A Logic for Representing Actions, Beliefs, Capabilities, and Plans

Becky Thomas
Computer Science Department
Stanford University
Stanford, CA 94305
bthomas@cs.stanford.edu

Abstract

When several artificial agents share an environment, they must be able to communicate in order to coordinate their efforts. It is widely recognized that efficient communication requires the ability to reason about other agents' beliefs and plans [6]; we argue that it is also important to reason about the capabilities of other agents.

We present here a logic in which to represent an agent's beliefs, its plans concerning future actions and goals, and its capabilities. While there have been several proposals for representing agents' beliefs and intentions or plans, we believe that it is novel to include representation of agents' capabilities in such a system.

1 Introduction

When we have a group of artificial agents working in the same environment, whether they are working cooperatively on the same project or engaged in separate projects, it will be necessary for them to be able to communicate in order to coordinate their efforts. It is widely recognized that efficient communication requires the ability to reason about other agents' beliefs and plans [6]. We argue that it is also important to reason about the capabilities of other agents. For example, in cooperative work, one agent may request that a second agent perform some task. Certainly the second agent should not agree unless it believes that it will be capable of performing the task. In fact, the first agent may have chosen the second agent based on its belief that the second agent has the requisite capability.

Our motivation stems from our work on Agent-Oriented Programming (AOP) [16], in which agents' internal states consist of sets of beliefs, capabilities, and so forth. Agents communicate by sending typed messages, such as requests, informing messages, responses to requests, and so on. An agent's program is a set of rules for changing its internal state, based on its current internal state and any incoming messages.

Our work has focused on using AOP in domains like building construction, where several agents, each with a different area of expertise (plumbing, electrical systems, *etc.*), work cooperatively. It quickly became obvious that these agents would have to reason about capabilities. It would do no good to request that a plumbing-expert agent do some wiring; rather the agents needed to know which agent is capable of performing wiring tasks. The goal of this paper, then, is to provide a way to model the beliefs, plans, and capabilities of agents, to facilitate communication. Although we will use a logic for this purpose, this is not to say that the agents will use this, or any, logic in their reasoning.

We examine here a snapshot of an agent's mental state. When the snapshot was taken, the agent had a certain model of the world. It had various beliefs about the world and about itself, was capable of performing various actions, and had plans concerning its future actions. We will not address in this paper the question of how these beliefs, capabilities, and plans might change with time, although others involved in the AOP project have considered some of these questions; see [9]. There is a large body of work on belief revision; the reader is referred to [4; 10; 14]. [2], [5], and [15] address the question of how an agent's plans and/or intentions might evolve.

Cohen and Levesque [5] and Rao and Georgeff [15] have made proposals for representing beliefs and intentions. Neither includes explicit representation of capabilities. The systems that we are aware of for reasoning about actions tend to address capability *to perform actions* only indirectly, although several address an agent's capability *to bring about states of affairs.* They assume either that (a) an agent is always capable of performing any action in its repertoire, but an action's consequences depend on whether its preconditions were met, or (b) an agent is capable of performing an action exactly when its preconditions are true. We assume that whether an agent can perform an action at all is separate from whether performing that action will have the intended effects. We will discuss some of this work in section 5.

In this paper we have made some simplifying assumptions. Although one motivation of this work is the problem of communicating agents, we present the single-agent case. We believe that expanding to the multiple-agent case will prove to be fairly straightforward and unenlightening. We use discrete time, and assume that at each time the agent executes exactly one action, which could be the null action. Finally, as in [16], we look only at primitive actions which take one tick to execute. however, unlike [16], this

logic was designed with the expectation that it would be extended to handle higher-level actions built from lower-level ones (see [12] for a description of how this might be done). These higher-level actions obviously may take longer to execute. We will say more in section 5 about this extension.

The rest of this paper is organized as follows: in section 2 we discuss the basic temporal logic we will use. Section 3 presents our model of actions, and section 4 presents operators for beliefs, capabilities, and plans. In section 5 we discuss related work and the effects of relaxing some of the assumptions mentioned above. Section 6 is a concluding summary.

2 Base Language

Our base language is a propositional temporal language.

Syntax We have a set of time symbols T and a set of propositional symbols P.

- If $\mathtt{t}_1$, $\mathtt{t}_2 \in \mathtt{T}$ then $\mathtt{t}_1 < \mathtt{t}_2$ is a wff.
- If $\mathtt{p} \in \mathtt{P}$ and $\mathtt{t} \in \mathtt{T}$ then $\mathtt{p}^{\mathtt{t}}$ is a wff.
- If $\mathtt{t}_1$,$\mathtt{t}_2 \in \mathtt{T}$ and $\Phi^{\mathtt{t}}1$ is a wff then $\Box^{\mathtt{t}}2\Phi^{\mathtt{t}}1$ is a wff.
- If $\Phi^{\mathtt{t}}$ and $\Psi^{\mathtt{t}}$ are wffs then so are $\neg\Phi^{\mathtt{t}}$ and $\Phi^{\mathtt{t}} \wedge \Psi^{\mathtt{t}}$.

