# 2

# A first-order formalisation of knowledge and action and action for a multi-agent planning system

K. Konolige
Artificial Intelligence Center
SRI International, USA

## 1. INTRODUCTION

We are interested in constructing a computer agent whose behaviour will be intelligent enough to perform cooperative tasks involving other agents like itself. The construction of such agents has been a major goal of artificial intelligence research. One of the key tasks such an agent must perform is to form plans to carry out its intentions in a complex world in which other planning agents also exist. To construct such agents, it will be necessary to address a number of issues that concern the interaction of knowledge, actions, and planning. Briefly stated, an agent at planning time must take into account what his future states of knowledge will be if he is to form plans that he can execute; and if he must incorporate the plans of other agents into his own, then he must also be able to reason about the knowledge and plans of other agents in an appropriate way. These ideas have been explored by several researchers, especially McCarthy & Hayes (McCarthy & Hayes 1969) and Moore (Moore 1980).

Despite the importance of this problem, there has not been a great deal of work in the area of formalizing a solution. Formalisms for both action and knowledge separately have been examined in some depth, but there have been few attempts at a synthesis. The exception to this is Moore's thesis on reasoning about knowledge and action (Moore 1980), for which a planner has been recently proposed (Appelt 1980). Moore shows how a formalism based on possible-world semantics can be used to reason about the interaction of knowledge and action. In this paper we develop an alternative formalism for reasoning about knowledge, belief, and action; we show how this formalism can be used to deal with several well-known problems, and then describe how it could be used by a plan constructing system.

### 1.1 Overview and Related Work

We seek a formalization of knowing and acting such that a description of their

interaction satisfies our intuitions. In the first section, we present a basic formalism for describing an agent's static beliefs about the world. We take a *syntactic* approach here: an agent's beliefs are identified with formulas in a first-order language, called the *object language* (OL). Propositional attitudes such as knowing and wanting are modelled as a relation between an agent and a formula in the OL. By introducing a language (the *metalanguage*, or ML) whole prime object of study is the OL, we are able to describe an agent's beliefs as a set of formulas in the OL, and express partial knowledge of that theory. An agent's reasoning process can be modelled as an inference procedure in the OL: from a base set of facts and rules about the world, he drives a full set of beliefs, called his *theory of the world.*

The syntactic approach to representing propositional attitudes is well-known in the philosophy literature, and in the artificial intelligence field McCarthy (McCarthy 1979) has developed a closely related approach. The formalism developed here differs mainly in that it explicitly identifies propositional attitudes as relations on sentences in an object language, and uses provability in the OL as the model of an agent's reasoning process. We are able to present quite complex deductions involving the beliefs of agents (see the *Wise Man Puzzle* in Appendix A, for example) by exploiting the technique of **semantic attachment** to model directly an agent's reasoning process. We are indebted to Weyhrauch (Weyhrauch 1980) for an introduction to this technique, and for the general idea of using ML/OL structures to represent agents.

Finally, our work differs from McCarthy's in its careful axiomatization of the relation between ML and OL, and incorporates solutions to several technical problems, including reasoning about *belief-nesting* (beliefs about beliefs; Creary (Creary 1979) has also described a solution, and a cleaner approach to representing quantified OL expressions in the ML. (This latter subject is not directly relevant to this paper, and will be reported in (Konolige 1981).)

An alternative to the syntactic approach to representing propositional attitudes is the *possible-world* approach, so called because it utilizes Kripke-type possible-world semantics for a modal logic of knowledge and belief. Moore (Moore 1980) has shown how to reason efficiently about propositional attitudes by using a first-order axiomatization of the possible-world semantics for a modal logic. Our objections to the possible-world approach are twofold: first, the possible-world semantics for representing propositional attitudes is complex and at times unintuitive; to deduce facts about an agent's knowledge, one must talk about the possible-worlds that are compatible with what the agent knows. Ultimately, we suspect that the syntactic approach will prove to be a simpler system in which to perform automatic deduction, but further research in both areas is needed to decide this issue. A second objection is that it seems to be difficult to modify possible-world semantics for the modal logic to model adequately inference processes other than logical deduction. The possible-world approach uses the modal axiom that every agent knows the consequences of his knowledge, and this is obviously not true, if only because real agents have

resource limitations on their reasoning processes. The syntactic approach does not suffer from this criticism, because it is possible to describe explicitly in the ML the inference procedure an agent might use.

The second part of this paper integrates the syntactic approach to representing knowledge and belief with a *situation calculus* description of actions (McCarthy & Hayes 1969). We concentrate on many of the interactions between knowledge and action presented in Moore's thesis (Moore 1980). Simply stated, Moore's account is that an agent's beliefs in any situation arise from at least three sources: direct observation of the world, persistence of beliefs about previous situations, and beliefs about what events led to the current situation. By formalizing this assumption, he shows how to model in an intuitively plausible way the knowledge an agent needs to perform actions, and the knowledge that he gains in performing them. Although we subscribe to his notions on how knowledge and action should interact, for the reasons stated above we feel that the possible-world approach Moore uses to formalize these ideas, while elegant, may not have the same intuitive appeal as the syntactic approach.

The main contribution of this paper is to show that the syntactic approach, when integrated with a situation calculus description of actions, can adequately formalize Moore's criteria for the interaction of knowledge and belief. An important benchmark is to formalize the idea of a test: an agent can perform an action and observe the result to figure out the state of some unobservable property of the world. We conclude the second section with just such an example.

In the third section we consider the application of these results to a planning system, in particular one that would require an agent to take account of other agents' plans in forming his own. We come to the conclusion that such a planning system may not be significantly different from current situation calculus planners in its method of search, but does require considerably more sophistication in the deductions it performs at each node in that search.

## 2. AGENTS' BELIEFS AND FIRST-ORDER THEORIES

In this section we lay the basic groundwork for our syntactic approach to representing and reasoning about agents' beliefs. We will model an agent's beliefs about the world as a set of statements (or *theory*) in some first-order language with equality. This is not to say that an agent actually represents the world as a set of first-order statements; we are not concerned here with the details of the internal representation of a computer or human agent with respect to its environment. All we seek is a way of modelling the beliefs of an agent in a manner that will make reasonable predictions about the agent's behaviour, and still be formally tractable. To this end we assume that we can represent an agent's beliefs about the world as a set of statements in a first-order language, and model the derivation of new beliefs by an agent as an inference process in those statements.

Consider an example from the blocks-world domain; let $A_0$ be the name of an agent. $A_0$ will have some set of beliefs about the state of the blocks-world. We represent $A_0$'s beliefs as a list of well-formed formulas (wffs) in a first-order

language with equality. We call this list of wffs $A_0$'s *theory of the world*. For example, suppose $A_0$ believes that block $B$ is on block $C$, and that he is holding block $D$. Then we would have:

$A_0$'s Theory of the Blocks-World
$ON(B,C)$
$HOLDING(A_0, D)$

where *ON* and *HOLDING* have the appropriate interpretations.

Besides specific facts about the state of the world, $A_0$ also has some general rules about the way the world is put together. For instance, $A_0$ may know the rule that if any block $x$ is on any block $y$, then the top of $y$ is not clear. Using this rule together with specific beliefs about the world, he may be able to deduce that $C$ is not clear. This can be modelled as a process of extending $A_0$'s initial set of beliefs about the world to include the deduced information:

| $A_0$'s Facts and Rules about the World | $\Rightarrow$ | $A_0$'s Theory of the World |
|---|---|---|
| $ON(B,C)$ | | $ON(B,C)$ |
| $HOLDING(A_0,D)$ | | $HOLDING(A_0,D)$ |
| $\forall xy\, ON(x,y) \supset \sim CLEAR(y)$ | | $\forall xy\, ON(x,y) \supset \sim CLEAR(y)$ |
| | | $\sim CLEAR(C)$ |
| | | ... |

Thus an agent's theory of the world will be the closure of a set of facts and rules about the world, under some suitably defined inference procedure. We will call the set of basic facts and rules from which all other beliefs are derived the *base set* of the theory. Note that the inference procedure that derives the consequences of the base set need not be logical deduction; it is readily demonstrated that people do not know all the consequences of their beliefs, that they derive contradictory consequences, etc. We recognize that the problem of deriving the consequences of beliefs for more realistic inference procedures is a thorny and unsolved one, and do not intend to pursue it here. For the purposes of this paper we have chosen logical deduction as the inferential procedure: an agent will be able to deduce the logical consequences of his beliefs.

#### 2.1 Metalanguage and Object Language

If we were always to have complete knowledge of an agent's beliefs, then it would be possible to use a simple list of facts and rules to represent the base set of those beliefs. However, it is often the case that our knowledge is incomplete; we may know that an agent either believes fact $P$ or fact $Q$, but we don't know which. Such a description of an agent's beliefs cannot be modelled by a list of facts. So the modelling process must be extended to a *description* of an agent's beliefs. Since beliefs are wffs in a first-order language, a *metalanguage* can be used to describe a collection of such wffs (Kleene 1967). The basic idea is to have terms in the metalanguage to denote syntactic expressions in the first-order language used to encode an agent's beliefs. The latter first-order language

is called the *object language*, or OL, since it is the object of study of the metalanguage (ML). Predicates in the metalanguage are used to state that an expression of the object language is in an agent's theory of the world. The full expressive power of the metalanguage is available for describing a given theory of the object language.

It is natural to choose a first-order language for the metalanguage, since we will be interested in proof procedures in the ML as well as the OL. Let ML be a sorted, first-order language with variables restricted to range over particular sorts. The domain of discourse of the ML will be both the syntactic expressions of the OL, as well as the domain of discourse of the OL. Thus the ML will be able to state relationships that hold between OL expressions and the actual state of the world.

A basic division of sorts of the ML is between terms that denote individuals in the world, and terms that denote expressions in the OL. Among the former will be terms that denote agents ($A_0, A_1, \ldots$) and agents' theories of the world; all these will be called $T_I$ terms. We will use the function *th* of one argument, an agent, to denote that agent's theory of the world.

The other major sort of terms will denote *formulas* of the OL; these will be referred to as $T_F$ terms. Restricting our attention for the moment to sentential formulas of OL, there will be terms in ML that denote propositional letters in OL, and constructors in ML for putting together more complicated formulas from these letters. For example, $P'$ in ML denotes the propositional letter $P$ of the OL,† and the ML term $and(P',Q')$ denotes the sentence $P \wedge Q$ of the OL. These ML constructors from an *abstract syntax* (McCarthy 1962) for OL expressions.

Writing names of formulas using *and*, *or*, *not*, and *imp* as constructors is somewhat cumbersome. For the most part we will use a syntactic abbreviation, enclosing an OL formula in **sense quotes**,‡ to indicate that the standard ML term for that formula in intended. For example, we will write:

$$\ulcorner P \wedge Q \urcorner \quad \text{for} \quad and(P',Q')$$
$$\ulcorner P \supset (Q \vee R) \urcorner \quad \text{for} \quad imp(P',or(Q',R'))$$
and so on.

