A Computational Theory of Belief Introspection

Kurt Konolige
Artificial Intelligence Center
SRI International
Menlo Park, California 94025

Abstract

Introspection is a general term covering the ability of an agent to reflect upon the workings of his own cognitive functions. In this paper we will be concerned with developing an explanatory theory of a particular type of introspection: a robot agent's knowledge of his own beliefs. The development is both descriptive, in the sense of being able to capture introspective behavior as it exist; and prescriptive, in yielding an effective means of adding introspective reasoning abilities to robot agents.

1 Introduction

Introspection is a general term covering the ability of an agent to reflect upon the workings of his own cognitive functions. In this paper we will be concerned with developing a theory of a particular type of introspection: an agent's knowledge of his own beliefs. There are at least two reasons why it is important to develop such a theory, one descriptive and the other prescriptive. As Collins and his coworkers have shown (in [1]), an agent often reasons about his own beliefs and nonbeliefs in deciding the answer to a posed query; hence a descriptively adequate account of agents' beliefs must deal with introspection. The second reason is that researchers attempting to build artificial agents must imbue these agents with introspective knowledge if they are to act in an intelligent manner. Moore [11] gives the example of an agent who must introspect about his beliefs in order to form a correct plan to achieve a goal.

In this paper we offer an explanatory theory of belief introspection based on the concept of a *belief subsystem* as developed in Konolige [4], [5]. Put simply, a belief subsytem is the computational structure within an artificial agent responsible for representing his beliefs about the world. Because the belief subsystem is "at hand" and available to the agent, it is possible for the agent to gain knowledge of his beliefs by simply making recursive calls to this belief subsystem, perhaps with ever-decreasing resource allocations. This, in a nutshell, is the model of introspection we adopt. Its advantages are that it is an adequate explanatory theory of belief introspection, and that it is immediately prescriptive: the theory shows how artificial agents that exhibit introspective reasoning of the requisite sort can be built.

Given the importance of introspective reasoning, it is perhaps surprising that the problem of finding a good explanatory basis for belief introspection in artificial agents has scarcely been addressed. In Section 3 we review two approaches that differ from ours in being nonconstructive: an ideal agent's introspective reasoning is defined by putting constraints on her belief set. The disadvantage of such nonconstructive theories is that, in general, they do not extend to the case where an agent's reasoning powers are bounded by resource limitations.

This research was made possible in part by a gift from the System Development Foundation. It was also supported by Grant N00014-S0-C-0296 from the Office of Naval Research.

2 The Introspective Machine

We start developing a theory of belief introspection by considering the computational embodiment of belief in an artificial agent. We have argued elsewhere (*e.g.*, Konolige [4]) for the identification of a *belief subsystem* as a conceptually separate part of an agent's cognitive makeup. A belief subsystem **M** consists of a finite list of facts the agent believes to be true of the world (the *base set*), together with some computational apparatus for inferring consequences of these facts. **M** interacts with other cognitive systems of the agent (*e.g.*, a planning system) as a query-answering device. It accepts a query ϕ and attempts to show that ϕ be derived from its base set of facts. The *belief set* of an agent is the set of all queries that can be derived.

The queries presented to **M** are in an internal language L; the exact nature of this language is not important, but there must be expressions in it that refer to the agent's own beliefs. We take these expressions to be of the form $\Box\phi$, meaning the agent believes ϕ to be one of his own beliefs. Formulas of L not containing $\Box$ are called *nondoxastic*; the sublanguage of L consisting of all nondoxastic sentences is called the *underlying language*.

When presented with a query in the language L, we assume **M** operates by either matching the query against its base set, or applying inference rules to generate subqueries in a backward-chaining manner. For example, in trying to answer the query $P \vee Q$, it may split the disjunc-