By $\Box^{\mathtt{t}}1\Phi^{\mathtt{t}}2$ we mean that it is the case at time $\mathtt{t}_1$ that Φ will be (or is, or was) true at time $\mathtt{t}_2$ no matter what else may happen between $\mathtt{t}_1$ and $\mathtt{t}_2$. Notice that $\Phi^{\mathtt{t}}1 \wedge \Psi^{\mathtt{t}}2$ is a legal sentence only if $\mathtt{t}_1$=$\mathtt{t}_2$. We define $\vee$, $\rightarrow$, and $\Diamond$ as usual.

Semantics We have a set W of worlds, each with a unique associated time. Our basic semantic structure is a *capability tree*, which is a collection of worlds and a precedence relation over them. Intuitively, a tree is well-formed if only forward branching is allowed and every branch extends infinitely far into the past and future. We do not require that the structure be connected, so a more proper name would be a "capability forest," but for most of this paper we will assume a tree.

Formally, a capability tree is a tuple $\langle W, T, <, timeOf, pred, \pi \rangle$ where

- W is an infinite set of worlds,
- T is an infinite set of times[1],
- $<$ is a total order on times,
- $timeOf$ is a function from worlds to times,
- $pred$ is a function from worlds to worlds (from each world to its unique predecessor in the tree), where
  - $timeOf(pred(w)) + 1 = timeOf(w)$[1]
  - $\forall w \exists w_1 . w_1 = pred(w) \wedge \forall w \exists w_2 . w = pred(w_2)$
  - $pred^*(w) \stackrel{\text{def}}{=} \{w' \mid w' = w \text{ or } w' \in pred^*(pred(w))\}$
- π is a truth function which maps each $\langle$p,t$\rangle$ pair (where p is a primitive proposition and t is a time) into the set of worlds with associated time t where p is true.

We say that one world is reachable from another if there is some temporally monotonic path between them in the tree; that is,

$reachable(w_1, w_2) \equiv w_1 \in pred^*(w_2)$ or $w_2 \in pred^*(w_1)$.

Finally we define a function rat (for *reachable-at-time*):

$rat(w,t) = \{w' \mid reachable(w', w) \text{ and } timeOf(w') = t\}$.

We now define what it means for a pair M, w to model a sentence (where M is a capability tree and $w \in W$):

- $M, w \models \mathtt{t}_1 < \mathtt{t}_2$ iff $\mathtt{t}_1$,$\mathtt{t}_2 \in \mathtt{T}$ and $t_1 < t_2$. [2]
- $M, w \models \mathtt{p}^{\mathtt{t}}$ iff $w \in \pi(\mathtt{p},\mathtt{t})$.
- $M, w \models \neg\Phi^{\mathtt{t}}$ iff $timeOf(w) = t$ and $M, w \not\models \Phi^{\mathtt{t}}$.
- $M, w \models \Phi^{\mathtt{t}} \wedge \Psi^{\mathtt{t}}$ iff $M, w \models \Phi^{\mathtt{t}}$ and $M, w \models \Psi^{\mathtt{t}}$.
- $M, w \models \Box^{\mathtt{t}}1\Phi^{\mathtt{t}}2$ iff $timeOf(w) = t_1$ and for every $w' \in rat(w, t_2)$, it is the case that $M, w' \models \Phi^{\mathtt{t}}2$.

So our model of the agent's environment is a tree (or forest) with many possible futures. If the agent is using this model to reason at time t, then it may be that all the branching in the tree takes place after t, since the present and the past are completely determined. Note that this is not to say that the agent has complete knowledge of the present or past, as we will see in section 4.1.

3 Actions

An action is a function from a world to a set of worlds. For example, let us consider an action a. Suppose that applying a to world w gives us a set S of worlds. If our agent performs action a in world w (and $timeOf(w) = t$) then at time $t+1$ the agent will be in one of the worlds in S. Which of those worlds is the actual resulting world may depend on such external factors as chance or the actions taken by other agents.

As mentioned in section 1, we will consider only primitive actions, each taking exactly one unit of time to execute. We assume our agent performs exactly one action at each time, and that one available action is the null action. We discuss relaxing these assumptions in section 5.

[1] For this paper, we will assume T is the integers, but this choice is not forced on us.