The rule for translating sense-quote abbreviations into $T_F$ terms of the ML is to replace each predicate symbol $P$ of the sense-quote expression by the ML term symbol $P'$, and each Boolean connective by the corresponding ML Boolean constructor. As more sorts are introduced into the ML we will extend the sense-quote convention in various ways.

Finally, we introduce the ML predicates *TRUE*, *FACT*, and *PR*, each of which has an OL formula as one of its arguments. $TRUE(f)$, where $f$ is an OL formula, means that $f$ is actually true in the world under consideration. It is often

† The general convention will be to use primed terms in ML to denote the corresponding unprimed formulas in OL.

‡ They are called sense-quotes to indicate that the sense of the expression is wanted, rather than its truth-value. In (Kaplan 1971) these are called Frege quotes.

the case that we will want to describe a certain condition actually holding in the world, independent of whether some agent believes it or not; for instance, this is critical to our reasoning about events in the next section, where events are defined as transformations from one state of the world to another.

We intend *TRUE* to have the normal Tarskian definition of truth, so that the truth-recursion axioms are valid. Let the variables $f$ and $g$ range over OL expressions. Then we can write the metalanguage axioms for truth-recursion in the object language as follows:

$$\begin{aligned} &\forall f \sim TRUE(f) \equiv TRUE(not(f)) \\ &\forall fg\, TRUE(f) \vee TRUE(g) \equiv TRUE(or(f,g)) \\ &\forall fg\, TRUE(f) \wedge TRUE(g) \equiv TRUE(and(f,g)) \\ &\forall fg\, TRUE(f) \supset TRUE(g) \equiv TRUE(imp(f,g)). \end{aligned} \qquad (TR)$$

$FACT(t,f)$, where $t$ is an OL theory, means that $f$ is one of the base set formulas of the theory (from which the rest of the theory will be derived by deduction). Using *FACT*, agent $A_0$'s previously exhibited beliefs about the world could be described by the following ML predicates:

$$\begin{aligned} &FACT(th(A_0), \ulcorner ON(B,C) \urcorner) \\ &FACT(th(A_0), \ulcorner HOLDING(A_0,D) \urcorner) \\ &FACT(th(A_0), \ulcorner \forall xy\, ON(x,y) \supset \sim CLEAR(y) \urcorner). \end{aligned}$$

The last *FACT* predicate describes a rule that agent $A_0$ believes.

One special type of *FACT* that we will make frequent use of is a formula known to all agents. We define the predicate *CFACT* on OL expressions to mean that a true expression is a *FACT* for all agents, that is, a *Common FACT*:

$$\forall f\, CFACT(f) \supset \forall a\, FACT(th(a), f) \wedge TRUE(f). \qquad (CF1)$$

*CF1* doesn't completely axiomatize what we intend a common fact to be, however, since it doesn't say that every agent knows that every agent knows that every agent knows $f$, etc. But a fuller characterization of *CFACT* must wait until the technical machinery for describing belief-nesting is developed in a later subsection.

$PR(t,f)$ means that $f$ is provable in the theory $t$. As discussed previously, we will assume that *PR* gives the closure of sentences in OL that can be generated by logical deduction from an original set of *FACT*s. A simple axiomatization of *PR* can can be given for Hilbert-style (assumption-free) proofs. There is only one rule of inference, *Modus Ponens*:

$$\forall tfg\, PR(t,imp(f,g)) \wedge PR(t,f) \supset PR(t,g) \qquad (MP)$$

that is, from $P \supset Q$ and $P$ in the OL, infer $Q$. Since every *FACT* is an initial theorem of the theory, we assert that each of these is provable:

$$\forall tf\, FACT(t,f) \supset PR(t,f). \qquad (FP)$$

And in each theory the logical axioms of a Hilbert system need to be asserted; we assume a sufficient set for the sentential case.

*MP* and the Hilbert axioms will be used in ML proofs of the provability of OL statements; these axioms simulate a Hilbert-type proof system for an OL theory. This simulation is necessary because in general there will be an incomplete ML description of the OL theory, rather than a simple list of *FACT*s for that theory. In those special cases when a list of *FACT*s is available, it is possible to run the proof procedure on the OL theory directly. That is, since the intended meaning of the *PR* predicate is provability in the OL theory, we can check whether the *PR* predicate holds in the ML by running a theorem-prover in the OL. It also isn't necessary to use a Hilbert system, and we will feel free to exploit any system of natural deduction that is sound. The technique of using a computable model of the intended interpretation of a predicate to determine the truth of formulas involving the predicate is called *semantic attachment* (Weyhrauch 1980), and it will be used extensively to simplify proofs in later sections.

The provability predicate *PR* does not have the same characteristics as *TRUE*, and this is important in representing beliefs. For example, the fact that $P$ is not provable doesn't imply that $\sim P$ is provable. If we identify provability with belief, $\sim PR(th(A_0),\ulcorner P\urcorner)$ asserts that $P$ is not one of $A_0$'s beliefs about the word, but this does not imply $PR(th(A_0),\ulcorner\sim P\urcorner)$, i.e., that $A_0$ believes $\sim P$. Also, it is possible to express that either $A_0$ believes that $C$ is clear, or he believes that $C$ is not clear:

$$PR(th(A_0),\ulcorner CLEAR(C)\urcorner) \vee PR(th(A_0),\ulcorner\sim CLEAR(C)\urcorner) ;$$

this says something quite different from $PR(th(A_0),\ulcorner CLEAR(C) \vee \sim CLEAR(C)\urcorner)$; the latter is a tautology that every agent believes, while the former says something a lot stronger about $A_0$'s beliefs about the world.

Parallelling the truth recursion axioms *TR*, we can state rules for the provability of compound OL expressions in terms of their immediate subexpressions. Because of the nature of provability, the axioms for negation, disjunction, and implication, unlike their truth-theoretic counterparts, are not equivalences.

$$\begin{aligned} &\forall tsf \sim PR(t,f) \subset PR(t,not(f)) \\ &\forall tsfg[PR(t,f) \vee PR(t,g)] \supset PR(t,or(f,g)) \\ &\forall tsfg[PR(t,f) \wedge PR(t,g)] \equiv PR(t,and(f,g)) \\ &\forall tsfg[PR(t,f) \supset PR(t,g)] \subset PR(t,imp(f,g)) \ . \end{aligned} \qquad (PR)$$

These are all deducible from the logical axioms in the Hilbert proof system; for instance, the last assertation is just a restatement of *Modus Ponens*.

Another interesting connection between the *PR* and *TRUE* predicates can be drawn by looking at models of the OL. Suppose we have used *FACT* and *PR* to describe an agent's theory $T$ of the world. There will be some set of models that satisfy $T$, i.e., for which all of $T$'s theorems hold. The actual world will be one of these models just in case all $T$'s theorems hold for the world. This condition is statable in the ML as:

$$\forall f PR(T,f) \supset TRUE(f) \ .$$

In general this assertion will not be valid, that is, an agent's beliefs need not correspond to the actual world. By introducing the predicate *TRUE* in the ML, we are able to state the correspondence between a given theory of the world and the actual state of affairs in the world.

### 2.2 Knowledge and Belief

The *PR* and *TRUE* predicates can be used to state our fundamental definitions of knowing and believing for an agent. $BEL(a,f)$ means that agent $a$ believes $f$; $KNOW(a,f)$ means that agent $a$ knows $f$. Then we have the definitions:

$$\begin{aligned} &\forall a f\, BEL(a,f) \equiv PR(th(a),f) \\ &\forall a f\, KNOW(a,f) \equiv BEL(a,f) \wedge TRUE(f)\ . \end{aligned} \tag{B1}$$

That is, we identify belief with provability in an OL theory, and knowledge as a belief that actually holds in the world. In model-theoretic terms, a sentence is known to an agent if the sentence holds in all of his models, and the actual world is a model for that sentence. The definition of a common fact in *CF*1 means that all common facts are known to all agents.

We already know that the inference process used in deriving new beliefs from old ones is only approximated as logical consequence, yet we should still expect this approximation to correctly model some of the characteristics we attribute to belief. For instance, if a rational agent believes that $P \supset Q$, and he doesn't believe $Q$, then it should be the case that he doesn't believe $P$. Translating to the above notation yields the sentence:

$$BEL(A_0,\ulcorner P \supset Q \urcorner) \wedge \sim BEL(A_0,\ulcorner Q \urcorner) \supset \sim BEL(A_0,\ulcorner P \urcorner)\ .$$

To illustrate the use of axioms for belief and provability given so far, we exhibit a natural deduction proof of this sentence in ML.

| | | |
|---|---|---|
| 1. | $BEL(A_0,\ulcorner P \supset Q \urcorner)$ | given |
| 2. | $PR(th(A_0),\ulcorner P \supset Q \urcorner)$ | 1,*B*1 |
| 3. | $\sim BEL(A_0,\ulcorner Q \urcorner)$ | given |
| 4. | $\sim PR(th(A_0),\ulcorner Q \urcorner)$ | 3,*B*1 |
| 5. | $PR(th(A_0),\ulcorner P \urcorner) \supset PR(th(A_0),\ulcorner Q \urcorner)$ | 2,*PR* |
| 6. | $\sim PR(th(A_0),\ulcorner P \urcorner)$ | 4,5 contrapositive |
| 7. | $\sim BEL(A_0,\ulcorner P \urcorner)$ | 6,*B*1 |

This particular proof in the ML cannot be done by semantic attachment to the OL, because it involves reasoning about what isn't provable in the OL theory.

At this point we have presented the basic ideas and definition for a syntactic approach to representing and reasoning about agent's beliefs. The rest of this section is devoted to exploring various technical issues that arise when extending the previous analysis to talking about individuals.

### 2.3 Individuals

By restricting ourselves to the case of sentential formulas in OL, we have been

able to present the basic concepts for representing the beliefs of an agent more simply. Additional complications arise when dealing with terms in the OL that denote individuals rather than truth-values. But a ML encoding of these terms is necessary in order to express such concepts as *agent $A_0$ knows who B is*.

To talk about the individuals that the OL refers to, we introduce an additional sort into the ML, whose denotation will be the *function terms* of the OL. This sort will be called $T_T$, and consists of the following members:

(1) variables $\alpha, \beta, \ldots$;
(2) $\{f^n(t_1, \ldots, t_n)\}$, where $t_i \in T_T$ [n-ary OL function];
(3) $\eta(t)$, where $t \in T_I$ [the 'standard name' function]; (*TT*)
(4) nothing else.

The ML variables $\alpha, \beta, \ldots$, range over OL function terms. For example, we can state that $A_0$ believes a particular block is on $C$ by asserting the ML expression:

$$\exists \alpha \; BEL(A_0, ON'(\alpha, C')) .$$

In this expression there are two ML terms in $T_T$, namely, $\alpha$ and $C'$. $C'$ is a 0-ary function (or constant) in $T_T$ that denotes the constant term $C$ in OL.† $ON'$ is a type of ML term that hasn't been used explicitly before; it is a member of $T_F$ because it names an OL formula. It takes two arguments, each of which is an ML term denoting an OL term, and constructs an OL formula that is the OL predicate $ON$ of these arguments. So the ML term $ON'(\alpha, C')$ denotes the OL expression $ON(\mathcal{A}, C)$ where $\mathcal{A}$ is the OL term denoted by $\alpha$.

