To show $\Gamma(U)$ is minimal, note that it is the set of tautological consequences of W and some set γ_i of conclusions of defaults. To make it smaller, we would have to eliminate some of the γ_i. But it is clear from the discussion below that the only way a γ_i could be present is if the third condition defining $\Gamma(U)$ holds; thus all γ_i must be present, and $\Gamma(U)$ is minimal.

We can reduce the definition of extensions (2) to use only the kernel:

$$U = \{\phi \in \mathcal{L}_0 \mid A \cup \neg L\overline{U} \vdash_S \phi\}\,.$$

This gives a fixed-point condition defining extensions as

$$U = \Gamma(U)$$

which is the same as for default logic. ■

This is a simple translation of DL into a minimal AE logic. It is the same as the translation in [5] (except for the use of $\alpha\wedge\alpha$ instead of α), but there it was necessary to limit the extensions of the AE logic to strongly grounded ones, a syntactic method based on the form of the premises. No such method is needed here.

The stipulation on the form of $L(\alpha\wedge\alpha)$ is necessary to prevent derivations that arise from the interaction of modal atoms. Consider the two theories:

$$\{\neg Lp \supset p, Lp \supset p\}$$
$$\{\neg Lp \supset p, L(p\wedge p) \supset p\}$$

The first one has an RN extension $\mathrm{Cn}(p)$, because p is a tautological consequence of the initial constraints. On the other hand, it is not a consequence of the second set of constraints, because $\neg Lp$ and $L(p\wedge p)$ are consistent from the view of propositional logic. Since there is no way to derive p by any of the rules, $\mathrm{Cn}(p)$ cannot be an extension; yet assuming $\neg Lp$ leads to the derivation of p and a contradiction. So the second set has no extensions.

To get autoepistemic logic, we need to include more assumptions about beliefs in the fixed point equation 2. Let us define *open RN extensions* as solutions of the equation

$$T = \{\phi \mid A \cup LT_0 \cup \neg L\overline{T}_0 \vdash_{\mathrm{RN}} \phi\}\,, \qquad (4)$$

where LT_0 is the set $\{L\phi \mid \phi \in T_0\}$. Actually, the presence of the Up rule is redundant here. From results in [5], it is easy to show the following proposition.

Proposition 6 *T is an open RN extension of A iff it is the kernel of an AE extension of A.*

The kernel of an AE extension is just the part of the extension from $\mathcal{L}_0$. The kernel completely determines the extension.

So the basic difference between AE and default logic is based on the groundedness of the extensions, that is, AE logic lets an agent assume belief in a proposition α, and use that assumption to derive the very same proposition as part of the final set of beliefs. In default logic, all derivations must be ideally grounded, so that assumptions are of the form $\neg L\phi$.

The circular reasoning possible in AE logic was noted in [5], and two increasingly stronger notions, moderate and strong groundedness, were defined as a means of throwing out extensions that exhibit such reasoning. Moderately grounded extensions of A are defined as those AE extensions are also minimal stable sets containing A. Strongly grounded extensions use a syntactic method to eliminate all inferences from facts to belief propositions, e.g., even with the premise set

$$A = \{La \supset a, \neg La \supset a\} \tag{5}$$

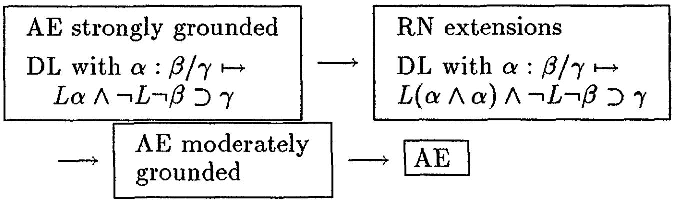
there is no derivation of a, because La and $\neg La$ are not allowed to interact. This means that different sets A, even if they are propositionally equivalent, can generate different extensions. Strongly grounded extensions are equivalent to default logic extensions under the simple translation of default rules:

$$\alpha : \beta_1, \cdots \beta_n/\gamma \mapsto L\alpha \wedge \neg L \neg \beta_1 \wedge \cdots \wedge \neg L \neg \beta \supset \gamma \, . \tag{6}$$

Note the difference with the translation of (3): $L\alpha$ instead of $L(\alpha \wedge \alpha)$.

Here, rather than defining restrictions on extensions, we have taken the approach of trying to find the minimal reflective principles that will allow an agent full knowledge of his beliefs, at the same time trying to make them as grounded as possible. The result is a logic that is somewhere between moderately and strongly grounded AE extensions, and which can imitate the groundedness conditions of default logic.