[2] We will adopt the convention of using `typewriter font` for syntactic objects and *math font* for semantic objects. When a syntactic object and a semantic object have the same name, it is to be understood that our interpretation maps the one to the other. Throughout this paper, we will leave the interpretation implicit.

Syntax We have a set A of primitive actions. If $\mathtt{a} \in \mathtt{A}$ and $\mathtt{t} \in \mathtt{T}$ then $\mathtt{did(t,a)}$ is a wff. (For consistency of notation, we will write this wff as $\mathtt{did^t(a)}$.) $\mathtt{did^t(a)}$ is true iff the agent has just finished performing action a at time t.

Semantics We add to our model a set A of functions[3]; if $a \in A$ then $a : W \mapsto 2^W$. If $A = \{a_1, a_2, \ldots, a_n\}$ and $w \in W$ then

- $a_i(w) \subseteq \{w' \mid w = pred(w')\}$, $1 \leq i \leq n$,
- $a_1(w) \cup a_2(w) \cup \cdots \cup a_n(w) = \{w' \mid w = pred(w')\}$, and
- $a_i(w) \cap a_j(w) = \emptyset$ if $i \neq j$.

This last requirement, that the results of different actions be completely disjoint, may seem overly constraining at first. However, worlds resulting from different actions will always be distinct if only by virtue of the did predicate, whose meaning we now define:

- $M, w \models \mathtt{did^t(a)}$ iff $timeOf(w) = t$ and $w \in a(pred(w))$.

From these definitions, we have the following:

Proposition 1 *The following sentences are valid:*

- $\mathtt{did^t(a_1) \vee did^t(a_2) \vee \cdots \vee did^t(a_n)}$.
- $\mathtt{\neg(did^t(a_i) \wedge did^t(a_j))}$, $\mathtt{i \neq j}$.

Notice that, due to the branching time structure we are using, we can say what has happened in the past, but the only way we can talk about the future is by saying what is true in every possible future (or in some possible future). That is, there is a fact of the matter about the past, but the future is yet to be determined. Intuitively, this is why we included $\mathtt{did^t(a)}$ in our language rather than for example $\mathtt{does^t(a)}$. There is a fact of the matter about what the agent just did. Whether the agent *will* perform some action is a question that we cannot answer without peering into the future, and whether an agent *plans to* perform some future action is a question about the agent's mental state, not about the world.

4 The Agent's Mental State

Now that we have a way of talking about the agent's environment, we turn to describing the agent. We represent the information the agent possesses about itself and its environment as a set of *beliefs*. The actions it has available to it and the states of affairs it can bring about by performing these actions are described in terms of its *capabilities*. Finally, the actions that the agent means to perform and the states of affairs it means to bring about are described in terms of its *plans*.

Remember that we are looking only at a snapshot of the agent's internal state and will not address the question of how its beliefs, capabilities, and plans evolve over time.

[3]See footnote 2.

4.1 Beliefs

We add a modal operator representing belief. We will leave open the precise choice of modal system, as this decision may depend on characteristics of the domain or the agents. For example, for an agent whose reasoning consists of theorem-proving, and whose beliefs are just the axioms contained in its knowledge base, the axioms of positive and negative introspection would probably be appropriate. We will insist, however, that the agent cannot believe contradictions.

We use a standard possible-worlds model. We have a collection of accessibility relations $\mathcal{B}^t$, one for each $t \in T$. These accessibility relations are just sets of ordered pairs of worlds. If we have a forest of unconnected capability trees, accessible worlds are not required to be in the same tree. (Belief is the only operator for which other trees in the forest might be relevant.)

Syntax If $\Phi^{\mathtt{t}_1}$ is a wff and $\mathtt{t}_2 \in \mathtt{T}$ then $\mathtt{B^{t_2}}\Phi^{\mathtt{t}_1}$ is a wff.

Semantics For each $t \in T$, we add to our model $\mathcal{B}^t \subseteq 2^{W^t \times W}$, where $W^t = \{w \mid w \in W \text{ and } timeOf(w) = t\}$. For every world $w \in W^{t_1}$ and every $t_2 \in T$, there exists some $w' \in W^{t_2}$ such that $(w, w') \in \mathcal{B}^{t_1}$.

- $M, w \models \mathtt{B^{t_1}}\Phi^{\mathtt{t}_2}$ iff $w \in W^{t_1}$ and for all $w' \in W^{t_2}$ such that $(w, w') \in \mathcal{B}^{t_1}$, $M, w' \models \Phi^{\mathtt{t}_2}$.

From these definitions, we have the following:

Proposition 2 *The following sentences are valid:*

K $\mathtt{B^{t_1}}\Phi^{\mathtt{t}_2} \wedge \mathtt{B^{t_1}}(\Phi^{\mathtt{t}_2} \rightarrow \Psi^{\mathtt{t}_2}) \rightarrow \mathtt{B^{t_1}}\Psi^{\mathtt{t}_2}$.

