Modeling Rational Agents within a BDI-Architecture

February, 1991

Technical Note 14

By:

Anand S. Rao
Australian Artificial Intelligence Institute

Michael P. Georgeff
Australian Artificial Intelligence Institute

This research was partly supported by a *Generic Industrial Research and Development Grant* from the Department of Industry, Technology and Commerce, Australia.
This paper is to appear in the proceedings of the *Second International Conference on Principles of Knowledge Representation and Reasoning, KR91* edited by Allen, J., Fikes, R., and Sandewall, E., published by Morgan Kaufmann, San Mateo, CA, 1991.

Abstract

Intentions, an integral part of the mental state of an agent, play an important role in determining the behavior of rational agents as they seek to attain their goals. In this paper, a formalization of intentions based on a branching-time possible-worlds model is presented. It is shown how the formalism realizes many of the important elements of Bratman's theory of intention. In particular, the notion of intention developed here has equal status with the notions of belief and desire, and cannot be reduced to these concepts. This allows different types of rational agents to be modeled by imposing certain conditions on the persistence of an agent's beliefs, goals, and intentions. Finally, the formalism is compared with Bratman's theory of intention and Cohen and Levesque's formalization of intentions.

1 INTRODUCTION

The role played by attitudes such as beliefs (B), desires (D) (or goals (G)), and intentions (I) in the design of rational agents has been well recognized in the philosophical and AI literature [Bratman, 1987; Bratman *et al.*, 1988; Georgeff and Ingrand, 1989]. Systems and formalisms that give primary importance to intentions are often referred to as BDI-architectures. While most philosophical theories treat intentions as being reducible to beliefs and desires, Bratman argues convincingly that intentions play a significant and distinct role in practical reasoning. He treats intentions as partial plans of action that the agent is committed to execute to fulfill her goals.

Some of the philosophical aspects of Bratman's theory were formalized by Cohen and Levesque [Cohen and Levesque, 1990]. In their formalism, intentions are defined in terms of temporal sequences of an agent's beliefs and goals. In particular, an agent *fanatically committed* to her intentions will maintain her goals until either they are believed to be achieved or believed to be unachievable; an agent with a *relativized commitment* to her intentions is similarly committed to her goals but may also drop them when some specified conditions are believed to hold.

In this paper, we present an alternative possible-worlds formalism for BDI-architectures. There are three crucial elements to the formalism. First, intentions are treated as first-class citizens on a par with beliefs and goals. This allows us to define different strategies of commitment with respect to an agent's intentions and thus to model a wide variety of agents. Second, we distinguish between the *choice* an agent has over the actions she can perform and the *possibilities* of different outcomes of an action. In the former case, the agent can choose among outcomes; in the latter case, the environment makes that determination. Third, we specify an interrelationship between beliefs, goals, and intentions that allows us to avoid many of the problems usually associated with possible-worlds formalisms, such as commitment to unwanted side effects.

In the following sections, we briefly outline the formalism and describe some of its more important features. We then define a number of different commitment strategies and show how these affect agent behavior.

2 INFORMAL SEMANTICS

We choose to model the world using a temporal structure with a branching time future and a single past, called a *time tree*. A particular time point in a particular world is called a *situation*.

Event types transform one time point into another. Primitive events are those events directly performable by the agent and uniquely determine the next time point in a time tree. Non-primitive events map to non-adjacent time points, thus allowing us to model the partial nature of plans. Their potential for decomposition into primitive events can be used to model hierarchical plan development.

The branches in a time tree can be viewed as representing the *choices* available to the agent at each moment of time. For example, if there are two branches emanating from a particular time point, one labeled e_1, say, and the other e_2, then the agent has a choice of executing e_1 and moving to the next time point along the branch of the time tree labeled with e_1, or of executing e_2 and likewise moving along its associated branch.

Of course, the agent may attempt to execute some event, but fail to do so. We thus distinguish between the successful execution of events and their failure and label the branches accordingly. As we shall see later, this distinction is critical in having an agent act on her

Figure 1: Temporal modalities

intentions without requiring her to be successful in her attempts.