It is now possible to give a full definition of $T_F$ terms:

(1) variables $f, g, \ldots$;
(2) $\{f^n(t_1, \ldots, t_n)\}$, where $t_i \in T_F$ [Boolean constructors, e.g., *and*];
(3) $\{g^n(t_1, \ldots, t_n)\}$, where $t_i \in T_T$ [predicate constructors, e.g., $ON'$];
(4) nothing else.

and $T_I$ terms:

(1) variables $x, y, \ldots$;
(2) $\{f^n(t_1, \ldots, t_n)\}$, where $t_i \in T_I$ [individual constants and functions];
(3) $\Delta(t)$, where $t \in T_T$ [the denotation function]; (*TI*)
(4) nothing else.

We will also find it convenient to extend the notion of sense-quote abbreviations to handle ML terms involving $T_T$ variables. The previous rules are expanded in the following way: all function symbols in the sense-quote expression are replaced by their primed forms, while any symbols used as variables in the surrounding ML expression remain unchanged. For example, the sense-quote expression in $\exists \alpha \; KNOW(A_0, \ulcorner ON(\alpha, b(C)) \urcorner)$ is to be understood as a syntactic

† We extend the prime convention to cover ML terms in $T_T$ as well as $T_F$; that is, $t'$ in ML denotes the unprimed term $t$ in OL.

abbreviation for the ML term $ON'(\alpha, b'(C'))$. We have not yet said what happens to $T_I$ variables in sense-quote expressions; this must wait until standard names are explained in the next subsection.

The introduction of $T_T$ terms into the ML completes the descriptive power of ML for OL expressions. It also lets us handle some of the well-known denotational puzzles in the philosophy literature. One of the simplest of these is the Morningstar-Eveningstar description problem. Both Morningstar and Eveningstar are actually the planet Venus seen at different times of the day. An agent $A_0$ believes that they are not the same; further, he doesn't have any knowledge about either being the planet Venus. Let *MS*, *ES*, and *VENUS* be OL terms that denote the Morningstar, the Eveningstar, and Venus, respectively. The following set of ML formulas describes this situation:

$$TRUE(\ulcorner ES = VENUS \urcorner)$$
$$TRUE(\ulcorner MS = VENUS \urcorner)$$
$$BEL(A_0, \ulcorner MS \neq ES \urcorner)$$
$$\sim BEL(A_0, \ulcorner ES = VENUS \urcorner)$$
$$\sim BEL(A_0, \ulcorner MS = VENUS \urcorner) \ .$$

It is perhaps easiest to explain this set of sentences in model-theoretic terms. The intended interpretation of the OL terms, *ES*, *MS*, and *VENUS* is the same object, namely the planet Venus. The two *TRUE* predicates establish this, since they assert that these three terms denote the same individual in the world. On the other hand, the first *BEL* predicate asserts that in the models of $A_0$'s theory of the world, *MS* and *ES* denote different individuals. This means that the actual world cannot be among the models of this theory. Further, the last two *BEL* predicates assert that *ES* and *MS* are not provably equal to *VENUS* in this theory; hence there will be some models of the theory for which $ES = VENUS$ holds, some for which $MS = VENUS$ holds, and some for which neither holds. From this we conclude that not only is $A_0$ mistaken as to the equality of *ES* and *MS*, he also is unsure about whether either is the same as *VENUS*. McCarthy (1979) lists some other philosophical puzzles that can be handled in a syntactic formulation.

### 2.4 Knowing Who Someone Is

One of the problems that any formal treatment of belief must confront is that of describing when an agent knows who or what something is. For example, the following two English sentences say something very different about the state of $A_0$'s knowledge:†

(1) "$A_0$ knows who murdered John."
(2) "$A_0$ knows that someone murdered John."

The police would certainly be interested in talking to $A_0$ if the first statement were true, while the second statement just means that $A_0$ read the local tabloid.

† A similar problem appears in (Quine 1971).

We might paraphrase the first statement by saying that there is some individual who murdered John, and $A_0$ knows who that individual is. The second statement can be true without $A_0$ having any knowledge about the particular individual involved in the murder.

How is the distinction between the two sentences above to be realized in this formalism? The second sentence is easy to represent:

$$BEL(A_0, \ulcorner \exists\, x\, MURDERED(x, JOHN) \urcorner) \ . \qquad (W1)$$

This simply says that $A_0$ believes in the existence of an individual who murdered John. It might be supposed that the first sentence could be represented in the following way:

$$\exists\, \alpha\ BEL(A_0, \ulcorner MURDERED(\alpha, JOHN) \urcorner) \qquad (W2)$$

$W2$ says that there is a *MURDERED* predicate in $A_0$'s theory of the world relating some individual ($\alpha$'s denotation) and John. Unfortunately, this isn't quite strong enough; if the denotation of $\alpha$ is the OL term *murderer(JOHN)*, then $W2$ is virtually a tautology, and doesn't say that $A_0$ knows who murdered John. Indeed, if the OL expression in $W1$ is skolemized, it becomes obvious that $W1$ and $W2$ are equivalent.

What seems to be going on here is that different names have a different status as far as identifying individuals is concerned. "Bill" is a sufficient description for identifying John's murderer, whereas "John's murderer" is not. The question of what constitutes a sufficient description is still being debated in the philosophical literature. But for the purposes of this paper, it will suffice if we have a name that is guaranteed to denote the same individual in every model of the OL. By asserting a predicate involving this name in $A_0$'s theory of the world, it will be possible to encode the fact that $A_0$ believes that predicate for the given individual. Names that always denote the same individual are called *standard names.*

The formal method of establishing standard names is straightforward. Consider the set of all individuals involved in the situation we wish to consider.† Include in the OL a set of constant symbols, the *standard name symbols*, to be put in one-one correspondence with these individuals. The language OL will be partially interpreted by specifying this correspondence as part of any model of the language; this means that the only models of OL we will consider are those that are faithful to the standard name mapping.

In the metalanguage, we introduce the *standard name function* $\eta$ of one argument (see the definition of $T_T$ terms above). This function returns the standard name of its argument. Generally we will use lowercase Greek letters from the later part of the alphabet as ML variables for OL standard names [$\mu$, $\nu$, ...]. The metalanguage statement of "$A_0$ knows who the murderer of John is" then becomes:

$$\exists x \mu\ (\eta(x) = \mu) \wedge KNOW(A_0, \ulcorner MURDERED(\mu, JOHN) \urcorner) \ . \qquad (W3)$$

† We restrict ourselves to countable sets here.

Because $\mu$ denotes a standard name, the only models of this statement are those in which the same individual $x$ murdered John. This is in contrast to $W1$ and $W2$ above, which allow models in which any individual murdered John. An immediate consequence is that $W1$ and $W2$ are derivable from $W3$, but not the other way round.

So in order to assert that $A_0$ knows who or what some individual $B$ is, we write in the ML:†

$$\exists x \mu\ (\eta(x) = \mu) \wedge KNOW(A_0, \ulcorner B = \mu \urcorner)\ .$$

By modifying the sense-quote translation rules slightly, it is possible to write OL expressions involving standard names much more compactly. The modification is to assume that any ML variable of type $T_I$ occurring within a sense-quote gets translated to the standard name of that variable. With this rule, for example, the above assertion comes out as $\exists x\, KNOW(A_0, \ulcorner B = x \urcorner)$.

We will use the predicate $KNOWIS(a,\beta)$ to mean that the agent $a$ knows who or what the OL term denoted by $\beta$ refers to. The definition of $KNOWIS$ is:

$$\forall a \beta\, KNOWIS(a,\beta) \equiv \exists x\, KNOW(a, \ulcorner \beta = x \urcorner)\ . \qquad (KW)$$

Note that the property of being a standard name is a relation between a term of the OL and models of this language, and hence cannot be stated in the OL. The use of a metalanguage allows us to talk about the relation between the OL and its models.

One of the proof-theoretic consequences of using standard names is that every theory can be augmented with inequalities stating the uniqueness of individuals named by standard names. In the metalanguage, we write:

$$\forall x y\, x \neq y \supset \forall t\, PR(t, \ulcorner x \neq y \urcorner)\ . \qquad (SN)$$

Formally, the definition of a standard name can be axiomatized in the ML by introducing the denotation function $\Delta$.‡ $\Delta(\alpha)$, where $\alpha$ denotes an OL term, is the denotation of $\alpha$ in the actual world; it is the inverse of the standard name function, since it maps an OL term into its denotation. There is an intimate relation between the denotation function and equality statements in OL formulas describing the world:

$$\forall \alpha \beta\, TRUE(\ulcorner \alpha = \beta \urcorner) \equiv \Delta(\alpha) = \Delta(\beta) \qquad (D1)$$

that is, two OL terms are equal in the actual world just in case they denote the same individual; $D1$ can be viewed as a definition of the intended interpretation of equality. The prime purpose of the denotation function is to tie together the

† This analysis essentially follows that of (Kaplan 1971), with the extension of standard names to all individuals in the domain, rather than just numbers and a few other abstract objects. There are problems in using standard names for complex individuals, however (see Kaplan 1971).

‡ This is Church's denotation predicate in function form (Church 1951); since a term can have only one denotation, it is simpler to use a function.

denotation of terms in the OL and the ML. For standard names, it can be used to state that the denotation of a standard name is the same individual in all situations, something that cannot be done with equality predicates in the OL:

$$\forall x\ \Delta(\eta(x)) = x\ . \qquad (D2)$$

For example, by asserting $\eta(VENUS) = VENUS'$ in ML, we fix the denotation of the OL term $VENUS'$ to be the individual denoted by the ML term *VENUS* in all models of the OL.

The introduction of standard names with fixed denotations across all models makes the task of relating the OL to the ML easier. By introducing this 'common coin' for naming individuals, we are able to write expressions of the OL that represent beliefs without constantly worrying about the subtle consequences of the denotational variance of terms in those expressions. Standard names will play an important role in describing belief-nesting (beliefs about beliefs), in describing executable actions, and in simplifying the deduction process.

#### 2.5 The Object Language as Metalanguage

In this subsection we extend the OL to include a description of another object language OL′. Thus extended, the OL can be viewed as a metalanguage for OL′. The reason we want to do this is that it will be necessary for representing an agent's view of a world that is changing under the influence of events. In the next section we will show how an agent can model the way in which the world changes by describing what is true about different states of the world connected by events. But to describe these states of the world, or *situations*, the agent's theory must talk about sentences of another language holding in a given situation.