D $\mathtt{\neg B^{t_1} false^{t_2}}$.

That is, the agent is logically omniscient (a well-known problem with possible-worlds models of belief) and never believes contradictions. As mentioned above, we can choose other semantic restrictions to suit a particular application. For more about modal systems for belief, see [7].

In addition, we will place a few more constraints on $\mathcal{B}^t$:

- If $(w_1, w_2) \in \mathcal{B}^t$ and $timeOf(w_1) = timeOf(w_2)$, then $w_1 \in a(pred(w_1))$ iff $w_2 \in a(pred(w_2))$ for every $a \in A$.
- If $(w_1, w_2) \in \mathcal{B}^{t_1}$, $(w_2, w_3) \in \mathcal{B}^{t_2}$, and $t_1 \leq t_2$, then $(w_1, w_3) \in \mathcal{B}^{t_1}$.
- If $(w_1, w_2) \in \mathcal{B}^{t_1}$ then $(w_1, pred(w_2)) \in \mathcal{B}^{t_1}$.
- If $(w_1, w_2) \in \mathcal{B}^{t_1}$ then $(w_1, w_3) \in \mathcal{B}^{t_1}$ for some $w_3 \in \{w | w_2 = pred(w)\}$.

From these constraints, we have

Proposition 3 *The following sentences are valid:*

- $\mathtt{did^t(a) \equiv B^t did^t(a)}$.
- $\mathtt{(t_1 \leq t_2) \rightarrow (B^{t_1}}\Phi^{\mathtt{t'} \rightarrow \mathtt{B^{t_1}B^{t_2}}}\Phi^{\mathtt{t'}})$.
- $\mathtt{B^{t_1}}\Box^{\mathtt{t_2}}\Phi^{\mathtt{t_3}} \rightarrow \mathtt{B^{t_1}}\Phi^{\mathtt{t_3}}$.

- $B^{t1}\Phi^{t2} \rightarrow B^{t1}\Diamond^{t3}\Phi^{t2}$.

The first says that the agent always knows what action it just performed, and the second that the agent believes that its current beliefs will persist. The third and fourth specify the relationship between the belief operator and the necessity and possibility operators.

Since we are only discussing a snapshot of the agent's mental state, we will not discuss how the agent's beliefs actually evolve as time passes. This is an important and difficult problem, as the literature on belief revision attests. As seen above, however, we can examine the agent's *current* beliefs about its future beliefs, and indeed about all aspects of its future mental state. Others who work on the Agent Oriented Programming project have considered the evolution of beliefs under a constraint of minimal learning; see [9].

4.2 Capabilities

We will examine two types of capability: CanDo, which applies to actions, and CanAch (for "can achieve"), which applies to sentences. It is important to keep these two types of capability separate, and a planning agent will want to be able to reason about both. An agent will consider whether it CanAch a proposed goal Φ in deciding whether to adopt Φ as a goal. If it provably cannot achieve Φ, then it should not adopt the goal. If Φ is being considered as a means of achieving some higher-level goal Ψ then the agent should seek other means of achieving Ψ. Reasoning about whether it CanDo a given action, on the other hand, usually occurs when the agent is seeking means to achieve goals which have already been adopted.

By $\text{CanDo}^{t}(a)$, we mean that the agent can perform action a at time t. When we say $\text{CanAch}^{t}\Phi^{t'}$ we mean that the agent is capable (at time t) of some series of actions which will guarantee that Φ will be true at time t′, no matter what else happens between t and t′.

Syntax If $a \in A$, $t_1 \in T$, and Φ^{t2} is a wff then $\text{CanDo}^{t1}(a)$ and $\text{CanAch}^{t1}\Phi^{t2}$ are wffs.

Semantics We define CanDo and CanAch as follows:

- $M, w \models \text{CanDo}^{t}(a)$ iff $timeOf(w) = t$ and there exists w' such that $w' \in a(w)$.
- $M, w \models \text{CanAch}^{t1}\Phi^{t2}$ iff
  - $t_1 = t_2$ and $M, w \models \Phi^{t2}$, or
  - $t_1 < t_2$ and there exists $a \in A$ such that $M, w \models \text{CanAch}^{t1}(\text{CanDo}^{t2-1}(a) \wedge \Box^{t2-1}(\text{did}^{t2}(a) \rightarrow \Phi^{t2}))$[4], or
  - there exists t′ such that $t_1 \leq t' < t_2$ and $M, w \models \text{CanAch}^{t1}\ \Box^{t'}\Phi^{t2}$.