We use a formalism similar to Computation Tree Logic, CTL*, [Emerson and Srinivasan, 1989] to describe these structures.[1] A distinction is made between *state formulas* and *path formulas*: the former are evaluated at a specified time point in a time tree and the latter over a specified path in a time tree. We introduce two modal operators, *optional* and *inevitable*, which operate on path formulas. A path formula ψ is said to be *optional* if, at a particular time point in a time tree, ψ is true of at least one path emanating from that point; it is *inevitable* if ψ is true of all paths emanating from that point.[2] The standard temporal operators $\bigcirc$ (next), $\diamond$ (eventually), $\Box$ (always), and U (until), operate over state and path formulas.

These modalities can be combined in various ways to describe the options available to the agent, such as shown in Figure 1. For example, the structure shown in the figure could be used to represent the following statements: it is *optional* that John will *eventually* visit London (denoted by p); it is *optional* that Mary will *always* live in Australia (r); it is *inevitable* that the world will *eventually* come to an end (q); and it is *inevitable* that one plus one will *always* be two (s).

Belief is modeled in the conventional way. That is, in each situation we associate a set of *belief-accessible* worlds; intuitively, those worlds that the agent *believes* to be possible. Unlike most conventional models of belief, however, each belief-accessible world is a time tree. Multiple belief-accessible worlds result from the agent's lack of knowledge about the state of the world. But within each of these worlds, the branching future represents the choice (options) still available to the agent in selecting which actions to perform.

Further insight into the approach is provided by comparing the above possible-worlds model with a conventional decision tree. In this case, each arc emanating from a *chance* node of a decision tree corresponds to a possible world, and each arc emanating from a *decision* node to the choice available within a possible world. A formal comparison of our possible-worlds model with the decision-tree representation is carried out elsewhere [Rao and Georgeff, 1990a].

Similar to belief-accessible worlds, for each situation we also associate a set of *goal-accessible* worlds to represent the goals of the agent. Although, in the general case, desires can be inconsistent with one another, we require that goals be consistent. In other words, goals are chosen desires of the agent that are consistent. Moreover, the agent should believe that the goal is achievable. This prevents the agent from adopting goals that she believes

[1] Elsewhere [Rao and Georgeff, 1990b] we use an explicit notion of time to describe these structures.

[2] In CTL*, E and A are used to denote these operators.

Figure 2: Subworld relationship between beliefs and goals

are unachievable and is one of the distinguishing properties of goals as opposed to desires. Cohen and Levesque [Cohen and Levesque, 1987] call this the property of *realism*.

In this paper, we adopt a notion of *strong realism*. In particular, we require that the agent believe she can optionally achieve her goals, by carefully choosing the events she executes (or, more generally, that get executed by her or any other agent). We enforce this notion of compatibility by requiring that, for each belief-accessible world w at a given moment in time t, there must be a goal-accessible world that is a *sub-world* of w at time t. Figure 2 illustrates this relation between belief- and goal-accessible worlds. The goal-accessible world $g1$ is a sub-world of the belief-accessible world $b1$.

Intentions are similarly represented by sets of *intention-accessible* worlds. These worlds are ones that the agent has *committed* to attempt to realize. Similar to the requirement for belief-goal compatibility, the intention-accessible worlds of the agent must be compatible with her goal-accessible worlds; an agent can only intend some course of action if it is one of her goals. Consequently, corresponding to each goal-accessible world w at time t, there must be an intention-accessible world that is a *sub-world* of w at time t. Intuitively, the agent chooses some course of action in w and commits herself to attempt its execution.

In this framework, different belief-, goal-, and intention-accessible worlds represent different possible scenarios for the agent. Intuitively, the agent believes the actual world to be one of her belief-accessible worlds; if it were to be belief world $b1$, then her goals (with respect to $b1$) would be the corresponding goal-accessible world, $g1$ say, and her intentions the corresponding intention-accessible world, $i1$. As mentioned above, $g1$ and $i1$ represent increasingly selective choices from $b1$ about the desire for and commitment to possible future courses of action.

While for every belief-accessible world there must be a goal-accessible world (and similarly for intentions), the converse need not hold. Thus, even if the agent believes that certain facts are inevitable, she is not forced to adopt them as goals (or as intentions). This means that goals and intentions, while having to be consistent, need not be closed under the beliefs of the agent.

In this way, an agent believing that it is inevitable that pain (p) always accompanies having a tooth filled (f), may yet have the goal (or intention) to have a tooth filled without also having the goal (or intention) to suffer pain. This relationship between belief, goal, and intention-accessible worlds is illustrated by the example shown in Figure 3. Although the agent believes that *inevitably always* ($f \supset p$), she does not adopt this as a goal nor as an intention. Similarly, although the agent adopts the goal (and intention) to achieve f, she does not thereby acquire the goal (or intention) p.