Let us define one fixed point logic $S1$ to be *included in* another $S2$ ($S1 \rightarrow S2$) if for any premise set the extensions of $S1$ are always extensions of $S2$, and for some premise set there is an extension of $S2$ that is not an extension of $S1$. $S1$ is the stronger nonmonotonic logic if we define ϕ as a consequence of a premise set just in case ϕ is in every extension of the premises. The relationship among the various AE logics can be diagrammed as follows:


AE strongly grounded
DL with $\alpha : \beta/\gamma \mapsto L\alpha \wedge \neg L \neg \beta \supset \gamma$
RN extensions
DL with $\alpha : \beta/\gamma \mapsto L(\alpha \wedge \alpha) \wedge \neg L \neg \beta \supset \gamma$
AE moderately grounded
AE

(7)

Nested belief

So far we have preferred to forego the complications of beliefs about beliefs, using the language $\mathcal{L}_1$ that contains no nesting of modal operators. This language and its semantics can be extended in a straightforward way. Let $\mathcal{L}$ be the propositional modal language formed from $\mathcal{L}_0$ by the recursive addition of atoms of the form $L\mu$, with $\mu \in \mathcal{L}$.

The semantic equations for a stable set (1) are modified to take away the restriction of beliefs being in $\mathcal{L}_0$:

The premises are in Γ.
$\phi \in \Gamma \rightarrow L\phi \in \Gamma$ (8)
$\phi \notin \Gamma \rightarrow \neg L\phi \in \Gamma$

Any set from $\mathcal{L}$ that satisfies these conditions, and is closed under tautological consequence, will be called a stable set for A (in contrast to $\mathcal{L}_1$-stable, which does not consider nested modal atoms).

Consider a premise set A that is drawn from $\mathcal{L}_1$, as before. In every RN extension of A there is complete knowledge of what facts are believed or disbelieved, i.e., $L\phi$ or $\neg L\phi$ is present for every nonmodal ϕ. The addition of the nested modal atoms should make no difference to this picture, except to reflect the presence of the belief atoms in the correct way. So, for example, if La is in an RN extension S, then LLa should be in the extension when we consider $\mathcal{L}$; and similarly $L\neg La$ should be present if $\neg La$ is not in S. This much is easily accomplished by removing the restriction on Reflective Up, and giving it its usual name from modal logic.

Rule Necessitation. From ϕ infer $L\phi$.

This rule will add positive modal atoms; but we need also to add negative ones. For example, if La is in an extension, and the extension is consistent, then $\neg La$ is not in it, and this fact should be reflected in the presence of $\neg L \neg La$. In fact we want to infer $\neg L\mu$ for *every* sentence μ that will not be in the extension, given that we have full knowledge of the belief atoms from $\mathcal{L}_1$. Suppose that there is a sentence $La \vee \neg Lb \vee c$ that is not in S, where c is a nonmodal sentence. This implies that, for stable S, $\neg La \in S$, $Lb \in S$, and $\neg Lc \in S$. So from these latter sentences we should infer $\neg L(La \vee \neg Lb \vee c)$. This is what the following rule does.

Rule Fill. From $L\alpha_i$, $\neg L\beta_j$, $\neg L\gamma$, and $\mu \supset (\bigvee_i L\alpha_i \vee \bigvee_j \neg L\beta_j \vee \gamma)$, infer $\neg L\mu$.

The system NRN consists of the rules Taut, Necessitation, and Fill. The basic properties of NRN extensions are that they are minimal stable sets, the rules are essential, and they are conservative extensions of RN fixed points.

Proposition 7 *If for every set $A \subseteq \mathcal{L}$, the S-extension of A is a stable set containing A, then Taut, Necessitation, and Fill are admissible rules of S.*

Proof. Taut and Nec are the same as for Proposition 2. For Fill, note that every consistent stable set containing the premises to the rule cannot contain μ, and so must contain $\neg L\mu$. ∎

Proposition 8 *Every NRN extension of A is a stable set for A.*

Proof. Assume that T is a consistent NRN extension of A. By rule Nested Reflective Up, the first part of the semantic definition is satisfied. For negative modal atoms, we proceed by induction on the level of nesting of L. By definition and the rule Nested Reflective Up, either $L\phi$ or $\neg L\phi$ is in T for every nonmodal ϕ. Suppose a sentence $s = (\bigvee_i L\alpha_i \vee \bigvee_j \neg L\beta_j \vee \gamma) \in \mathcal{L}_1$ is not in T. Then each of $\neg L\alpha_i$, $L\beta_j$ and $\neg L\gamma$ is in T. By rule Fill, $\neg L\mu$ is in T for any $\mu \supset s$. Hence for every sentence $\nu \in \mathcal{L}_1$, the negative semantic rule is satisfied, and either $L\nu$ or $\neg L\nu$ is in T. By induction, it can be shown that the semantic rule is satisfied for all levels of nesting. ∎

Extensions that are stable sets are also minimal, as for the nonnested language.