So an agent is capable of an action if there is at least one world which might result from the performance of that action. The agent is capable of achieving some state of affairs Φ^{t} if either it is now time t and Φ is true, or there is a series of actions, not necessarily contiguous, which would if performed guarantee Φ^{t} and furthermore the agent is capable of performing all of these actions in the appropriate situations. This definition is quite strong, as it requires that the agent be able to bring about Φ^{t} no matter what other events might happen in the meantime.

We will just point out that an obvious extension to the language we have presented thus far is to allow quantification over actions. In that case, we might change the second disjunct of the definition of CanAch^{t} to be:

- $t_1 < t_2$ and $M, w \models \text{CanAch}^{t1}$ $(\exists\, a.\text{CanDo}^{t2-1}(a) \wedge \Box^{t2-1}(\text{did}^{t2}(a) \rightarrow \Phi^{t2}))$,

which would say the agent is capable of achieving Φ if there is some (possibly conditional) plan for achieving Φ; that is, which action the agent takes may depend on which world the agent ends up in. However, for this paper, we will allow only linear plans.

From the above definition, we have

Proposition 4 *The following sentences are valid:*

- $\text{CanDo}^{t}(a) \equiv \Diamond^{t}\text{did}^{t+1}(a)$.
- $\text{did}^{t}(a) \rightarrow \Box^{t}\text{CanDo}^{t-1}(a)$.
- $\Box^{t1}\Phi^{t2} \rightarrow \text{CanAch}^{t1}\Phi^{t2}$.
- $\text{CanAch}^{t1}\Phi^{t2} \rightarrow \Diamond^{t1}\Phi^{t2}$.
- $\neg\text{CanAch}^{t}\ \text{false}^{t'}$.
- $\text{CanDo}^{t}(a) \equiv \text{CanAch}^{t}\text{did}^{t+1}(a)$.
- $t_1 < t_2 \rightarrow \neg\text{CanAch}^{t2}\Phi^{t1}$.

Notice that we get axiom **D** for CanAch, but not axiom **K**. In particular, $\text{CanAch}^{t1}(\Phi^{t2} \wedge \neg\Phi^{t2})$ is not satisfiable, but $\text{CanAch}^{t1}\Phi^{t2} \wedge \text{CanAch}^{t1}\neg\Phi^{t2}$ is.

4.3 Plans

We introduce three operators, PD, PND, and PA for representing the agent's plans. The abbreviations stand for "plan to do," "plan not to do," and "plan to achieve," respectively. We are not using the word "plan" in the typical AI sense of the word. Following the usage of [1], by "plans to ..." we mean not just that the agent in question knows about a plan, but that the agent means to carry out (a particular instantiation of) that plan. As mentioned above, in this paper we limit our agent to linear plans; in a later paper, we will discuss conditional plans as well. Since in this paper we are examining only a snapshot of the agent's mental state, we will not here address the question of how the agent's disposition to execute the planned action(s) persists, although this persistence is obviously central to the usefulness of making plans. (For more on the usefulness and uses of plans, see [13]).

[4] Of course we have not defined temporal terms like "t-1;" but we can replace t-1 with t′ and conjoin $(t' < t \wedge (t'' < t \rightarrow t'' \leq t'))$.

Syntax If $a \in A$, $t_1 \in T$ and $\Phi^{t}2$ is a wff then $PD^{t}1(a)$, $PND^{t}1(a)$, and $PA^{t}1\Phi^{t}2$ are wffs.

Semantics We add to our model an accessibility relation $\mathcal{P}$ over worlds. We require that if $(w, w') \in \mathcal{P}$ then $w = pred(w')$, and we further require that $\mathcal{P}$ be serial (so there is always some $\mathcal{P}$-accessible world). In other words, for each world, $\mathcal{P}$ picks out some of its successor worlds.

- $M, w \models PD^{t}(a)$ iff for all worlds w' such that $(w, w') \in \mathcal{P}$, $M, w' \models did^{t+1}(a)$.
- $M, w \models PND^{t}(a)$ iff for all worlds w' such that $(w, w') \in \mathcal{P}$, $M, w' \not\models did^{t+1}(a)$.
- $M, w \models PA^{t}1\Phi^{t}2$ iff
  - $t_1 = t_2$ and $M, w \models B^{t}2\Phi^{t}2$, or
  - $t_1 < t_2$ and there exists $a \in A$ such that $M, w \models PA^{t}1(PD^{t}2^{-1}(a) \wedge \Box^{t}2^{-1}(did^{t}2(a) \rightarrow \Phi^{t}2))$, or
  - there exists t' such that $t_1 \leq t' < t_2$ and $M, w \models PA^{t}1\ \Box^{t'}\Phi^{t}2$.