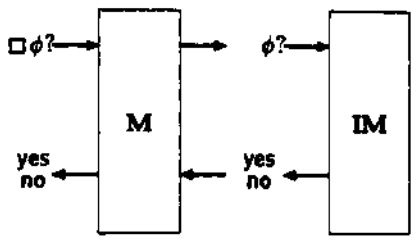

Figure 1: An Introspective Belief Subsystem

tion into two separate queries, and try to answer each of these. During the course of generating subqueries, it come upon one that is a question about its own state, *i.e.*, of the form $\Box\phi$. Such a subquery can be answered by making a recursive call to the belief subsystem again, posing the query ϕ. Conceptually we can think of this recursive call as a call on a new belief subsystem **IM** (the *introspective machine*). The **IM** may have different characteristics than **M** — for instance, it may have only a subset of the facts available to **M**, or even have facts that contradict those in **M** (according to Hintikka [3], people can have introspective beliefs of this sort). If the **IM** must answer a query about *its* self-beliefs, then it relies on another machine, the **IIM**; and so on, creating a hierarchy of belief subsystems. We write $\mathbf{I}^n\mathbf{M}$ to indicate the nth element of the hierarchy, with $\mathbf{M} = \mathbf{I}^0\mathbf{M}$. A belief subsystem that relies on an introspective machine to answer queries about self-beliefs in this manner is called an *introspective belief subsystem*.

A query ϕ that is answered affirmatively by **IM** means that the agent, upon introspecting on his own beliefs, comes to the conclusion that she believes ϕ — that is, $\Box\phi$ is one of her beliefs. A negative answer, on the other hand, means that she doesn't believe her belief subsystem computes ϕ, and so in this case $\neg\Box\phi$ is one of her beliefs. Figure 1 illustrates the workings of the introspective machine by showing the way in which **M** responds to the query $\Box\phi$. **M** poses the subquery ϕ to the introspective machine. If **IM** answers *yes*, then $\Box\phi$ is accepted as a belief, and **M** also answers affirmatively. If **IM** answers *no*, $\Box\phi$ is not a belief.

There are no restrictions on the inference rules that a belief subsystem uses, except that they should be *sound* with respect to the semantics of the underlying language (they need not be complete). In particular, we wish to exclude rules of an introspective nature, because we want all properties of introspection to arise from the interaction of **M** and **IM**. For example, if the underlying language is propositional, the rules should respect the truth-functional semantics of the boolean connectives. In the case of a first-order underlying language, rules such as those in Kripke [6] or Konolige [4] may be used, disallowing the rules explicitly dealing with modal operators.

Proposition 1 *Suppose* **M** *is an introspective belief subsystem whose base sentences are nondoxastic and consistent. Then* **M** *is consistent.*

The proof[1] of this proposition follows from noting that **M** is atomically consistent. For nondoxastic atoms P, at most one of P or $\neg P$ will be derivable from the base sentences. For atoms $\Box\phi$, the responses given in Figure 1 indicate that either $\Box\phi$ or $\neg\Box\phi$ will be provable, but not both. Note that if **IM** is inconsistent, $\Box\phi$ and $\Box\neg\phi$ will be provable in **M** for some sentence ϕ.

2.1 Ideal Agents

An ideal agent should have perfect knowledge of her own beliefs. This motivates the following definition.

Definition 1 *An* ideal introspective belief subsystem **M** *satisfies three criteria:*

1. *The belief set of* **M** *is consistent.*
2. *The inference rules are complete.*
3. $\mathbf{I}^n\mathbf{M} = \mathbf{M}$ *for all* $n > 0$.

The first condition is that an ideal agent's beliefs not be contradictory. An interesting case of inconsistency occurs when the base set of **M** contains doxastic sentences. For example, suppose the base set of consists solely of the sentence $\Box P$. Now the query $\Box P$ will be answered affirmatively in **M** (by direct matching). The query $\neg\Box P$ can also be proven, because it generates the query P to **IM**. Since **IM** has the same base set as **M**, it answers P negatively, and so **M** affirms $\neg\Box P$.

Not all doxastic base sentences lead to inconsistency, of course; sometimes their presence is required for useful inference. Moore [12] gives the following example of an agent's introspective reasoning: "I don't have any brothers, because if I did, I would know about them, and I have no such knowledge." If we let P stand for "I have no brothers," then the agent's base set includes the axiom $\neg\Box P \supset \neg P$. Now given the query $\neg P$, the agent's belief subsystem would use the axiom and set up the goal of proving $\neg\Box P$. This generates a query P for the **IM**, which answers negatively. Hence $\neg\Box P$ is proven, and so is the original query $\neg P$.

The completeness of the inference rules is with respect to the semantics of the underlying language. If the rules are complete, the belief set is closed under the appropriate notion of logical consequence for the underlying language. In the case of a propositional language, it is closed under truth-functional consequence.