Proposition 9 *If the rules S are sound with respect to stable sets, and the S-extension of A is a stable set, then it is a minimal stable set for A.*

Proof. Same as for Proposition 4. ∎

Proposition 10 *If $A \subseteq \mathcal{L}_1$, then the kernel of every RN extension is the kernel of an NRN extension, and conversely, the kernel of every NRN extension is the kernel of an RN extension.*

Proof. The converse is obvious, since the rules NRN include RN. For the original direction, assume we have an RN extension S, which contains $L\phi$ or $\neg L\phi$ for every $\phi \in \mathcal{L}_0$. From the proof of Proposition 8, it is clear that the set $T = \{\mu \mid S \vdash_{\text{NRN}} \mu\}$ is a stable set for A, and further it is an NRN extension, since all elements of its kernel are derivable from A and $\neg L\overline{S}$. ∎

Finally, we can show that the Fill rule is redundant if the schema K ($[L\phi \wedge L(\phi \supset \psi)] \supset L\psi$) is present.

Proposition 11 *The rule Fill is admissible in any system containing K, Taut and Necessitation.*

Proof. Suppose each of $\neg L\alpha_i$, $L\beta_j$ and $\neg L\gamma$ is in A, together with K and all instances of K. Let $\mu = \bigwedge_i \neg L\alpha_i \wedge \bigwedge_j L\beta_j$. By Taut and Up, $L[\mu \wedge (\mu \supset \gamma) \supset \gamma]$ is derivable, and from schema K and $\neg L\gamma$ we have $\neg L[\mu \wedge (\mu \supset \gamma)]$. Since we also have $L\mu$ by Up, this gives (using K) $\neg L(\mu \supset \gamma)$. Again by K and Taut, we could derive $\neg L\nu$ for any ν such that $\nu \supset (\mu \supset \gamma)$ is a tautology. ∎

Because nested modal atoms are propositionally distinct from nonnested ones, it is possible to derive new translations from default logic to sentences of $\mathcal{L}$ such that all extensions are strongly grounded and hence equivalent to default logic extensions. There are many ways to do this; all that is required is to translate from $\alpha : \beta/\gamma$ to a sentence in which α and β are put under different nestings of modal operators that correspond to the single nesting semantics. For example, three such translations are:

$$\begin{array}{ll} a) & LL\alpha \wedge \neg L \neg \beta \supset \gamma \\ b) & L\alpha \wedge \neg LL \neg \beta \supset \gamma \\ c) & L\alpha \wedge L \neg L \neg \beta \supset \gamma \end{array} \qquad (9)$$

Reflective reasoning principles

The systems RN and NRN are minimal rules that might be used by an agent reasoning about its own beliefs. They have the nice characteristic of giving minimal stable sets, and so are somewhere between strongly and moderately grounded. But are there other reflective reasoning principles that could be incorporated? In this section we will give a partial answer to this question by examining several standard modal axiomatic schemata, and showing how some of them are appropriate as general reasoning principles, while others must be regarded as specific assumptions about the relation of beliefs to the world.

The most well-known modal schemata are the following.

$$\begin{array}{ll} K. & L(\phi \supset \psi) \supset (L\phi \supset L\psi) \\ T. & L\phi \supset \phi \\ D. & L\phi \supset \neg L \neg \phi \\ 4. & L\phi \supset LL\phi \\ 5. & \neg L\phi \supset L \neg L\phi \end{array} \qquad (10)$$

The first question we could ask is: which of these schemata are sound with respect to the semantics of amalgamated belief sets? It should be clear that K, 4 and 5 are all sound, since if their antecedents are true of a stable set, then so are their consequents. The schema D is true only of consistent stable sets, as we might expect, since it says that a sentence can be in a belief set only if its negation is not.

The schema T, on the other hand, is not semantically valid. It is possible for an agent to believe a fact ϕ, but that fact may not be true in the real world. Asserting T for a particular fact ϕ says something about the agent's knowledge of how his beliefs are related to the world, and causes different reasoning patterns to appear in an agent's inferences about his own beliefs.