Notice that the definition of the PA operator looks similar to the definition for CanAch, although it includes an additional belief operator. This difference emphasizes the fact that the agent's capability depends on the world, while the agent's plans are entirely "mental" phenomena. According to the definition of PD, for every world in the agent's model, there is *at most* one chosen action that the agent plans (in the current snapshot of the agent's mental state) to perform, should it ever find itself in that world. For some worlds, the agent may not plan to perform any particular action. On the other hand, there may be many actions which an agent plans *not* to perform. (In fact, if $|A| = n$ then there may be as many as $n - 1$ actions which the agent plans not to perform.)

From the above definitions, we have:

Proposition 5 *The following sentences are valid:*

- $PD^{t}(a) \rightarrow CanDo^{t}(a)$.
- $\neg CanDo^{t}(a) \rightarrow PND^{t}(a)$.
- $PA^{t}1\Phi^{t}2 \rightarrow B^{t}1\,CanAch^{t}1\Phi^{t}2$.
- $\neg PA^{t}\ false^{t'}$.
- $PD^{t}(a_i) \rightarrow PND^{t}(a_j)$, $i \neq j$.
- $\neg(PD^{t}(a) \wedge PND^{t}(a))$.
- $PND^{t}(a_1) \wedge \ldots \wedge PND^{t}(a_{n-1}) \rightarrow PD^{t}(a_n)$ (where $A = \{a_1, a_2, \ldots, a_n\}$).
- $PD^{t}(a) \rightarrow PA^{t}did^{t+1}(a)$.
- $PA^{t}did^{t+1}(a) \rightarrow (PD^{t}(a) \vee B^{t}\Box^{t}did^{t+1}(a))$.

In addition, we will now place further semantic restrictions on the relationship between the belief relation $\mathcal{B}^{t}$ and our new relation $\mathcal{P}$:

- If $(w_1, w_2) \in \mathcal{B}^{t}$ then $(pred(w_2), w_2) \in \mathcal{P}$.
- If $(w_1, w_2) \in \mathcal{B}^{t}$ and $timeOf(w_1) = timeOf(w_2)$ then there exists a w_3 such that $(w_1, w_3) \in \mathcal{P}$ and $w_3 \in a(w_1)$ iff there exists a w_4 such that $(w_2, w_4) \in \mathcal{P}$ and $w_4 \in a(w_2)$, for all actions a.

From these conditions, we get

Proposition 6 *The following sentences are valid:*

- $B^{t}1PD^{t}2(a) \rightarrow B^{t}1\,did^{t}2^{+1}(a)$.
- $B^{t}1PND^{t}2(a) \rightarrow B^{t}1\neg did^{t}2^{+1}(a)$.
- $PD^{t}(a) \equiv B^{t}PD^{t}(a)$.
- $PND^{t}(a) \equiv B^{t}PND^{t}(a)$.
- $PA^{t}1\Phi^{t}2 \rightarrow B^{t}1\Phi^{t}2$.

Again, it is outside the scope of this paper to consider how the agent's plans evolve, and this is a complicated question. While it seems clear that plans should tend to persist, it also seems clear that they should sometimes be reconsidered. The reader is referred to the discussions of this point in [2; 13; 5].

Notice that although we can say when an agent has a plan for achieving some state of affairs, we do not necessarily know what that plan is. Furthermore, plans are not objects in our language, the way actions are. We believe that the plan operators defined above will suffice for describing the agents we are interested in. These agents will be described in a paper to appear on the agent oriented programming language PLACA (which stands for PLAnning Communicating Agents).

5 Discussion and Related Work

In [15], Rao and Georgeff present a logic for beliefs, goals, and intentions. For them, goals are states of the world that the agent desires and has chosen, while intentions are goals that the agent is committed to realizing. Our PlantoDo and PlantoAch operators are meant to capture something slightly different from either. For us, an agent plans to achieve some Φ if it not only has chosen to pursue Φ but also has constructed a plan for achieving Φ *and* means to carry out that plan. Rao and Georgeff use a branching time structure rather like ours, in which actions (or *events*) move the agent from one *situation* (our worlds) to another; in their system as in ours an agent can attempt some action but fail to bring about the desired effects. They examine the evolution of intentions, once formed, while we model only the static mental state of an agent.