Figure 3: Belief, Goal, and Intention Worlds

The semantics of beliefs, goals, and intentions given above is formalized in Section 3. It thus remains to be shown how these attitudes determine the actions of an agent and how they are formed, maintained, and revised as the agent interacts with her environment. Different types of agent will have different schemes for doing this, which in turn will determine their behavioral characteristics. We consider some of these schemes and their formalization in Section 4.

3 FORMAL THEORY

3.1 SYNTAX

CTL* [Emerson and Srinivasan, 1989] is a propositional branching-time logic used for reasoning about programs. We extend this logic in two ways. First, we describe a first-order variant of the logic. Second, we extend this logic to a possible-worlds framework by introducing modal operators for beliefs, goals, and intentions. While Emerson and Srinivasan [Emerson and Srinivasan, 1989] provide a sound and complete axiomatization for their logic, we do not address the issue of completeness in this paper. Our main aim is to present an expressive semantics for intentions and to investigate certain axioms that relate intentions to beliefs and goals within this structure.

Similar to CTL*, we have two types of formulas in our logic: *state formulas* (which are evaluated at a given time point in a given world) and *path formulas* (which are evaluated along a given path in a given world). A state formula can be defined as follows:

- any first-order formula is a state formula;
- if ϕ_1 and ϕ_2 are state formulas and x is an individual or event variable, then $\neg\phi_1$, $\phi_1 \vee \phi_2$, and $\exists x\ \phi_1(x)$ are state formulas;

- if e is an event type then *succeeds*(e), *fails*(e), *does*(e), *succeeded*(e), *failed*(e), and *done*(e) are state formulas;
- if ϕ is state formula then BEL(ϕ), GOAL(ϕ) and INTEND(ϕ) are state formulas; and
- if ψ is a path formula, then *optional*(ψ) is a state formula.

A path formula can be defined as follows:

- any state formula is also a path formula; and
- if ψ_1 and ψ_2 are path formulas, then $\neg\psi_1$, $\psi_1 \vee \psi_2$, $\psi_1 \mathsf{U} \psi_2$, $\Diamond\psi_1$, $\bigcirc\psi_1$ are path formulas.

Intuitively, the formulas *succeeded*(e) and *failed*(e) represent the immediate past performance, respectively successfully and unsuccessfully, of event e. The formula *done*(e) represents the immediate past occurrence of e, either successfully performed or not. The formulas *succeeds*(e), *fails*(e), and *does*(e) are similarly defined but refer to the immediate future occurrence of events. The operators BEL, GOAL, and INTEND represent, respectively, the beliefs, goals, and intentions of the agent.

3.2 POSSIBLE-WORLDS SEMANTICS

We first provide the semantics of various state and path formulas. This will be followed by the semantics of events and, finally, the possible-worlds semantics of beliefs, goals, and intentions.

Definition 1 : An interpretation M is defined to be a tuple, M = $\langle$W, E, T, $\prec$, U, $\mathcal{B}$, $\mathcal{G}$, $\mathcal{I}$, $\Phi\rangle$. W is a set of worlds, E is a set of primitive event types, T is a set of time points, $\prec$ a binary relation on time points,[3] U is the universe of discourse, and Φ is a mapping of first-order entities to elements in U for any given world and time point. A situation is a world, say w, at a particular time point, say t, and is denoted by w_t. The relations, $\mathcal{B}$, $\mathcal{G}$, and $\mathcal{I}$ map the agent's current situation to her belief, goal, and intention-accessible worlds, respectively. More formally, $\mathcal{B} \subseteq$ W $\times$ T $\times$ W and similarly for $\mathcal{G}$ and $\mathcal{I}$. Sometimes we shall use $\mathcal{R}$ to refer to any one of these relations and shall use $\mathcal{R}_t^w$ to denote the set of worlds $\mathcal{R}$-accessible from world w at time t. Figure 4 shows how the belief relation $\mathcal{B}$ maps the world $w0$ at time t_1 to the worlds $b1$ and $b2$. In other words, $\mathcal{B}_{t_1}^{w0} = \{b1, b2\}$.