The third condition is the requirement that ideal agents have perfect introspective knowledge. We enforce this by assuming that an agent's view of her own belief subsystem (**IM**) is exactly the same as the real subsystem (**M**). Because the introspective machine has its own introspective

[1] Space requirements preclude more than sketches of most proofs in this paper.

machine (IIM), this too must be the same as M, and so on to arbitrary introspective levels.

Proposition 2 *In an ideal introspective belief subsystem,*

PI: *If* M *responds* yes *to* ϕ, *it responds* yes *to* $\Box\phi$.

NI: *If* M *responds* no *to* ϕ, *it responds* yes *to* $\neg\Box\phi$.

To prove this proposition, note that the IM has exactly the same structure as M, and so has exactly the same behavior on a query ϕ. *PI* abbreviates *positive introspection*; informally, it says that if an agent believes a proposition, she believes that she believes it. Similarly, *NI*, or *negative introspection*, says that an agent has knowledge of what she does not believe.

In the best of all possible worlds, we could actually implement such an introspective belief subsystem, and so provide artificial agents with an ideal mechanism for reasoning about their own beliefs. Unfortunately, except under fairly strict conditions on the underlying language, there does not exist any realizable computational structure that will implement an ideal introspective belief subsystem.

Definition 2 *A belief subsystem* M *is* decidable *if there exists an algorithm that will return* yes *when* ϕ *is derivable and* no *when not; it is* semi-decidable *if there exists an algorithm which will return* yes *when* ϕ *is derivable; it is* undecidable *if it is not semidecidable.*

When we talk about the decidability of an introspective belief subsystem M, we normally take derivability to include the derivation of introspective beliefs via IM. Sometimes, however, we want to refer to the decidability of M without the introspective rules; to make this clear, we say "the decidability of nonintrospective M."

Proposition 3 *Let* M *be an ideal introspective belief subsystem. If nonintrospective* M *is not decidable,* M *is undecidable.*

The proof is simple: suppose M is semidecidable. Then there is an algorithm for determining that M returns yes on $\neg\Box\phi$ and $\Box\phi$, where ϕ is an arbitrary sentence. Thus there is an algorithm for determining whether ϕ is derived or not by IM, contradicting the assumption that IM is not decidable.

Proposition *4 If the underlying language is propositional, and its base $et i$ nondoxastie, an ideal introspective belief subsystem is decidable.*

The proof here is straightforward: any query will have a finite maximum embedding n of self-belief operators. One need only look at the (decidable) theorems produced by the first n levels of the introspective machine. As long as queries do not include any quantification into the context of the self-belief operator, we can extend this result to any underlying language which can be decided by reduction to the propositional calculus. For example, monadic predicate calculus (PC) and the class of $\exists\forall$-sentences have this property.

These two propositions to some extent delimit the nature of decidability for introspective subsystems. A natural question to ask is if Proposition 4 can be extended to the case of *any* decidable underlying language. The answer to this has important consequences for adding introspective ability to artificial agents, because these agents are (nonintrospectively) decidable: they must answer a belief query in a finite amount of time.

Proposition 5 // *the underlying language is monadic PC, and its base set is nondoxastic, an ideal introspective belief subsystem is decidable.*

The proof of this proposition relies on Kripke's result in [7] that monadic modal PC is not decidable. The difference between monadic modal PC and propositional modal languages is that the former allows quantifying into the modal context. As we mentioned, queries without quantifying-in are decidable for monadic PC. Thus the presence of quantifying-in seems to pose an inherently difficult computational problem for introspective systems. Yet the expressivity of quantifying-in is desirable in many applications; Levesque [9] gives the example of a question-answering system in which sentences of the form $\exists x[P(x) \wedge \neg\Box P(x)]$ express the fact that there are individuals with property *P* whose identity is unknown to the database.

Proposition 5 is discouraging, since it means that in constructing introspective agents, we must either use a very weak underlying language, or give up some of the three conditions of ideality. We discuss the latter method in the next section. Note that even without Proposition 5, there are reasons for developing the theory of non-ideal agents. First, even with a very weak underlying language and a decidable subsystem, an agent may have limited resources for derivation of beliefs, and can only compute an approximation to the conditions of Definition 1. Second, we mentioned that human agents are not always ideal agents, and we would like to model their cognitive behavior.