While this work is similar in some ways to that of Cohen and Levesque [5], there are some important differences. For example, Cohen and Levesque have as an axiom of their system $BEL(\Phi) \rightarrow GOAL(\Phi)$, while Rao and Georgeff have $GOAL(\Phi) \rightarrow BEL(\Phi)$ (where Φ is an O-formula). Cohen and Levesque place this restriction on their model in order to capture *realism*, that is, in order to enforce the condition that an agent's goals must be consistent with its beliefs. Rao and Georgeff capture realism (in fact, *strong realism*, which requires

that the agent believe it can optionally achieve its goals) in a different way; their axiom shown above says that if the agent has a goal that optional(Φ) be true then it believes optional(Φ). While we have no goal operator, our $\mathtt{PA}^{\mathtt{t}}1\Phi^{\mathtt{t}}2 \rightarrow \mathtt{B}^{\mathtt{t}}1\Phi^{\mathtt{t}}2$ looks similar to Rao and Georgeff's axiom. We are after a different intuition, though; we mean not just that the agent believes that it is possible that it will succeed in achieving its goal, but that it believes that it will in fact succeed.

Cohen and Levesque's model is different from ours, using linear-time worlds instead of branching-time, and using modal temporal operators (next-time, every-subsequent-time, *etc.*) as opposed to our explicit dates. They allow actions to be combined into higher-level actions; we intend to make a similar extension to our language but have not yet done so. Cohen and Levesque are interested in exploring the issues of agents' commitment to their goals and they define intentions in terms of persistent goals. We are providing a representation which combines actions, beliefs, capabilities, and plans in order to facilitate the kinds of reasoning that must be done by planning agents (in particular to help us see what counts as a consistent "mental state"), so our work is at a lower level.

Singh's simple *know-how* operator K' [17] is similar to our $\mathtt{CanAch}^{\mathtt{t}}$ operator, with two main differences: (1) K'p is atemporal, so it doesn't matter when p is achieved, while our operator specifies *when* p is to be made true, and (2) Singh's system has no analog to CanDo; an agent *knows how* to achieve p "if it has the right set of basic actions,"[5] and one cannot reason about whether the agent's capability to perform them depends on the current situation. We believe that artificial agents will usually be given deadlines and so our temporal CanAch operator makes sense. In addition, we find our semantics simpler and more intuitive than Singh's.

In [3], Brown presents a logic for representing ability which uses a possible-worlds semantics in which the accessibility relation is from worlds to *sets* of worlds, which he calls clusters. Each cluster can be thought of as the set of worlds which might result from the performance of some action; if a proposition p is true at every world in some accessible cluster, then the agent in question is capable of bringing about p. This is obviously similar to our $\mathtt{CanAch}^{\mathtt{t}}$ operator, but there are some important differences. Time is not explicitly represented in Brown's system. For Brown, to say that an agent can achieve p means that there is some action which the agent could perform now which would cause p to be true. In his system, one cannot describe the result of a series of actions. Like Singh, Brown does not have an operator which corresponds to our CanDo, but each accessible cluster corresponds to an action which the agent can take. It would be easy to add an explicit CanDo operator to his system. Given that Brown's motivation is so different from ours (he is interested in "claims to ability that are morally relevant"), it is interesting that the two systems have even this much similarity.

McCarthy and Hayes [11] discuss three flavors of capability, two of which are relevant to agent oriented programming systems. In their paper, the system under consideration (in our case, a collection of agents) is represented as a set of interacting finite automata. Each automaton has an internal state and may have some inputs and/or outputs, each of which may be connected to another automaton or to the world external to the system. If automaton p is part of system S, then we say p can achieve some Φ if there is some sequence of outputs from p which will cause system S to enter a state where Φ holds.[6] If we think of each automaton as an agent, and S as a system of agents, then we can see directly the correspondence between their concept of capability and our CanAch. McCarthy and Hayes do not discuss the relation between CanDo and CanAch, although later in the paper they discuss knowledge preconditions for actions or strategies.

We see our definition of "planning to ..." as a first step towards a definition of intention; basically, an agent who has an intention to achieve Φ must come eventually to have a plan for achieving Φ, or else it must drop its intention.[7] Once we add the evolution of agents' mental states to our picture, we can think of adopting a (high-level) intention to achieve Φ as the starting point of a process which ends when the agent comes to have a plan for achieving Φ. The agent may begin with no particular idea of how to bring Φ about, but over time it will form a high-level plan for achieving Φ and then refine that plan until it consists of executable actions. This refinement may happen gradually, and the agent may for example refine first the parts of the plan which will be executed earliest. Whatever the strategy for plan refinement, an agent whose intention to achieve Φ persists must come up with a plan for achieving Φ, and do so in time to execute the plan appropriately.

Our system is fairly easy to extend to a multiple-agent system. Once we add other agents to the system, we automatically lose our assumption that only one action is performed at a time. We must then introduce a result function from worlds and n-tuples of actions to sets of worlds (that is, from the current world and the actions performed by all agents to the set of possible resulting worlds). The value of this function at a given world and tuple of actions can be found by applying each of the individual action functions to the current world and then taking the intersection of the result-

[5] [17], p. 345

[6] The difference between their two flavors of capability, which they call *cana* and *canb*, is a difference in p's allowable transition functions. In the first case, p's output can depend on the history of external inputs received by S, while in the second, p's output can only depend on (the history of) p's input.