Before trying to extend the formal apparatus of the OL to describe another OL, it is helpful to examine more closely the relation between the ML as a means of studying the OL and as a means of describing the actual world. This is because the structure of an ML/OL pair will be very similar no matter what the depth of embedding; and the simplest such structure to study is obviously the topmost one. Although we initially characterized the ML's domain of discourse as including that of the OL, it appears that we have not made much use of this characterization. In describing the models of OL, however, it was necessary to pick out the model that was the actual world; this was done with the predicate *TRUE*. And it was impossible to state the definition of a standard name without appealing to terms in the ML that referred to individuals in the actual world. So, in fact, we have already used the ML to characterize the actual state of the world and the individuals that populate it.

We have stated that agent's beliefs are represented as first-order therories of the world. The ML is, by the above argument, just such a theory; but whose theory of the world is it? One useful interpretation is to take what we will call the *egocentric view*: a theory in the ML is identified as the theory of a particular agent. That is, suppose we were to build a computer agent and invest him with an ML/OL structure as a way of representing other agent's beliefs. Then the

nonlogical axioms of the ML would constitute the computer agent's theory of the world. The interpretation of the ML predicate *TRUE* would be "what the computer agent believes about the world", and of the predicate *KNOW*, "what another agent believes that agrees with what the computer agent believes." In this interpretation, there is no sense of absolute truth or knowledge; the beliefs of one agent are always judged relative to those of another.

Suppose we identify the agent $A_0$ with the ML; what interpretation does the OL theory $th(A_0)$ have? Interestingly enough, it is $A_0$'s introspective description of his own beliefs. Unlike other agent's theories of the world, $th(A_0)$ shares an intimate connection with formulas that hold in the ML. For a rational agent, it should be the case that if he believes $P$, then he believes that he believes $P$. We can state this connection by the following rule of inference:

> Belief attachment: If the agent $a$ is identified with the ML, then from $TRUE(f)$ infer $BEL(th(a), f)$ .

Introspection will be useful when we consider planning, because a planning agent must be able to reflect on the future state of his beliefs when carrying out some plan.

If the metalanguage is intended to describe the actual world, then it is reasonable to ask what the relation is between models of the ML and models of its OL, and whether this connection can be formalized in the ML. We start by adding predicate symbols to the ML whose intended meaning is a property of the actual world, rather than of the OL and its models. Consider such a predicate $P$ of no arguments, and let its intended meaning be "222 Baker Street Apt 13 is unoccupied:" that is, the actual world satisfies $P$ just in case this apartment is indeed unoccupied. In the OL there is also a predicate symbol $P$ of no arguments whose meaning we wish to coincide with that of the ML predicate $P$. The fact that these symbols are the same is an orthographic accident; they come from different languages and there is thus no inherent connection between them. However, because the ML can describe the syntax and semantics of the OL, it is possible to axiomatize the desired connection. Let $P'$ be the ML term (in $T_F$) denoting the OL predicate $P$. Then $P$ in the ML and OL have the same meaning if:

$$P \equiv TRUE(P') \qquad (R1)$$

is asserted in the ML. For suppose the actual world satisfies $P$ in the ML; then $TRUE(P')$ must also hold, and hence by the meaning of *TRUE*, the actual world is also a model for $P$ in the OL. Similarly, if the actual world falsifies $P$ in the ML, $TRUE(not(P'))$ must hold, and the actual world falsifies $P$ in the OL also. So the proposition named by $P'$ holds just in case Apt. 13 at 222 Baker Street is unoccupied, and thus the meanings of $P$ in the ML and $P$ in the OL coincide.

For predicates that have arguments, the connection is complicated by the need to make sure that the terms used in the ML and OL actually refer to the same individuals. So, for example, if $P$ is an ML predicate of two arguments that we wish to mean the same as the OL predicate $P$, we would write:

$$\forall \alpha \beta \, TRUE(\ulcorner P(\alpha,\beta) \urcorner) \equiv P(\Delta(\alpha),\Delta(\beta)) ; \qquad (R2)$$

that is, since the denotation function $\Delta$ gives the individuals denoted by the OL terms $\alpha$ and $\beta$, $P$ in the ML agrees with $P$ in the OL on these individuals. Using standard names, $R2$ could be rewritten as:

$$\forall x y \, P(x,y) \equiv TRUE(\ulcorner P(x,y) \urcorner) \qquad (R3)$$

since, by $D2$, $\Delta(\eta(x)) = x$, $\Delta(\eta(y)) = y$. Note that the standard name convention for sense-quotes is in force for $R3$.

Using *TRUE* and equivalence, axioms like $R3$ cause predicate symbols to have a 'standard meaning' across the ML and OL, in much the same way that $D2$ formalizes standard names using the denotation function and equality. But while nonstandard names are a useful device for encoding an agent's beliefs about individuals that the agent may have misidentified (recall the Morningstar–Eveningstar example), nonstandard predicates don't seem to serve any useful purpose. So we will assume that for every predicate symbol $P$ in ML, there is a function symbol of the form $P'$ whose denotation is the OL predicate $P$, and there is an axiom of the form $R2$ equating the meaning of these predicates.

To make the OL into a metalanguage for OL′, we simply introduce sorts that denote OL′ expressions into the OL, in exactly the same way that it was done for the ML. In addition, the various axioms that tie the ML and OL together (*MP*, $D1$, etc.) must also be asserted in the OL. Unfortunately, this also means that the ML itself must have a new set of terms denoting terms in the new OL; the machinery for describing embedded ML/OL chains rapidly becomes confusing as the depth of the embedding grows. So in this paper we will supply just enough of the logical machinery to work through the examples by introducing two conventions; readers who want more detail are referred to Konolige (1981).

We will extend the convention of sense-quote abbreviation to include ML variables of the sort $T_F$ (denoting formulas of the OL). When these occur in sense-quotes, they are to be translated as the *standard name* of the variable; hence they denote the name of an expression. To take an example, we will complete the axiomatization of *CFACT*:

$$\forall f \, CFACT(f) \supset CFACT(\ulcorner CFACT(f) \urcorner) \qquad (CF2)$$

$CF2$ asserts that if $f$ is a common fact, then every agent knows it is a common fact. The sense-quote term $\ulcorner CFACT(f) \urcorner$ denotes the OL expression $CFACT(f')$, where $f'$ is the standard name of the OL′ expression corresponding to $f$.

Finally, every axiom is a common fact:

$$CFACT(\ulcorner A \urcorner), \text{ A an axiom.} \qquad (CF3)$$

In practice, we hope that the depth of embedding needed to solve a given problem will be small, since the complexity needed for even the three-level structure of ML, OL, and OL′ is substantial. Also, the technique of semantic

attachment can be used to reduce the complexity of reasoning about embedded structures by attaching to a particular level of an embedded structure and reasoning in that language. In Appendix A we use embedded ML/OL structures to solve the wise man puzzle, which involves reasoning to a depth of embedding of three (ML, OL, and OL'); we exploit semantic attachment to simplify the reasoning involved.

## 3. THE INTERACTION OF ACTIONS AND BELIEFS

The previous section laid the groundwork for a syntactic treatment of knowledge and belief in a static world. This must be integrated with a formal treatment of actions in order to accomplish our original task of formalizing the interaction of knowledge and action. We examine the following two questions:

- What knowledge is required by an agent successfully to perform an action?
- What knowledge does an agent gain in performing an action?

The methodology we will use is to apply the *situation calculus* approach (McCarthy and Hayes 1969) first to formally describe the way in which the world changes as events occur. It will then be assumed that this formal system is a reasonable approximation to the way an agent reasons about changes in the world: this means that it becomes part of an agent's rules about the world. By simply attributing a facility for reasoning about events to agents, it turns out that we are able to answer both these questions formally, and that this formalization corresponds well with our intuitions about real agents. This is essentially the same method that was used by Moore (Moore 1980); here, we show that it can be successfully carried out for a syntactic formalization of knowledge and belief.

Once the formal requirements for reasoning about events have been specified, we consider how an agent might plan to achieve a goal using his knowledge of actions. We conclude that planning is inherently a process of self-reflection: that is, in order to construct a plan, an agent must reflect on what the state of his beliefs will be as the plan is undergoing execution. Such a self-reflection process is represented naturally by an ML/OL structure in which the planning agent is identified with the ML, and his future states are theories of the OL. We will show how it is possible to construct plans within this representation, and extend it to include plans that involve other cooperative agents.

### 3.1 Situations

In the situation calculus approach, events are taken to be relations on *situations*, where situations are snapshots of the world at a particular moment in time. It is natural to identify situations with models of a language used to describe the world; in this case, we will use the language OL of the previous section, because the ML for describing models of the OL is already laid out. In the ML, situations will be named by terms, generally the constants $\{S_0, S_1, \ldots\}$. A formula $f$ of the OL holds in a situation $s$ when the situation satisfies $f$; the ML predicate $H(s,f)$

will be used to indicate this condition. If the situation $S_0$ is singled out as being the actual world (and the initial world for planning problems), then *TRUE* can be defined in terms of $H$:

$$\forall f\, TRUE(f) \equiv H(S_0, f) \ . \qquad (H1)$$

Since $H$ describes satisfiability in a model, the truth-recursion axioms *TR* are valid for $H$ as well as *TRUE*.

If we consider agents to be part of domain of discourse, then their beliefs can change from one situation to the next, just as any other inessential property of an agent might. But if an agent's beliefs change from situation to situation, then the theory that is used to model these beliefs must also change. One way to represent an agent's changing beliefs is to ascribe a different theory to an agent in each situation to model his beliefs in that situation. In the ML, we will write $ths(a,s)$ to denote agent $a$'s beliefs in situation $s$; if $S_0$ is taken to be the actual world, then it is obvious that $\forall a\, ths(a,S_0) = th(a)$.

But we might now ask what situation the expressions in each of these theories are about. Suppose that the OL sentence $P$ is a member of $ths(A_0, S_1)$, and thus one of $A_0$'s beliefs in situation $S_1$. We would naturally want $P$ to be property that $A_0$ believes to hold of situation $S_1$ (and not $S_0$ or some other situation). That is, $ths(a,s)$ represents agent $a$'s beliefs *in* situation $s$, *about* situation $s$. In informal usage we will call the situation we are focussing on the *current situation*, and say 'the agent $a$ in situation $s$' when we are referring to the agent's beliefs in that situation. Later we will show how to represent an agent's beliefs about situations other than the one he is currently in.

For each situation, an agent's beliefs in that situation are specified by a theory. Given this arrangement, we define the new predicates $B$ and $K$ as similar to *BEL* and *KNOW*, but with a situation argument:

$$\begin{aligned} &\forall a s f\, B(a,s,f) \equiv PR(ths(a,s), f) \\ &\forall a s f\, K(a,s,f) \equiv B(a,s,f)\, H(s,f) \ . \end{aligned} \qquad (B2)$$

$B(a,s,f)$ means that in situation $s$ agent $a$ believes that $f$ holds in $s$; $K$ is similar, with the condition that $f$ actually holds in $s$. Note that the underlying predicates *FACT* and *PR* do not have to be changed, since they are defined on theories of OL rather than models. Thus the properties of *BEL* and *KNOW* described in the previous section also hold for $B$ and $K$ in any particular situation. *BEL* and *KNOW* can be defined as $B$ and $K$ in the situation $S_0$.