2.2 Real Agents

In Figure 1, a belief subsystem had to respond either yes or no to every query. In a computational setting with finite resource bounds, it may not be possible to do this in a consistent way. For example, if the underlying language is PC, there are some (nondoxastic) queries that do not have a derivation, and hence the belief subsystem should respond no; but there is no algorithm for determining this in a finite amount of time. To accommodate this situation, we allow a subsystem to return und (undecided) as one of its answers.

Let R be a resource bound. If M derives a query ϕ within this bound, we write $\mathbf{M}(\phi, R)$:yes; if it decides that ϕ is not derivable, we write $\mathbf{M}(\phi, R)$:no; and if it cannot

decide one way or the other within the bounds R, we write $\mathbf{M}(\phi, R)$:und. (We abbreviate $\forall r.\mathbf{M}(\phi, r)$:$x$ by $\mathbf{M}(\phi)$:x.) Note that real agents are computationally oriented; the inference rules specify which derivations are possible, but the subsystem has the option of responding und if its resources are not sufficient to actually compute a derivation.

The response of the introspective machine in Figure 1 is extended in the following way: whenever IM returns und on ϕ, M returns und on both $\Box\phi$ and $\neg\Box\phi$. We can summarize the response of M to self-belief queries of the form $\Box\phi$ and $\neg\Box\phi$ by considering the behavior of the IM on ϕ. R is the bound for the self-belief query, and R' for the introspective query.

$\mathbf{IM}(\phi, R')$:yes	$\rightarrow$	$\mathbf{M}(\Box\phi, R)$:yes,	$\mathbf{M}(\neg\Box\phi, R)$:no
$\mathbf{IM}(\phi, R')$:no	$\rightarrow$	$\mathbf{M}(\Box\phi, R)$:no,	$\mathbf{M}(\neg\Box\phi, R)$:yes
$\mathbf{IM}(\phi, R')$:und	$\rightarrow$	$\mathbf{M}(\Box\phi, R)$:und,	$\mathbf{M}(\neg\Box\phi, R)$:und

Note that we want to leave open the possibility that a real agent has no knowledge of some of her own beliefs, and this is where the "undecided" answer plays a crucial role. If IM returns und, then M will be undecided about its introspective belief.

One obvious result of the imposition of resource bounds is that condition (2) of Definition 1 must be abandoned for sufficiently hard underlying languages. Further, we may also have to give up condition (3). Given resource bounds, the behavior of IM may differ significantly from M, even when they have the same base sentences and inference rules. For example, let the query to M be the sentence $\alpha \wedge \Box\beta$ with resource bound R. The control strategy of M might apply a rule to break this sentence into two conjunctive subqueries, α and $\Box\beta$. The solution of α may consume a large fraction or M's computational resources. Thus when it asks IM to solve the query β, it may give IM a significantly lower resource bound than R. Thus although M would respond yes to β posed *simpliciter*, it won't be able to derive the subquery $\Box\beta$, because IM does not have enough resources to do so.

If constraints (2) and (3) of Definition 1 do not hold for real agents, can we find weaker correspondents? For condition (2) we have already done the best we can, by assuming that the answers returned by a subsystem respect the intended semantics of the underlying language and self-belief operator. Because of this, real agents as we have defined them obey a *monotonicity* condition: for $R' > R$, the only difference in the behavior of a belief subsystem can be to change some undecided queries to yes or no. Thus a belief subsystem with a large resource bound is never further away from consequential closure than one with a small bound. However, we may not want real agents to abide by monotonicity — perhaps, if a query cannot be derived within a resource bound, we may want to jump to the conclusion that it cannot be derived.[2] In this sense our definition of belief subsystem may be too strict for some purposes.

We can obtain weaker versions of condition (3) by considering two interesting constraints between IM and M: *faithfulness* and *fulfillment*. Roughly, an IM is faithful if whenever it returns a definite answer (yes or no) on a query, M also returns the same answer on that query. Fulfillment is the converse: whenever M returns a yes (or no) on a query, IM must also.