[7] Our thinking on the relationship between intentions and plans has been greatly influenced by [1].

ing sets. Notice, however, that if the agents' actions might interfere with each other, then the sets returned by individual action functions will be much larger than in the single-agent case, as they must include all the worlds that might result from the action done alone or done with any other possibly concurrent actions. For a deeper discussion of this point, see [12] and [8]. The extension of belief to a multiple-agent system is widely understood [7], and the extensions of the capability and plan operators are not difficult. In this new system, each agent can have beliefs about the mental states of other agents. An agent, once given a catalogue of communicating actions, can incorporate these actions into its plans in order to communicate with other agents to enlist their aid, exchange information, or respond to requests. (For more about the types of communicative actions which will be incorporated into a working system, see [16].)

In this paper, we assumed that each action takes one clock tick to execute, and that our agent executes only one action at a time. The effects of relaxing the first assumption are not severe, if we know how long it takes to execute a given action. (If the execution time of a given action varies slightly, we can adjust by using the maximum time. If the execution time of an action can vary widely, our system requires more substantial modifications.) The definitions of `CanDo` and `CanAch` will change slightly; we will need a function to give the execution time required for each action. We can then simply modify the definitions of these operators to take into account this execution time. The definitions of `PlantoDo` and `PlantoAch` will require greater, but not severe, modification. If we relax the second assumption, so that a single agent may perform several actions at once, we can model this agent as a group of agents, each performing one action, thus reducing to the multiple-agents case discussed above (perhaps each effector can be treated as an agent). Alternatively, we could create a new "superaction" for every set of primitive actions which an agent might perform simultaneously, thus reducing to the one-action-at-a-time case. We suspect that the first method will be simpler.

6 Conclusions and Future Work

We have presented a logic for representing the mental state of an agent, which consists of its beliefs, its capabilities, and its plans. In future work, we shall add intentions to the language, move to a multiple-agent framework, and extend the plan-to operators to cover partially-refined plans. We are implementing a working demonstration system of simulated agents who plan and act in a simple shared domain.

7 Acknowledgements

Thanks to Yoav Shoham, Nils Nilsson, and Martha Pollack for valuable discussions, and to anonymous reviewers from another conference for helpful comments. This work was supported by the Air Force Office of Scientific Research.

References

[1] M. E. Bratman. *Intention, Plans, and Practical Reason.* Harvard University Press, Cambridge, MA, 1987.

[2] M. E. Bratman, D. J. Israel, and M. E. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4(4):349–355, November 1988.

[3] M. A. Brown. On the logic of ability. *Journal of Philosophical Logic*, 17:1–26, 1988.

[4] T. S.-C. Chou and M. Winslett. Immortal: a model-based belief revision system. In J. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 99–110, 1991.

[5] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(3):213–261, 1989.

[6] P. R. Cohen, J. Morgan, and M. E. Pollack, editors. *Intentions in Communication.* System Development Foundation Benchmark Series. MIT Press, 1990.

[7] J. Y. Halpern and Y. Moses. A guide to the modal logics of knowledge and belief: Preliminary draft. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 480–490, 1985.

[8] F. Lin and Y. Shoham. Concurrent actions in the situation calculus. In *Proceedings of the Fourth International Workshop on Nonmonotonic Reasoning*, 1992.

[9] F. Lin and Y. Shoham. On the persistence of knowledge and ignorance: A preliminary report. Unpublished manuscript, 1992.

[10] J. P. Martins and S. C. Shapiro. A model for belief revision. *Artificial Intelligence*, 35(1):25–79, 1988.

[11] J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. L. Webber and N. J. Nilsson, editors, *Readings in Artificial Intelligence*, pages 431–450. Morgan Kaufman, 1981.

[12] R. N. Pelavin. Planning with simultaneous actions and external events. In *Reasoning About Plans.* Morgan Kaufmann Publishers, 1991.

[13] M. E. Pollack. The uses of plans. *Artificial Intelligence*, 57(1):43–68, 1992.

[14] A. S. Rao and N. Y. Foo. Formal theories of belief revision. In R. J. Brachman, H. J. Levesque, and R. Reiter, editors, *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 369–380, 1989.

[15] A. S. Rao and M. P. Georgeff. Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 473–484, 1991.

[16] Y. Shoham. Agent-oriented programming. Technical Report STAN-CS-90-1325, Stanford University, 1990.

[17] M. P. Singh. A logic of situated know-how. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 343–348, 1991.