Several extensions to the formalism presented in the first section must be made to deal with situations. A new denotation function $\delta$ takes a situation argument as well as an OL term; $\delta(s,\alpha)$ is the denotation of $\alpha$ in situation $s$. $\Delta(\alpha)$ gives the denotation of a $\alpha$ in situation $S_0$, and is definable as $\delta(S_0,\alpha)$. The appropriate forms of $D1$ and $D2$ are:

$$\begin{aligned} &\forall s \alpha \beta\, H(\ulcorner s, \alpha = \beta \urcorner) \equiv \delta(s,\alpha) = \delta(s,\beta) \\ &\forall s x\ \ \delta(s, \eta(x)) = x \ . \end{aligned} \qquad (D3)$$

This last says that standard names always have the same interpretation in every situation. Nonstandard names can change their denotation in different situations, e.g., the block denoted by "the block $A_0$ is holding" may be changed by $A_0$'s actions.

Finally, we require the appropriate versions of $R1$-$R6$, where these axioms are appropriately generalized to refer to all situations.

### 3.2 Observables

Following Moore (Moore 1980), we recognize three ways that an agent can acquire beliefs in a situation:

- He can observe the world around him.
- His beliefs about past situations persist in the current situation.
- He can reason about the way in which the current situation arose from events that occurred in previous situations.

In the next few subsections we describe how an agent's beliefs persist and how he reasons about events; here we formalize what it means for a property of the world to be observable.

It is certainly true that there are many properties of the world we live in that are not directly observable; for example, consider a gas oven whose pilot light is completely encased and hence not visible. Whether this pilot light is on or off isn't an observable property, but there are other observations that could be made to test what the state of the pilot light is, e.g., by turning on the oven and observing whether it lights. What we actually consider to be observable depends on how we formalize a given problem domain; but it is important for a planning agent to be able to make the distinction between properties of the world he can observe directly, and those he must infer.

One of the reasons that it is handy to have a separate theory representing the beliefs of an agent in each situation is that we then have a way of describing the effect of observable properties on an agent's beliefs. Formally, we can state that a property is observable by asserting that in every situation, subject to certain preconditions that are required for the felicitous observation of the property, an agent knows whether that property holds or not. For example, in the OL let $o$ be an oven, and let $LIT(o)$ mean that $o$ is lit. Then $LIT(o)$ is asserted to be observable by:

$$\forall a o s\, H(s,\ulcorner AT(a,o)\urcorner) \supset [K(a,s,\ulcorner LIT(o)\urcorner) \vee K(a,s,\ulcorner \sim LIT(o)\urcorner)]; \quad (O1)$$

that is, if the agent is actually at the oven, he knows either that it is lit, or that it is not lit. Recall from the previous section on knowledge and belief that $O1$ says something very strong about the state of $a$'s knowledge, and is not derivable from the tautology $K(a,s,\ulcorner LIT(o) \vee \sim LIT(o)\urcorner)$.

### 3.3 Events Types

Event types are relations on situations; a given event type describes the possible states of the world that could result from an event occurring in any initial state.

We will use the three-place predicate $EV$ in the metalanguage to describe event types: $EV(e,s_i,s_f)$, where $e$ is an event type and $s_i$ and $s_f$ are situations, means that $s_f$ results from an event of type $e$ occurring in $s_i$. An event is an instance of an event type,† but generally we will not have to distinguish them for the purposes of this paper, and we will use 'event' for 'event type' freely.

Generally the events of interest will be agents' actions, and these will be constructed in the ML using terms representing actions, agents, and the objects involved in the action (the *parameters* of the action). If *act* is an action, then $do(a,act)$ is the event of agent $a$ performing this action. Consider the situation calculus axiomatization of a simple blocks-world action, $puton(x,y)$, where the parameters of the action are blocks:

$$\forall axys_i s_f\, EV(do(a,puton(x,y)),s_i,s_f) \supset H(s_i,\ulcorner CLEAR(y)\urcorner) \wedge H(s_i,\ulcorner HOLDING(a,x)\urcorner) \wedge H(s_f,\ulcorner ON(x,y)\urcorner) \wedge H(s_f,\ulcorner \sim HOLDING(a,x)\urcorner) \qquad (PO1)$$

$$\forall axys_i s_f\, EV(do(a,puton(x,y)),s_i,s_f) \supset [\forall f\, SAF(f) \wedge f \neq \ulcorner CLEAR(y)\urcorner \wedge f \neq \ulcorner HOLDING(a,x)\urcorner \supset H(s_i,f) \equiv H(s_f,f)] \qquad (PO2)$$

The form of $PO1$ is conditional, so the right-hand side describes the conditions under which situations $s_i$ and $s_f$ are related by the event of $a$ putting $x$ on $y$. The first two conjuncts on the right-hand side are essentially preconditions for the event to occur, since they state conditions on the intial situation $s_i$ that must be satisfied for $EV$ to hold. The preconditions are that $CLEAR(\eta(y))$ and $HOLDING(\eta(a),\eta(x))$ must hold in situation $s_i$; note that the standard names for the parameters are indicated by the sense-quote convention. If the preconditions are not met, then there is no situation $s_f$ that is the successor to $s_i$ under the event $e$. The rest of the conjuncts describe which formulas of the OL are to hold in the new situation $s_f$.

$PO2$ specifies that all formulas of a certain type that hold in $s_i$ are also to hold in $s_f$. It is thus a *frame axiom* for the event $e$, describing which aspects of the situation $s_i$ remain unchanged after the event occurs. The predicate $SAF$ stands for *S*imple *A*tomic *F*ormula; it picks out those formulas of the OL that are composed of atomic predicates over standard names. Although $SAF$ applies only to non-negated atomic formulas, the frame axiom carries over negated atomic formulas as well, since $H(s,not(f))$ is equivalent to $\sim H(s,f)$.‡ Among

† For example, "Borg's winning of Wimbledon yesterday was fortuitous" is a statement about a single event, but "Borg winning Wimbledon has happened five times" describes an event type that had five particulat instances.

‡ The axiomatization of events given here is a standard one in the AI literature on formal planning, and there are well-known problems involving the use of frame axioms like the one above. We are not attempting to add any new insight to this particular aspect of planning; but we are interested in having a formal description of events to integrate with our theory of belief, and this seems to be the best formulation currently available.

the nicer features of this axiomatization is that events whose outcomes are conditional on the initial state can be easily described. For instance, consider the event of an agent turning on a gas oven that has a pilot light. If the pilot light is on, the oven will be lit; if the pilot light is off, the oven will have whatever status, lit or unlit, it had before the event occurred (the oven may already have been on). Let $PL(o)$ be an OL predicate meaning "the pilot light of oven $o$ is on"; and let $LIT(o)$ mean "oven $o$ is lit". Then the event of an agent turning on $o$ can be described as:

$$\begin{aligned}\forall a s_i s_f o\, EV(do(a, light(o)), s_i, s_f) \supset \\ H(s_i, \ulcorner AT(a,o) \urcorner) \wedge \\ H(s_i, \ulcorner PL(o) \urcorner) \supset H(s_f, \ulcorner LIT(o) \urcorner) \wedge \\ H(s_i, \ulcorner \sim PL(o) \urcorner) \supset [H(s_f, \ulcorner LIT(o) \urcorner) \equiv H(s_i, \ulcorner LIT(o) \urcorner)]\end{aligned} \tag{LT1}$$

$$\begin{aligned}\forall a s_i s_f o\, EV(do(a, light(o)), s_i, s_f) \supset \\ [\forall f SAF(f) \wedge f \neq \ulcorner LIT(o) \urcorner \supset H(s_i, f) \equiv H(s_f, f)]\ .\end{aligned} \tag{LT2}$$

The second conjunction of $LT1$ gives the result of the event on case the pilot light is on: the oven will be lit. The third conjunction says that if the pilot light is off, the oven will be lit in $s_f$ just in case it was lit in $s_i$, i.e., its status doesn't change. $LT2$ is the frame axiom.

### 3.4 Reasoning about Situations and Events

The axiomatization of events as relations on situations enables us to talk about what is true in the world after some events have occurred starting from an initial situation (which we will generally take to be $S_0$). What it doesn't tell us is how an agent's beliefs about the world will change; nothing in the *PO* or *LT* axioms gives any insight into this. It might be suspected that, as events are described by axioms as changing the actual state of the world, this description might be extended to cover agents' *theories* as well, e.g., changing $A_0$'s theory in situation $S_0$ ($ths(A_0, S_0)$) into his theory in situation $S_1$ ($ths(A_0, S_1)$).† But there is no obvious or well-motivated way to make modifications to axioms like *PO* and *LT* so that they take into account agents' beliefs about a situation rather than what actually holds in the situation.‡ What is needed here is a principled way of deriving the changes to an agent's beliefs that result from an event, given a description of the event as a relation on situations. Credit for the recognition of

† Indeed, it might be though that the most widely known AI planning system, *STRIPS*, has just such a mechanism in its add/delete list approach to describing events. However, closer examination reveals that because *STRIPS* makes the assumption that it has a *partial model* in the sense of (Weyhrauch 1980), and it is actually slightly less descriptive than the situational approach described above (Nilsson 1980).

‡ There is one proposal that is suggested by the our use of $H$ to refer to the actual situation and $PR$ to statements that an agent believes about a situation, namely, to replace all predicates involving $H$ with the corresponding ones involving $PR$. However, it can be shown that the substitution of $PR(ths(A_0, s), \ldots)$ for $H(s, \ldots)$ yields counterintuitive results for $A_0$'s beliefs.

this problem belongs to Robert Moore, and we will formalize the solution he presented in his thesis, the main points of which follow.

The solution to this difficulty lies in making the observation that agents are reasoning entities. Consider how agent $A_0$ might reason about some event $E$; let us suppose the event is that agent $A_0$ turned on the oven in situation $S_0$, and that the result was that the oven was not lit in situation $S_1$. What should $A_0$'s beliefs be in situation $S_1$? First, by observation, he knows that the oven isn't lit. He also believes (in $S_1$) that the current situation resulted from the event $E$ occurring in situation $S_0$. So $A_0$ reasons as follows: if, in situation $S_0$, the pilot light of the oven had been on, then in $S_1$ the oven would be lit, since he turned it on. But the oven isn't lit; hence the pilot light couldn't have been on in $S_0$, and remains not on in $S_1$.

There are several important things to note about this analysis. The first is that, as suggested previously, $A_0$'s beliefs in situation $S_1$ comes from only three sources: observation ("the oven is not lit"), persistence of beliefs about previous situations ("if in $S_0$ the pilot light had been on ..."), and beliefs about the way events change the world. This latter is equivalent to having some form of $LT1$ as part of $A_0$'s beliefs in situation $S_1$. From these three sources $A_0$ is able to generate a new set of beliefs for $S_1$.