Definition 3 *An introspective belief subsystem* M *is* faithful *if it has the following properties for every introspective pair* $\mathbf{I}^n\mathbf{M}$, $\mathbf{I}^{n+1}\mathbf{M}$:

positive faithfulness (*pfa*):
$$\exists r.\mathbf{I}^{n+1}\mathbf{M}(\phi, r)\text{:yes} \rightarrow \sim\mathbf{I}^n\mathbf{M}(\phi)\text{:no}$$
negative faithfulness (*nfa*):
$$\exists r.\mathbf{I}^{n+1}\mathbf{M}(\phi, r)\text{:no} \rightarrow \sim\mathbf{I}^n\mathbf{M}(\phi)\text{:yes}$$

Proposition 6 *In a faithful introspective belief subsystem* M,

pfa: $\exists r.\mathbf{M}(\Box\phi, r)$:yes $\rightarrow \forall r. \sim\mathbf{M}(\phi, r)$:no

nfa: $\exists r.\mathbf{M}(\neg\Box\phi, R)$:yes $\rightarrow \forall r. \sim\mathbf{M}(\phi, r)$:yes

This proposition follows readily from the definition of faithfulness and the monotonicity of responses with increasing resource bounds. Faithfulness is about the weakest constraint we can impose on introspective systems, and is a kind of soundness condition on introspective reasoning. That is, IM should not contradict M, in the sense that if IM ever decides a query, M should never decide the opposite.

Definition 4 *An introspective belief subsystem* M *is* fulfilled *if it has the following properties for every introspective pair* $\mathbf{I}^n\mathbf{M}$, $\mathbf{I}^{n+1}\mathbf{M}$ *and resource* R:

positive fulfillment (*pfu*):
$$\mathbf{I}^n\mathbf{M}(\phi, R)\text{:yes} \rightarrow \mathbf{I}^{n+1}\mathbf{M}(\phi, R)\text{:yes}$$
negative fulfillment (*nfu*):
$$\mathbf{I}^n\mathbf{M}(\phi, R)\text{:no} \rightarrow \mathbf{I}^{n+1}\mathbf{M}(\phi, R)\text{:no}$$

Proposition 7 *In a fulfilled introspective belief subsystem* M,

pfu: $\mathbf{M}(\phi, R)$:yes $\rightarrow \exists r \geq R.\mathbf{M}(\Box\phi, r)$:yes

nfu: $\mathbf{M}(\phi, R)$:no $\rightarrow \exists r \geq R.\mathbf{M}(\neg\Box\phi, R')$:yes

This proposition follows from the definition of fulfillment and the monotonicity of definite responses. Fulfillment is a kind of completeness property for introspection, in the sense that, if M derives ϕ, there is some resource bound at which it will also derive $\Box\phi$ (or $\neg\Box\phi$, if ϕ is not a belief).

[2] This type of nonmonotonic reasoning differs from that of McCarthy [16], which is based instead on the notion of a minimal model of a theory. McCarthy is not concerned with the problem of resource-limited derivation, but rather with the inability or undesirability of stating all conditions which do not obtain in a situation.

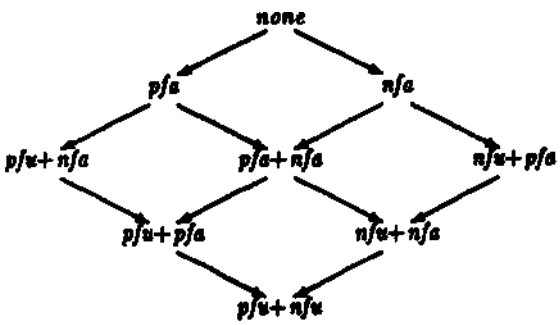

Figure 2: Lattice of domination relations

Fulfillment and faithfulness are not independent. For example, if we take the contrapositive of *pfu*, we get (for arbitrary R) $\sim \mathbf{IM}(\phi, R)$:yes $\rightarrow \sim \mathbf{M}(\phi, R)$:yes. But if IM responds no to a query ϕ, it never responds yes to the same query, so we also have $\exists r.\mathbf{IM}(\phi, r)$:no $\rightarrow \sim \mathbf{IM}(\phi, R)$:yes. By the elementary laws of propositional logic, we then can derive $\exists r.\mathbf{IM}(\phi, r)$:no $\rightarrow \sim \mathbf{M}(\phi, R)$:yes from the contrapositive of *pfu*, and this is the condition *nfa*, because R is arbitrary. In a similar manner, we can show that negative fulfillment implies positive faithfulness (*nfu* $\rightarrow$ *pfa*).