Here is a short example of how the sentence $Lp \supset p$ could be used by an agent. Consider the propositions:

p = The copier repairman has arrived
q = The copier is ok

Suppose an agent believes that if he has no knowledge that the repairman has arrived, the copier must be ok. Further he believes that the copier is broken. We represent this as:

$$A = \{\neg q, \neg Lp \supset q\} \ . \qquad (11)$$

The premises A do not have any NRN or AE extension, because while Lp is derivable, p is not. One solution is to give the agent confidence in his own beliefs, e.g.,

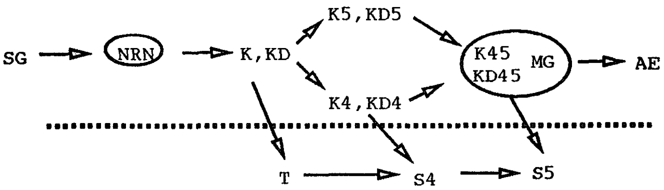
$$A' = \{\neg q, \neg Lp \supset q, Lp \supset p\} \ . \qquad (12)$$

Now there is an NRN-extension in which p is true, since from Lp the agent can derive p. It is as if the agent says, "I believe that p, therefore p must be the case."

Although one might not want to use this type of reasoning in a particular agent design, the point is that T sanctions a certain type of reasoning about the connection of beliefs to the world, and is thus a "nonlogical" axiom, similar to $\neg Lp \supset q$.

Different modal systems can be constructed by combining the different modal schemata with the inference rules Taut and Necessitation. Using our previous definition of inclusion, we show the following relations among the different versions of S-extensions.

Proposition 12 *The following diagram gives all the inclusion relations of ideally grounded extensions based on the modal systems formed from the schemas K, T, D, 4, and 5.*



Proof. We will sketch the technique for two examples. The basic idea is to consider a theory containing variations of the pair of sentences $Lp \supset p$, $\neg Lp \supset p$. This theory has the single extension with kernel Cn(p). For the system K, consider the pair $Lp \supset p$, $\neg L(p \wedge p) \supset p$. This theory has no RN extensions. But it does have a K-extension, since in the system K one infers p. Hence K extensions and RN extensions are distinct. For the schema 4, consider the pair of sentences $LLp \supset p$, $\neg L(p \wedge p) \supset p$. No K or RN extensions exist; but there is a $K4$ extension, since in $K4$ the pair infers p. Similar pairs can be found for the other systems. ∎

The top half are systems whose extensions are all subsets of AE logic. SG stands for strongly grounded AE extensions, and MG for moderately grounded. The minimal ideally grounded system is NRN, and the maximum is K45 or KD45, which is equivalent to MG (see [5]). An ideal introspective agent would use KD45 extensions, which we call ideal extensions. Note that the schema D does not make any difference as far as ideally grounded extensions are concerned; in effect, the agent cannot use reasoning about self-belief to detect an incoherence in his beliefs.

In fact all of the systems from NRN to KD45 are very similar. Their only difference comes from premise sets that contain sentences of the form

$$\begin{gathered}\neg Lp \supset p\\ \alpha \supset p\,,\end{gathered}$$

where $\alpha \supset Lp$ is a theorem of the modal system. For example, in K we have $L(p \wedge p) \supset Lp$, and a premise set as above with $\alpha = L(p \wedge p)$ would distinguish K from NRN, in that the former would have an extension containing p. Similarly, $\alpha = \neg L \neg Lp$ could be used for $K5$. But the sentence $\neg Lp \supset p$ is generally not one that captures a useful introspective reasoning pattern, and would probably not occur by design in an application. There thus seems to be no practical difference between NRN and KD45, since the additional axioms do not result in potentially interesting reasoning patterns.

The second tier is present for formal completeness. The axiom schema T, we have argued, is a useful way of characterizing a domain-dependent and proposition-dependent connection between the agent's beliefs and the world. These systems do not respect sound autoepistemic reasoning, and are not included in AE logic: the extensions generated using instances of T can differ significantly from AE extensions. In fact, if the AE fixed-point equation (4) is supplied with the axiom schema T, then it degenerates into monotonic $S5$ [9, 10]. This is because it interacts with the positive assumptions LT_0, producing arbitrary ungrounded beliefs. In ideally grounded logic, the T schema can serve a useful representational purpose, and all modal systems, including $S5$, produce nonmonotonic fixed points.