The second thing to note is that none of $A_0$'s reasoning in $S_1$ could have taken place unless he believed that $S_1$ resulted from $S_0$ via the event $E$. Beliefs about what sequence of events led to the current situation play a very important role in reasoning about that situation, and, like other beliefs, they can be mistaken or inferred from other evidence. Suppose, for example, that $A_0$ suddenly sees the oven become lit. He might infer that the only way that could happen when it wasn't previously lit would be for an agent to turn it on; this is inferring that the situation where the oven is lit is connected by a certain event with a previous situation where the oven wasn't lit. We will not be concerned with this kind of inference here, although we note the possibility of doing *event recognition* in this framework. The events we are interested in are actions, and the assumption we will make for the remainder of this paper is that an agent knows what action it is that he performs in executing a plan.

A third aspect of this reasoning that is unusual is that the axiomatization of events is being used in a different way than a planning program would normally consider doing. Typically, a planner uses an event description like $LT1$ to form plans to light the oven, and the side condition that the pilot light be on is one of the things that can go wrong with the plan, and so must be taken into account as a subgoal. However, in the above example $A_0$ has used $LT1$ to reason about a property of the world that is not available to his direct observation, that is, as a *test*. This is an important characteristic for any formalism that combines a description of agent's beliefs with a description of events; a single description of an event should suffice for an agent to reason about it either as a means of effecting a change in the world, or as a test that adds to his beliefs about the world.

Finally, the precondition that $A_0$ be at the oven to turn it on translates naturally in this analysis into a precondition on $A_0$'s beliefs in situation $S_0$. If $A_0$ is to reason that situation $S_1$ is the successor to $S_0$ under the event $E$, he must believe that he was actually at the oven in situation $S_0$. For if he doesn't believe this, then he cannot use $LT1$ to infer anything about the results of his action.

We might summarize the analysis of this section in the following way: by making the simple assumption that an agent reasons about the way in which situations are related by events, we are able to characterize in a natural way the belief preconditions required for executing an action, and the effects of actions on the subsequent belief state of an agent. The interaction of observation and reasoning about situations gives an agent the power to plan actions that perform tests, as well as change the state of the world.

### 3.5 Formalizing Agents' Reasoning about Events

We now give a formalization that implements the ideas just laid out. The first requirement is that we be able to describe an agent $a$ in situation $s$ reasoning about other situations, especially the one just preceding. Since the formulas of $ths(a,s)$ all refer to properties of situation $s$, we must enrich the OL so that formulas in the OL can refer to different situations. Using the techniques of belief-nesting of the previous section, we add to the OL the predicate $H$ corresponding to the ML predicate of the same name. The OL expression $H(S_1,\ulcorner P\urcorner)$ means that the OL' formula $P$ holds in situation $S_1$, regardless of what theory this formula appears in.† With the addition of the $H$ predicate to the OL, the notion that all formulas in $ths(a,s)$ refer to properties of $s$ can be formalized as:

$$\forall sf\ PR(th(a,s),f) \equiv PR(th(a,s),\ulcorner H(s,f)\urcorner)\ . \qquad (H2)$$

$H2$ can be paraphrased by saying that an agent believes $P$ in situation $s$ just in case he believes that $P$ holds in situation $s$. Given $H2$, it is possible to describe agents' theories as consisting purely of formulas in $H$; but the added level of embedding puts this technique at a disadvantage with respect to using other predicates from OL to describe an agent's beliefs about the current situation.

It is also possible to formalize the notion that beliefs about previous situations persist, or are carried over into succeeding situation. Suppose that in situation $S_n$ an agent has a belief of the form, "in a previous situation $S_i$, $P$ was true". Then if $S_{n+1}$ is the successor to $S_n$ under some event, this belief is still valid. Formally, we can assert this with the ML axiom:

$$\forall s_i s_f e\, EV(e,s_i,s_f) \supset [\forall asf\, B(a,s_i,\ulcorner H(s,f)\urcorner) \supset B(a,s_f,\ulcorner H(s,f)\urcorner)]\ . \qquad (H3)$$

† We will take $S_0, S_1, \ldots$ to be standard names for situations in all languages. It will be assumed that standard names are always used to name situations.

The antecedent of the implication says that $s_i$ and $s_f$ must be connected by some event for beliefs to be carried over from $s_i$ to $s_f$; this is necessary because we don't want agents to inherit beliefs from their future states. By phrasing the beliefs in terms of the predicate $H$, $H3$ carries over beliefs about all situations previous to and including $s_i$.

One of the consequences of $H3$ is that once an agent forms a belief about a situation, he holds that belief about that situation for all time. Since beliefs can be mistaken, it might happen that an agent observes something that forces him to revise his previously held beliefs. In that case, $H3$ is too strong, and the resultant theory will be inconsistent. We recognize that the general problem of reconciling inconsistent beliefs that arise from different sources (called *belief revision*) is a hard one, involving both conceptual and technical issues, and it is not part of this research to say anything new about it.† Nevertheless, it is worthwhile to note that because the ML has terms that refer to agents' theories in different situations, it may be possible to describe a belief revision process formally in the ML.

### 3.6 An Example of a Test

Given the preceding techniques for describing what an agent believes to hold in situations other than the one he is currently in, we can show formally that $A_0$ can use the $LT$ axioms as a test to figure out whether the pilot light is on or not. In the initial situation $S_0$, we will assume that $A_0$ knows he is at the oven $O$ (where $O$ is the standard name for the oven), and realizes that it is not lit:

Initial Conditions in the ML

(1) $K(A_0,S_0,\ulcorner AT(A_0,O) \wedge \sim LIT(O)\urcorner)$ given

(2) $K(A_0,S_1,\ulcorner EV(do(A_0,light(O)),S_0,S_1)\urcorner)$ given

The style of proof we will exhibit will be natural deduction, with assumption dependencies noted in square brackets in the justification for a line of the proof. Given the initial conditions, we next show that $A_0$ can observe whether or not the oven is lit in situation $S_1$:

(3) $\forall f SAF(f) \wedge f \neq \ulcorner LIT(O)\urcorner \supset H(S_0,f) \equiv H(S_1,f)$ 2,B2,$LT2$

(4) $SAF(\ulcorner AT(A_0,O)\urcorner)$ definition of $SAF$

(5) $H(S_1,\ulcorner AT(A_0,O)\urcorner)$ 1,3,4,$B2$

(6) $K(A_0,S_1,\ulcorner LIT(O)\urcorner) \vee K(A_0,S_1,\ulcorner \sim LIT(O)\urcorner)$ 5,$O1$

Line 3 comes from the frame axiom for *light*, and lets us infer that $A_0$ is still at the oven in situation $S_1$ (line 5). The observation axiom $O1$ is then invoked to assert that $A_0$ will know what the state of the oven is in that situation.

Throughout this proof, we will be interested in two theories of the OL:

† Doyle (Doyle 1978) worked on this problem under the rubric "Truth Maintenance", and more recent work in nonmonotonic reasoning also considers this problem.

$ths(A_0,S_0)$ and $ths(A_0,S_1)$. Assertions in the ML involving $A_0$'s beliefs can be reasoned about by using semantic attachment to the appropriate OL theory. For example, line 1 above is attached to the following statements in $ths(A_0,S_0)$:

$A_0$'s Theory in Situation $S_0$

(7) $AT(A_0,O) \wedge \sim LIT(O)$ 1,B2, semantic attachment
(8) $H(S_0,\ulcorner \sim LIT(O) \urcorner)$ $1,B2,H2$, semantic attachment

Line 7 is the attachment of line 1 to $A_0$'s theory in $S_0$. Line 8 is derived from line 1 by the use of $H2$; it is useful because it will persist as a belief in the successor situation $S_1$. Generally, beliefs that an agent derives about the current situation can be inherited into succeeding situations by expressing these beliefs with the $H$ predicate.

At this point we do reasoning by cases. First assume the right disjunct of line 6; then for $A_0$'s beliefs in situation $S_1$ we have:

$A_0$'s Theory in Situation $S_1$

(9) $\sim LIT(O)$ [9]: assumed, semantic attachment
(10) $\sim H(S_1,\ulcorner LIT(O) \urcorner)$ [9]: 9, $H2$ semantic attachment
(11) $EV(do(A_0,light(O)),S_0,S_1)$ 2, semantic attachment
(12) $H(S_0,\ulcorner PL(O) \urcorner) \supset H(S_1,\ulcorner LIT(O) \urcorner)$ 11, $LT1$
(13) $\sim H(S_0,\ulcorner PL(O) \urcorner)$ [9]:10,12 contrapositive
(14) $H(S_0,\ulcorner \sim PL(O) \urcorner)$ [9]:13,$TR$ for $H$

The first part of the result is derived by line 14, namely, that if $A_0$ observes that the $O$ is not lit in situation $S_0$, then he knows that the pilot light was not on in situation $S_0$. This sequence of steps is interesting because it illustrates the intermixture of proof techniques in the ML and OL. Lines 9, 10, and 11 come from statements in the ML about $ths(A_0,S_1)$. Line 10 is derived from line 9 in the ML by the application of axiom $H2$. Line 11 says that $A_0$ believes that $S_1$ is the result of the $light(O)$ action occurring in $S_0$, and follows directly from line 2 and semantic attachment. Line 12 follows from line 11 and the event axiom $LT1$; it is assumed that $A_0$ believes this axiom. Finally, 13 and 14 follow, given that the truth-recursion axioms for $H$ are made available in all theories in the OL.

The left disjunct of line 6 can be reasoned about in the following way (since lines 11 and 12 did not involve any assumptions, they can be used in this part of the proof also):

$A_0$'s Theory in Situation $S_1$

(15) $LIT(O)$ [15]: assumed, sem. att.
(16) $H(S_1,\ulcorner LIT(O) \urcorner)$ [15]:15,$H2$, sem. att.
(17) $\sim H(S_0,\ulcorner LIT(O) \urcorner)$ 8, $H3$,$TR$ for $H$, sem. att.
(18) $\sim[H(S_1,\ulcorner LIT(O) \urcorner) \equiv H(S_0,\ulcorner LIT(O) \urcorner)]$ [15]:16,17
(19) $H(S_0,\ulcorner \sim PL(O) \urcorner) \supset H(S_1,\ulcorner LIT(O) \urcorner) \equiv H(S_0,\ulcorner LIT(O) \urcorner)$ 11, $LT1$
(20) $\sim H(S_0,\ulcorner \sim PL(O) \urcorner)$ [15]:18,19 contrapositive
(21) $H(S_0,\ulcorner PL(O) \urcorner)$ [15]:20,$TR$ for $H$

Here again, the first few lines (15, 16, and 17) are established by reasoning at the ML about $ths(A_0,S_1)$. Line 17 comes from an instance of axiom $H3$, which enables an agent's beliefs to persist through a sequence of situations. Line 19 comes from $A_0$'s knowledge of $LT1$, and line 20 is the key step: it establishes that under the assumption of $O$ being lit in $S_1$, the pilot light was on in $S_0$. Finally, the frame axiom $LT2$ will carry the pilot light's status in $S_0$ forward into $S_1$:

$A_0$'s Theory in Situation $S_1$

| | | |
|---|---|---|
| (22) | $\forall f\, SAF(f) \wedge f \neq \ulcorner LIT(O) \urcorner \supset H(S_0,f) \equiv H(S_1,f)$ | $11, LT1$ |
| (23) | $SAF(\ulcorner PL(O) \urcorner)$ | definition of $SAF$ |
| (24) | $H(S_0,\ulcorner PL(O) \urcorner) \equiv H(S_1,\ulcorner PL(O) \urcorner)$ | 22,23 |
| (25) | $PL(O)$ | [15]: 21,24, $H2$ |
| (26) | $\sim PL(O)$ | [9]: 14,24, $H2$ |

Line 25 is under the assumption of the left disjunct of line 6, and line 26 is under the right disjunct. In the ML we can derive several results from the preceding proof structure:

In the ML

| | | |
|---|---|---|
| (27) | $B(A_0,S_1,\ulcorner PL(O) \urcorner) \vee B(A_0,S_1,\ulcorner \sim PL(O) \urcorner)$ | 6,20,21 |
| (28) | $B(A_0,S_1,\ulcorner LIT(O) \urcorner) \supset B(A_0,S_1,\ulcorner PL(O) \urcorner)$ | 15,25 |
| (29) | $B(A_0,S_1,\ulcorner \sim LIT(O) \urcorner) \supset B(A_0,S_1,\ulcorner \sim PL(O) \urcorner)$ | 9,26 |

Line 27 says that in $S_1$, $A_0$ will either believe that the pilot light is on, or he will believe that is not on. Thus, by performing the action of lighting the oven, $A_0$ gains knowledge about the state of an unobservable, the pilot light. This is the desired result of agent $A_0$ using $LT1$ to perform a test of an unobservable property.

Lines 28 and 29 give belief analogues to the $LT$ axioms, which described the event of lighting the oven solely in terms of the actual situations before and after the event. These assertions show how the beliefs of $A_0$ change under the influence of the event $do(a, light(O))$. By suitably generalizing the preceding proof, it can be shown that 28 and 29 hold for all agents and initial situations.

$$\begin{aligned} \forall a o s_i s_f\ & EV(do(a,light(o)),s_i,s_f) \wedge K(a,s_i,\ulcorner AT(a,o) \wedge \sim LIT(o) \urcorner) \supset \\ & B(a,s_f,\ulcorner LIT(o) \urcorner) \supset B(a,s_f,\ulcorner PL(o) \urcorner) \\ & B(a,s_f,\ulcorner \sim LIT(o) \urcorner) \supset B(a,s_f,\ulcorner \sim PL(o) \urcorner). \end{aligned} \qquad (LT3)$$

$LT3$ is valid under the condition that $LT1$ is assumed to be believed by all agents. $LT3$ is one description of the way in which an agent's beliefs change in a situation that results from an oven-lighting event; it would be most useful to a planner as a lemma to be invoked if the state of the pilot were to be tested as a step in a plan. Another lemma about oven-lighting that would be useful to a planner would be one in which the belief preconditions to an action were made explicit; this would be used to plan actions that light the oven.

## 4. PLANS AND PLANNING

In the previous section we saw how to characterize the changes to an agent's beliefs produced by his observation of events. In this section we will consider how to use these results as part of the deductions that an agent needs to do to construct workable plans, i.e., plans that will accomplish their goals.

Consider how an agent might go about constructing workable plans. Using his description of various events (*PO*, *LT*, and others) he can try to find a sequence of actions that lead to the desired goals being true in some final situation. If we identify the planning agent with the ML, then a plan would be a sequence of situations connected by actions performed by that agent, such that the goals are true in the final situation. This doesn't seem to involve the planning agent in any reasoning about his beliefs; all he needs to do is describe how the actual world changes under the influence of his actions.

This isn't the whole story, though. The plan that is derived must be an *executable* plan; that is, if the plan is a sentence of actions, the agent must be able to execute each of those actions at request time. For instance, the action description *light*(*oven*(*John*)) will not be executable if $A_0$ doesn't know which oven is John's. For a plan to be executable by an agent, the agent must know what action is referred to by each of the *do*-terms in the plan. According to a previous section, this means that the agent must have the standard name for the action in his theory. But what are standard names for actions? Following Moore (Moore 1980), we take the viewpoint that actions can be analysed as a general procedure applied to particular arguments, e.g., *puton* is a general procedure for putting one block on top of another, and *puton*(*A*,*B*) is that procedure applied to the two blocks *A* and *B*. If we assume that all agents know what general procedure each action denotes, then the standard names for actions are simply the terms formed by the action function applied to the standard names of its parameters.† The condition that actions be executable forces the planning agent to make the critical distinction between his beliefs at planning time and his beliefs at execution time. A planning agent may not know, as he forms his plan, exactly what action a particular *do*-term in his plan denotes; but if he can show that at the time he is to execute that action, he will know what it is, then the plan is an executable one. Plans of this type occur frequently in common-sense reasoning; consider a typical plan $A_0$ might form to tell someone what time it is. The plan has two steps: first $A_0$ will look at his watch to find out what the time is, and then he will communicate this information to the requestor. At planning time, $A_0$ doesn't really know what the second action

† Actually, the condition that the parameters be standard names is too strong. Standard names have the property that every agent knows whether two individuals named by standard names are the same or not in every situation, but this condition is not strictly necessary for an action to be executable. Consider the action of requesting information from the telephone operator; surely it is not required that an agent be able to differentiate the operator from every other individual in his beliefs. If he were to dial the operator on two separate occasions, he would not necessarily be able to tell if he talked to the same operator or not.

is, because he doesn't know the time, and the time is an important parameter of the communication act. Yet he can reason that after looking at his watch, he will know the time; and so the plan is a valid one.

By this argument, an agent must analyse at planning time what the future states of his beliefs will be as he executes the plan. Thus the planning process intrinsically forces the agent into introspection about his future beliefs. Since we have identified the planning agent with the ML, it is natural to represent his future beliefs during the execution of the plan as OL theories in the situations that the planning process gives rise to. If the planning agent is $A_0$, then these theories are $ths(A_0,S_0)$ (the initial situation), $ths(A_0,S_1)$, etc., where each of the $S_i$ results from its predecessor via the execution of the next action in the plan. $A_0$'s planning process is basically a simulation of the plan's execution in which he reasons about the changes that both the actual world and his set of beliefs will undergo during the course of the plan's execution. By figuring out what his future states of belief will be, he can decide at planning time whether an action of the plan will be executable.

For $A_0$ to take other agents' plans into account in forming his own, he must be able to represent their future states of belief, in addition to his own. But this doesn't involve any additional representational complexity, since $A_0$ is already keeping track of his own beliefs during the simulated execution of the plan. In Konolige and Nilsson [1980] an example of a multi-agent plan is presented; currently we are working on formalizing such plans in the framework presented here.

Actually, this planning process bears a strong resemblance to typical implementations of a situation calculus approach to planning (Warren 1974). In these systems, events are axiomatized along the lines of *PO* and *LT*, and the planner searches for a sequence of situations that leads to the goal by doing theorem-proving with the event axioms; the search space is essentially the same in either approach. The main difference is in the relative complexity of reasoning that the two planning systems must be able to handle. In the approach described here, the effect of actions on the agent's beliefs in each situation greatly increases the deductive complexity of the planner and the work that it must do at each node in the search space of plans. The usefulness of lemmas such as *LT*3 that describe the effects of actions on an agent's belief state now becomes apparent: by summarizing the effect of actions on an agent's beliefs, they reduce the complexity of the deductions that must be performed at each step in the plan. Further savings can be realized by using the method of belief attachment described in the previous section: from $H(s,f)$ at the ML, infer $K(A_0,s,f)$. Most of the work of figuring out $A_0$'s future states of knowledge can be performed by reasoning about $H$ at the metalevel, rather than $K$, and this is considerably simpler. Finally, it should be noted that the executability requirement acts as a filter on plans. Thus a reasonable search strategy would be to first find a plan that works without taking into account its executability (and hence the future belief states of the planning agent), and then test it for executability.

## 5. CONCLUSION

To summarize the contributions of this paper: we have defined a syntactic approach to the representation of knowledge and belief in which the key element is the identification of beliefs with provable expressions in a theory of the object language. The technique of *semantic attachment* to the intended interpretation of the metalanguage provability predicate has been advanced as a method of simplifying proofs by directly modelling an agent's inference procedure, rather than simulating it.

To unify a formalization of knowledge and action, we have shown how to take Moore's account of their interaction and formalize it within the syntactic framework. The benchmark example was a presentation of a *test* in which an agent uses his knowledge of observable properties of the world and the way actions affect the world to discover the state of an unobservable property. Finally, we pointed out how the formalization could be used in a planning system.

While this paper is a step towards showing that the syntactic approach can be extended to an adequate formalization of the interaction of knowledge and action, there is still much work to be done in constructing a practical planner for a multi-agent environment that uses this formalism. Two areas in particular are critical. First, a suitable system for doing automatic deduction in the framework has to be worked out. Although we have advocated semantic attachment as a means of simplifying proofs, we have not yet explored the problem of controlling a deduction mechanism that uses this technique. The second area also involves control issues: how can a planner be designed to search the space of multi-agent possible plans efficiently? One of the ideas suggested by this paper is to derive lemmas of the form of $LT3$ that show the effect of actions on an agent's beliefs. With such lemmas, a planning system would have compiled the necessary results for contructing new brief states from previous ones.

## ACKNOWLEDGEMENTS

This research is a direct outgrowth of research conducted with N. Nilsson (Konolige and Nilsson 1980), and still reflects his influence. S. Rosenschein and N. Nilsson read previous drafts of this paper, and their criticisms and comments have contributed to the final form. Also, I have benefited from talks with P. Hayes, R. Moore, R. Weyhrauch, C. Talcott, and all the members of the planning group at the Artificial Intelligence Center at SRI. This research is supported by the Office of Naval Research under Contract No. N00014-80-C-0296.

## APPENDIX: THE WISE MAN PUZZLE

This is a solution to a simple version of the wise man puzzle, for whose statement we quote from McCarthy *et al.* (1980):

> A king wishing to know which of his three wise men is the wisest, paints white dots on each of their foreheads, tells them that at least one spot is

white, and asks each to determine the colour of his spot. After a while the smartest announces that his spot is white, reasoning as follows: "Suppose my spot were black. The second wisest of us would then see a black and a white and would reason that if his spot were black, the dumbest would see two black spots and would conclude that his spot is white on the basis of the king's assurance. He would have announced it by now, so my spot must be white."

We will simplify this puzzle by having the king ask each wise man in turn if he knows what colour his spot is, staring with the dumbest. The first two say "no", and the last says that his spot is white.

In formalizing the puzzle, we will take the three wise men to be $A_0, A_1$, and $A_2$, in order of increasing stupidity.† We will reason about the puzzle from $A_0$'s point of view, and show that $A_0$ knows that his spot is white after hearing the replies of the other two. We will not be concerned with the axiomatization of the speech act performed by the agents; it will be assumed that $A_0$'s model of the world changes appropriately to reflect this new information.