These two relations are the only interdependencies of the constraints. There are thus nine distinct combinations that can be arranged in a lattice as in Figure 2.2. The arrows indicate domination relations; the constraint *pfu+nfu* is thus the strongest of the possible conditions on introspective belief, in the sense that every introspective belief subsystem that obeys it also obeys every other possible combination of the faithfulness and fulfillment constraints. Note that positive fulfilled systems dominate negative faithful ones, and negative fulfilled systems dominate positive faithful ones.

Example 1 The use of introspective belief subsystems as a descriptive model of human belief will be illustrated with one example, drawn from Hintikka [3]. He argues that if someone believes ϕ, she also believes that she believes it (at least in the absence of strict resource limitations on reasoning). This is our condition of positive fulfillment, where the resource R is always taken to be arbitrarily large, and we consider only the first level of introspection (M and IM). Hintikka goes on to argue that people will often have false ideas about their own beliefs, e.g., an utterance of the form

> *S believes that eke believes that* ϕ *although she does not believe it* (1)

can be a true statement about the state of S's beliefs.[3] In terms of the introspective model, we would say that human belief subsystems are not positive faithful (and hence not negative fulfilled).

[3]This is sentence 83 on page 126 of Hintikka [S].

There is an additional curiousity to Hintikka's theory. Although the first level of introspection is characterized as being positive fulfilled but not necessarily positive faithful, it appears that subsequent levels are considered to be totally faithful. For example, the utterance

> 5 *believes the following: that she believes that she believes* ϕ*, although she does not believe it* (2)

which is the statement of (1) as applied to S's idea of herself, is taken to be always false. In our introspective model, this is a statement about self-belief sentences of the introspective machine IM. To capture this behavior, we simply let IM's concept of self-belief be positive faithful.

2.3 Computational Issues

We now present some of our computational results on introspective machines. Generally, we are interested in the problem of converting a nonintrospective belief subsystem into an introspective one; one can imagine retrofitting an existing knowledge base with a mechanism for reasoning about its own beliefs. The questions we pose will have the following form: given a particular introspective constraint (a point in the lattice of Figure 2.2), and perhaps other conditions on nonintrospective behavior, can we implement a belief subsystem obeying these constraints? That is, we would like to find an algorithm that will return a definite answer (yes or no) to every query, given the constraints, so that the introspective belief subsystem is decidable. We first make this notion of decidability precise for resource-limited agents.

Definition 5 *Let $R(\phi)$ be a function mapping queries into finite resource bounds. A belief subsystem* **M** *is* decidable *if there exists an algorithm and function R for* **M** *such that for all queries ϕ, $\sim \mathbf{M}(\phi, R(\phi))$:und; it is semidecidable whenever ϕ is derivable, $\mathbf{M}(\phi, R(\phi))$:yes; it is undecidable if it is not semidecidable.*

The following proposition relates real and ideal agents.

Proposition 8 *Suppose a real introspective belief subsystem* **M** *obeys the following constraints:*

1. **M** *is consistent.*
2. *The inference rules of* $\mathbf{I}^n\mathbf{M}$ *are complete for all* $n \geq 0$.
3. **M** *is fulfilled (*pfu *and* nfu *hold).*

Then **M** *is an ideal introspective belief subsystem iff it is decidable.*

The proof is to show that all three conditions of an ideal introspective agent in Definition 1 are satisfied. The first two obviously are. By inspection, we note that the constraint *pfu+nfu* means that all $\mathbf{I}^n\mathbf{M}$ have the same behavior; hence the third condition is satisfied. Finally, if M is decidable, there is a function $R(\phi)$ for which M always returns a definite answer; hence the belief set of M is the

same as that of an ideal agent. Note that a real agent is ideal only if she has an algorithm that will decide any query ϕ in the finite resource bound $R(\phi)$. Real agents are always computational.

Now let us assume the first two conditions of Definition 8 hold, and explore the computational nature of belief systems obeying various introspection conditions. By "nondoxastic M" we mean that the base set of every belief subsystem of M is nondoxastic.

Proposition 9 *Let the introspection constraint be* pfu+nfu *// the underlying language is*

1. *semidecidable,* M *is undecidable;*
2. *propositional, nondoxastic* M *is decidable;*
3. *monadic PC, nondoxastic* M is *undecidable.*

This proposition just collects the results of the last section (Propositions 3-5 with respect to real agents. Note that, except in the case of a propositional language, M must return und for some queries, no matter what resources are available. In these cases, real agents are not even approximations of ideal agents, since there is no limit in which their behavior becomes the same.