Modal nonmonotonic logics

Modal nonmonotonic logics are based on the following fixed point equation:

$$T = \{\phi \mid A \cup \neg L\overline{T} \vdash_S \phi\}\,,$$

where S is a modal system. McDermott [8] analyzed this equation for the systems T, S4, and S5. Subsequent investigations [12, 7] considered many other modal systems, including most of those mentioned in this paper. The difference with ideally grounded extensions is the presence of assumptions containing nested atoms, e.g., $\neg L \neg Lp$. For an ideal agent, this amounts to an assumption of Lp, since any stable set not containing $\neg Lp$ must contain Lp. In fact, modal nonmonotonic logics whose underlying modal system contains the schema 5 are all equivalent to AE logic. And as with AE logic, the schemas 5 and T combine to collapse the fixed point to monotonic $S5$.

From the point of view of ideally grounded extensions, the assumption set $\neg L\overline{T}$ is too "large." The schema 5, which in ideally grounded extensions is just a principle of reasoning about derived beliefs, in modal nonmonotonic logic also interacts with nested negative assumptions to produce positive ones. The inclusion diagram for ideally grounded extensions is almost the same as that for the normal modal systems serving as a deductive base (see [2]), except for the schema D. But all modal nonmonotonic logics containing the K and 5 schemas (but not T) are equivalent to weakly grounded AE logic because of their large assumption set, collapsing systems that are distinct in the ideally grounded case. Because of this, modal nonmonotonic logic misses the moderately grounded endpoint. In fact, no modal nonmonotonic logic produces only minimal stable sets: in the simplest system N, containing only the necessitation rule and no logical axioms, the premises $\{Lp \supset p, \neg L \neg Lp \supset p\}$ have two extensions, Cn() and Cn(p). Only the first of these is minimal.

Conclusion

We have presented the minimal logic (NRN) that an ideal introspective agent should use. It is minimal in the sense that the agent makes a minimal set of assumptions about his own beliefs, and employs a minimal set of rules necessary to guarantee that his beliefs are stable. An ideal introspective reasoner may enjoy more powerful rules of introspection, for example the modal schemas 4 and 5, but he should keep the assumptions about his beliefs to a minimum. The schema T is not a sound axiom for an introspective agent, but can be used to characterize a contingent connection between beliefs and the world.

The concept of ideally grounded extensions first appeared in [5], where the system KD45 was presented and proven equivalent to moderately grounded AE extensions.[1] Fixpoints of the systems T, S4 and S5 were introduced under the name of nonmonotonic ground logics in [14], and it was shown that the S5 logic was nondegenerate and consistent, i.e., does not reduce to monotonic S5, and always has an extension.

Ideally grounded logic might be employed in an analysis of metatheoretic systems, such as the DEMO and SOLVE predicates in logic programming [1, 3]. Using a predicate to represent provability can cause problems with syntax and consistency (see [4] for some comments). Instead, this research suggests using a modal operator, and defining a theory by the fixed point definition (2). Some appropriate notion of negation-as-failure would be used to generate the assumptions, and the rest of the fixed point could be calculated using the reflection rules.

References

[1] K. A. Bowen and R. A. Kowalski, Amalgamating language and metalanguage in logic programming, Computer and Information Science Report 4/81, Syracuse University (1981).

[2] B. F. Chellas, *Modal Logic: An Introduction* (Cambridge University Press, 1980).

[3] S. Costantini, Semantics of a metalogic programming language, *International Journal of Foundations of Computer Science* **1** (3) (1990).

[4] J. des Rivières and H. Levesque, The consistency of syntactical treatments of knowledge, in: J. Y. Halpern, ed., *Conference on Theoretical Aspects of Reasoning about Knowledge* (Morgan Kaufmann, 1986) 115–130.

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

[6] J. W. Lloyd, *Foundations of Logic Programming* (Springer-Verlag, Berlin, 1987).

[7] W. Marek, G. F. Schwarz, and M. Truszczynski, Modal nonmonotonic logics: ranges, characterization, computation, in: *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, Cambridge, MA (1991).

[8] D. McDermott, Non-monotonic logic II, *Journal of the ACM* **29** (1982) 33–57.

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

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

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

[12] G. F. Schwarz, Autoepistemic modal logics, in: *Conference on Theoretical Aspects of Reasoning about Knowledge*, Asilomar, CA (1990).

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

[14] M. Tiomkin and M. Kaminski, Nonmonotonic default modal logics, in: *Conference on Theoretical Aspects of Reasoning about Knowledge*, Asilomar, CA (1990).

[1] A slightly different fixed-point was used because of a technical difference in the form of monotonic inference in modal systems.