There are three situations in the puzzle: the initial situation $S_0$, the situation $S_1$ just after $A_2$ speaks, and the situations $S_2$ just after $A_1$ speaks. The frame axioms for these situation are simply that every agent knows he knew in the previous situation; these frame axioms are common knowledge.

We will identify $A_0$ with the ML, so that goal is to show: $H(S_2, \ulcorner W(A_0) \urcorner)$ in the ML. $W(a)$ is the predicate whose meaning is '$a$'s spot is white'. The initial conditions of the problem are:

(1) $W(A_1) \wedge W(A_2)$
(2) $CFACT(\ulcorner W(A_0) \vee W(A_1) \vee W(A_2) \urcorner)$
(3) $CFACT(\ulcorner K(A_2, S_0, \ulcorner W(A_0) \urcorner) \vee K(A_2, S_0, \ulcorner \sim W(A_0) \urcorner) \urcorner)$
(4) $CFACT(\ulcorner K(A_2, S_0, \ulcorner W(A_1) \urcorner) \vee K(A_2, S_0, \ulcorner \sim W(A_1) \urcorner) \urcorner)$
(5) $K(A_1, S_0, \ulcorner W(A_0) \urcorner) \vee K(A_1, S_0, \ulcorner \sim W(A_0) \urcorner)$
(6) $CFACT(\ulcorner \sim K(A_2, S_0, \ulcorner W(A_2) \urcorner) \urcorner)$
(7) $CFACT(\ulcorner \sim K(A_1, S_1, \ulcorner W(A_1) \urcorner) \urcorner)$

Line 1 says that $A_0$ observes white spots on $A_1$ and $A_2$; line 2 asserts that it is common knowledge that at least one spot is white. The next two lines state it is common knowledge that $A_2$ can observe whether the other two agent's spots are white or not. Line 5 says that $A_1$ knows the colour of $A_0$'s spot. And the last two lines express the effect of the first two agent's answers to the king on everyone's knowledge. This axiomatization will be sufficient to prove that $A_0$ knows his spot is white in $S_2$.

The first step in the proof is to show that $A_1$ knows, in situation $S_1$, that either his own or $A_0$'s spot is white; this by reasoning about $A_2$'s answer to the king. We will attach to $A_1$'s theory in situation $S_1$ (that is, $ths(A_1, S_1)$), and do our reasoning there:

† $A_0, A_1$ and $A_2$ are standard names for the wise men.

$A_1$'s Theory in Situation $S_1$

| | | |
|---|---|---|
| (8) | $\sim K(A_2,S_0,\ulcorner W(A_2)\urcorner)$ | 6, semantic attachment |
| (9) | $K(A_2,S_0,\ulcorner W(A_0) \vee W(A_1) \vee W(A_2)\urcorner)$ | 3, semantic attachment |
| (10) | $K(A_2,S_0,\ulcorner(\sim W(A_0) \wedge \sim W(A_1)) \supset W(A_2)\urcorner)$ | 9 |
| (11) | $K(A_2,S_0,\ulcorner \sim W(A_0) \wedge \sim W(A_1)\urcorner) \supset K(A_2,S_0,\ulcorner W(A_2)\urcorner)$ | 10, *MP* |
| (12) | $\sim K(A_2,S_0,\ulcorner \sim W(A_0) \wedge \sim W(A_1)\urcorner)$ | 8,11 contrapositive |

In these lines, we have used the fact that everyone knows that everyone knows common knowledge assertions. At line 12, $A_1$ realizes that $A_2$ doesn't know that both $A_0$ and $A_1$ lack white dots; if he did, he would have announced the fact.

Now $A_1$ uses the common knowledge that $A_2$ can observe the colour of $A_0$'s and $A_1$'s dots to reason that one of the latter has a white dot:

$A_1$'s Theory in Situation $S_1$

| | | |
|---|---|---|
| (13) | $K(A_2,S_0,\ulcorner \sim W(A_0)\urcorner) \wedge K(A_2,S_0,\ulcorner \sim W(A_1)\urcorner)$ | [13]: assumption |
| (14) | $K(A_2,S_0,\ulcorner \sim W(A_0) \wedge \sim W(A_1)\urcorner)$ | [13]:13, *PR* |
| (15) | $\sim K(A_2,S_0,\ulcorner \sim W(A_0)\urcorner) \vee \sim K(A_2,S_0,\ulcorner \sim W(A_1)\urcorner)$ | 13;12,14 contradiction |
| (16) | $\sim K(A_2,S_0,\ulcorner \sim W(A_0)\urcorner)$ | [16]: assumption |
| (17) | $K(A_2,S_0,\ulcorner W(A_0)\urcorner) \vee K(A_2,S_0,\ulcorner \sim W(A_0)\urcorner)$ | 3, semantic attachment |
| (18) | $K(A_2,S_0,\ulcorner W(A_0)\urcorner)$ | [16]: 16,17 |
| (19) | $\sim K(A_2,S_0,\ulcorner \sim W(A_1)\urcorner)$ | [19]: assumption |
| (20) | $K(A_2,S_0,\ulcorner W(A_1)\urcorner) \vee K(A_2,S_0,\ulcorner \sim W(A_1)\urcorner)$ | 4, semantic attachment |
| (21) | $K(A_2,S_0,\ulcorner W(A_1)\urcorner)$ | [19]: 19,20 |
| (22) | $K(A_2,S_0,\ulcorner W(A_0)\urcorner) \vee K(A_2,S_0,\ulcorner W(A_1)\urcorner)$ | 15,16,18,19,21 |
| (23) | $H(S_0,\ulcorner W(A_0) \vee W(A_1)\urcorner)$ | 22, *B*2 |
| (24) | $W(A_0) \vee W(A_1)$ | 23, frame axioms, *R*1 |

We first show here that $A_2$ doesn't know $A_0$'s spot is black, or he doesn't know that $A_1$'s spot is black (line 15). Assertions that follow from assumptions are indicated by a square bracketing of the assumption line number in their justification. Next we do an analysis by cases of line 15; in either case, line 22 holds: $A_2$ either knows $A_0$'s spot is white, or he knows $A_1$'s spot is white. From this $A_1$ concludes that either he or $A_0$ has a white spot (line 24). Note that the frame axioms were needed to show that the $W$ predicate doesn't change from situation $S_0$ to situation $S_1$.

At this point we are through analysing $A_1$'s theory of situation $S_1$, and go back to the ML to reason about situation $S_2$. By line 5, $A_1$ knows the colour of $A_0$'s dot, so we assume that he knows it is black:

At the Metalevel

| | |
|---|---|
| (25) $K(A_1, S_0, \ulcorner \sim W(A_0) \urcorner)$ | [25]: assumption |
| (26) $K(A_1, S_1, \ulcorner \sim W(A_0) \urcorner)$ | [25]: 25, frame axioms |
| (27) $K(A_1, S_1, \ulcorner \sim W(A_0) \supset W(A_1) \urcorner)$ | 24, frame axioms |
| (28) $K(A_1, S_1, \ulcorner W(A_1) \urcorner)$ | [25]: 26, 27, *MP* |
| (29) $\sim K(A_1, S_1, \ulcorner W(A_1) \urcorner)$ | 7, *CF*1 |
| (30) $\sim K(A_1, S_0, \ulcorner \sim W(A_0) \urcorner)$ | 25, 28, 29, contradiction |
| (31) $K(A_1, S_0, \ulcorner W(A_0) \urcorner)$ | 5, 30 |
| (32) $H(S_0, \ulcorner W(A_0) \urcorner)$ | 31, *B*2 |

Under the assumption that $A_1$ knows $A_0$'s spot is black, we derive the contradiction of lines 28 and 29. Therefore, by line 5, it must be the case that $A_1$ knows $A_0$'s spot to be white. This is the conclusion of line 32; since this is one of $A_0$'s beliefs, we are done.

## REFERENCES


Appelt, D., (1980). A planner for reasoning about knowledge and belief, *Proceeding of the First Annual Conference of the American Association for Artificial Intelligence*, Stanford, California.

Church, A., (1951). A formulation of the logic of sense and denotation, *Structure, Method and Meaning*, (Ed. Henle, P., *et al.*). New York: Liberal Arts Press.

Cohen, P. R., & Perrault, C. R., (1979). Elements of a plan-based theory of speech acts, *Cognitive Science*, 3, 52–67.

Creary, L. G., (1979). Propositional attitudes: Fregean representation and simulative reasoning, *IJCAI-6 (Tokyo)*, 176–181.

Doyle, J., (1978). Truth maintenance systems for problem solving, *Memo AI-TR-419*, Cambridge, Mass.: Artificial Intelligence Laboratory, MIT.

Kaplan, D., (1971). Quantifying in, *Reference and Modality*, 112–144, (Ed. Linsky, L.). Oxford: Oxford University Press.

Kleene, S. C., (1967). *Mathematical Logic*, New York: John Wiley.

Konolige, K., & Nilsson, N., (1980). Multiple agent planning systems, *Proceeding of the First Annual Conference of the American Association for Artificial Intelligence*, Stanford, California.

Kowalski, R., (1979). *Logic for Problem Solving*. New York: North-Holland.

McCarthy, J., (1962). Towards a mathematical science of computation, *Information Processing, Proceedings of the IFIP Congress*, 62, 21–28. Amsterdam: North-Holland Publishing Company.

McCarthy, J., & Hayes, P. J., (1969). Some philosophical problems from the standpoint of artificial intelligence, *Machine Intelligence*, 4, 463–502. (Eds. Meltzer, B., and Michie, D.). Edinburgh: Edinburgh University Press.

McCarthy, J., *et al.* (1978). On the model theory of knowledge, *Memo AIM-312*, Stanford: Computer Science Department, Stanford University.

McCarthy, J., (1979). First order theories of individual concepts and propositions, *Machine Intelligence*, 9, 120–147, (Eds. Hayes, J. E., Michie, D., and Mikulich, L. I.). Chichester: Ellis Horwood; and New York: Halsted Press.

Moore, R. C., (1977). Reasoning about knowledge and action, *IJCAI-5*, Cambridge Mass., 223–227. Pittsburgh: Department of Computer Science, Carnegie-Mellon University.

Moore, R. C., (1980). Reasoning about knowledge and action, *Artificial Intelligence Center Technical Note 191*. Menlo Park: SRI International.

Nilsson, N. J., (1980). *Principles of Artificial Intelligence*. Menlo Park: Tioga Publishing Co.

Quine, W. V. O., (1971). Quantifiers and propositional attitudes, *Reference and Modality*, 101–101, (Ed. Linsky, L.). Oxford: Oxford University Press.

Warren, D. H. D., (1974). WARPLAN: A system for generating plans, *DAI Memo 76*. Edinburgh: Department of Artificial Intelligence, University of Edinburgh.
Weyhrauch, R., (1980). Prolegomena to a theory of mechanized formal reasoning, *Artificial Intelligence*, 13.