Now suppose we are given a nonintrospective belief subsystem M(whose base set is nondoxastic), and we are asked to construct an introspective subsystem M' whose first component is M. We are free to choose the introspective components, as long as they satisfy conditions (1) and (2) of Proposition 8. The following proposition tells us the best we can do in terms of satisfying various introspective constraints.

Proposition 10 *Suppose the underlying language of* M *is decidable. Then if the introspection constraint is*

1. pfu+nfu, M' *is undecidable;*
2. pfu+pfa, M' *can be semidecidable;*
3. nfa+ pfa, M' *can be decidable.*

The first result is simply (1) of Proposition 9. The second says that if we only want to enforce positive fulfillment and positive faithfulness, the best we can do is to construct an introspective subsystem that is semidecidable. And finally, if the introspection constraint is simple faithfulness, we can construct a decidable M'. Of course, we can do better than this for particular underlying language* (e.f., propositional), but there exists a decidable language for which these bounds are strict (namely, monadic PC).

Let us put these results into perspective. If we are given a nonintrospective agent whose inference rules are complete and whose beliefs are decidable, the best we can do in retrofitting introspective reasoning is to make the agent's self-beliefs faithful. However, if we start with an agent whose rules are incomplete, or we are willing to give up completeness, we can enforce stricter introspective constraints. But now these constraints are relative to a much weaker notion of belief derivation. For example, suppose an agent has no inference rules at all, so that her only nonintrospective beliefs are the base sentences. Certainly we can form a decidable introspective belief subsystem in which *pfu+nfu* holds; $\Box\phi$ is a belief if ϕ is a member of the base sentences, and $\neg\Box\phi$ is a belief if not, and membership in the finite base set is decidable.

3 Comparison to Related Work

Our definition of an ideal introspective agent has many points of similarity with work by Halpern and Moses [2] and Moore [12]. In both these latter cases an underlying propositional language is used, and beliefs sets are defined nonconstructively as *stable sets* (Stalnaker [13], although his original definition did not include consistency).

Definition 6 *A* stable set *S obeys the following constraints:*

1. *S is consistent.*
2. *S is closed under truth-functional consequence*
3. *If $\phi \in S$, then $\Box\phi \in S$; and if $\phi \notin S$, then $\neg\Box\phi \in S$.*

Now we would like an ideal rational agent's beliefs to be a stable set. To build an agent with ideally rational beliefs, we require favorable answers to the following questions.

(a) Given a sentence α that represents the initial beliefs of an agent, what is the appropriate stable set containing a that should be the belief set of the agent?

(b) Is there an algorithm for computing it?

The answer to (a) is not as simple as might be supposed, because it involves finding a stable set that includes α, and makes the fewest assumptions about what the agent believes in addition to α. The presence of doxastic sentences in α complicates matters, and indeed Halpern and Moses differ from Moore in identifying an appropriate belief set. However, if α is consistent and nondoxastic, both approaches converge on a single stable set. Further, this stable set is identical to the belief set of an ideal introspective agent with base set α, so that by Proposition 4 there exists an algorithm for deciding membership in the stable set (the algorithm D^e of Halpern and Moses [2] decides the stable set in this case).

From Definition 1 and Proposition 2, an ideal introspective subsystem, if it exists, is a stable set. For the propositional case it yields the appropriate stable set in the sense of question (a) above, taking α to be the base set of M. Now we can use the results of Section 2.1 to analyze the computational nature of stable sets in the case of quantified languages. By Proposition 5, even for the

relatively simple case of monadic PC and nondoxastic a, the question of membership in the stable set is undecidable. Thus for these systems we must answer question (b) in the negative.

4 Conclusion

We have developed a theory of introspection based on the idea that an agent can use a model of her own belief subsystem to reason about self-belief. The theory can serve as a descriptive tool, since we can describe agents with varying degrees of self-knowledge; hence it may be useful to researchers interested in modelling the cognitive state of users *(e.g.,* in domains such as natural-language systems, tutoring systems, intelligent front ends to databases, and so on). The theory also is a guide to building agents with introspective capabilities, or retrofitting these capabilities onto existing artificial agents.

Introspective belief subsystems can be related to the standard propositional modal logics for belief, weak $S4$ and $S5$ (the axiom schema $\Box p \supset p$ is discarded). An ideal introspective agent is described by weak $S5$ plus a consistency axiom $\Box p \supset \neg\Box\neg p$, since by Proposition 2 both $\Box p \supset \Box\Box p$ and $\neg\Box p \supset \Box\neg\Box p$ are true of such agents. An introspective agent with complete inference rules obeying *p/s* is described by weak $S4$ plus consistency. There are no standard epistemic logics for agents which simply obey the faithfulness constraint; we could construct these by adding $\Box\Box p \supset \Box p$ and $\Box\neg\Box p \supset \neg\Box p$ to the modal logic K.

There are many interesting questions about introspective subsystems that have not been answered in this paper, especially relating to ideal agents. There is obviously a close connection between our definition of an ideal introspective agent and the autoepistemic theories of Moore [12], yet we have compared them only for the case of nondoxastic base sets. Given a (perhaps doxastic) sentence α, Moore defines T to be a *stable expansion* of α if T is equal to the set of truth-functional consequences of

$$\{\alpha\} \bigcup \{\Box p : p \in T\} \bigcup \{\neg\Box p : p \notin T\}.$$

Some sentences have no stable expansions, some have just one, and some have more than one. For example, $\alpha = (\neg\Box P \supset Q) \wedge (\neg\Box Q \supset P)$ has two stable expansions, one containing P, the other Q. What happens to an ideal introspective subsystem when α is its base set? Given the query P, $\mathbf{M}$ will try to prove $\neg\Box Q$, and issue the query Q to $\mathbf{IM}$. $\mathbf{IM}$ will then try to prove $\neg\Box P$, and issue the query P to $\mathbf{I^2M}$. Thus there is no terminating derivation of P, and similarly for Q. However, at some point we could notice that the query delivered to $\mathbf{I^nM}$ is exactly the same as that for $\mathbf{I^{n-2}M}$, and decide that there is no derivation of the query. If we decide this when the query is P, we will get the stable set containing Q; conversely, if we decide that Q is not derivable, we will arrive at the stable set containing P.

Although this example is suggestive, we do not yet have any definitive results on the relationship between Moore's autoepistemic theories and ideal introspective subsystems.

References

[1] Collins, A. M., Warnock, E., Aiello, N. and Miller, M. (1975) Reasoning from Incomplete Knowledge. In *Representation and Understanding,* Bobrow, D. G., and Collins, A., eds., Academic Press, New York.

[2] Halpern, J. Y. and Moses, Y. (1984). Towards a Theory of Knowledge and Ignorance: Preliminary Report. Computer Science Research Report RJ 4448, IBM Research Laboratory, San Jose, California.

[3] Hintikka, J. (1962). *Knowledge and Belief.* Cornell University Press, Ithaca, New York.

[4] Konolige, K. (1984). Belief and Incompleteness. Artificial Intelligence Center Technical Note 319, SRI International, Menlo Park, California.

[5] Konolige, K. (1984). *A Deduction Model of Belief and its Logics.* Doctoral thesis, Stanford University Computer Science Department Stanford, California.

[6] Kripke, S. A. (1959). A Completeness Theorem in Modal Logic. *Journal of Symbolic Logic* 24, pp. 1 14.

[7] Kripke, S. A. (1962). The Undecidability of Monadic Modal Quantification Theory. *Zeitschrift fur Mathematische Logik and Grundlagen der Mathematik* 8, pp. 113-116.

[8] Kripke, S. A. (1963). Semantical considerations on modal logics. *Acta Philsophica Fennica* 16, pp. 83-94.

[9] Levesque, H. J. (1982). A Formal Treatment of Incomplete Knowledge Bases. FLAIR Technical Report No. 614, Fairchild, Palo Alto, California.

[10] McCarthy, J. (1980). Circumscription — A Form of Nonmonotonic Reasoning. *Artificial Intelligence* 13, pp. 27-39.

[11] Moore, R. C. (1980). Reasoning About Knowledge and Action. Artificial Intelligence Center Technical Note 191, SRI International, Menlo Park, California.

[12] Moore, R. C. (1983). Semantical Considerations on Nonmonotonic Logic. Artificial Intelligence Center Technical Note 284, SRI International, Menlo Park, California.

[13] Stalnaker, R. (1980). A Note on Non-monotonic Modal Logic. Unpublished manuscript, Department of Philosophy, Cornell University.