Definition 2 : Each world w of W, called a *time tree*, is a tuple $<T_w, \mathcal{A}_w, \mathcal{S}_w, \mathcal{F}_w>$, where $T_w \subseteq$ T is a set of time points in the world w and $\mathcal{A}_w$ is the same as $\prec$, restricted to time points in T_w. A *fullpath* in a world w is an infinite sequence of time points (t_0, t_1,...) such that $\forall i$ $(t_i, t_{i+1}) \in \mathcal{A}_w$. We use the notation ($w_{t_0}$, w_{t_1},...) to make the world of a particular fullpath explicit. The arc functions $\mathcal{S}_w$ and $\mathcal{F}_w$ map adjacent time points to events in E. More formally, $\mathcal{S}_w$: $T_w \times T_w \mapsto$ E and similarly for $\mathcal{F}_w$. We require that if $\mathcal{S}_w(t_i, t_j) = \mathcal{S}_w(t_i, t_k)$, then $t_j = t_k$ and similarly for $\mathcal{F}_w$. Also, the domains of $\mathcal{S}_w$ and $\mathcal{F}_w$ are disjoint. Intuitively, for any two adjacent time points for which the arc function $\mathcal{S}_w$ is defined, its value represents the event that successfully occurred between those time points. Similarly, the value of the arc function $\mathcal{F}_w$ represents the failure of events occurring between adjacent time points.

[3] We require that the binary relation be total, transitive and backward-linear to enforce a single past and branching future.

Definition 3 : A *sub-world* is defined to be a sub-tree of a world with the same truth-assignment of formulas. A world w' is a *sub-world* of the world w, denoted by $w' \sqsubseteq w$, if and only if (a) $T_{w'} \subseteq T_w$; (b) for all $u \in T_{w'}$, $\Phi(q, w', u) = \Phi(q, w, u)$, where q is a predicate symbol; (c) for all $u \in T_{w'}$, $\mathcal{R}_u^w = \mathcal{R}_u^{w'}$; and (d) $\mathcal{A}_{w'}$ is $\mathcal{A}_w$ restricted to time points in $T_{w'}$ and similarly for $\mathcal{S}_{w'}$ and $\mathcal{F}_{w'}$. We say that w' is a *strict sub-world* of w denoted by $w' \sqsubset w$ if and only if $w' \sqsubseteq w$ and $w \not\sqsubseteq w'$.

Now consider an interpretation M, with a variable assignment v.[4] We take v_d^i to be that function that yields d for the variable i and is the same as v everywhere else. The semantics of first-order formulas can be given as follows:

$M, v, w_t \models q(y_1, ..., y_n)$ iff $\langle v(y_1), ..., v(y_n)\rangle \in \Phi[q, w, t]$ where $q(y_1, ..., y_n)$ is a predicate formula.
$M, v, w_t \models \neg\phi$ iff $M, v, w_t \not\models \phi$.
$M, v, w_t \models \phi_1 \vee \phi_2$ iff $M, v, w_t \models \phi_1$ or $M, v, w_t \models \phi_2$.
$M, v, w_t \models \exists i\phi$ iff $M, v_d^i, w_t \models \phi$ for some d in U.
$M, v, (w_{t_0}, w_{t_1}, ...) \models \phi$ iff $M, v, w_{t_0} \models \phi$.
$M, v, (w_{t_0}, w_{t_1}, ...) \models \bigcirc\psi$ iff $M, v, (w_{t_1}, ...) \models \psi$.
$M, v, (w_{t_0}, w_{t_1}, ...) \models \Diamond\psi$ iff $\exists k$, $k{\geq}0$ such that $M, v, (w_{t_k}, ...) \models \psi$.
$M, v, (w_{t_0}, w_{t_1}, ...) \models \psi_1 \mathsf{U} \psi_2$ iff
(a) $\exists k$, $k{\geq}0$ such that $M, v, (w_{t_k}, ...) \models \psi_2$ and for all $0 \leq j < k$, $M, v, (w_{t_j}, ...) \models \psi_1$ or (b) for all $j \geq 0$, $M, v, (w_{t_j}, ...) \models \psi_1$.
$M, v, w_{t_0} \models$ *optional*(ψ) iff there exists a fullpath $(w_{t_0}, w_{t_1}, ...)$ such that $M, v, (w_{t_0}, w_{t_1}, ...) \models \psi$.

The formula *inevitable*(ψ) is defined as $\neg$*optional*$(\neg\psi)$ and $\Box\psi$ is defined as $\neg\Diamond\neg\psi$. The definition of U (until) given above is that of *weak* until, which allows fullpaths in which ψ_1 is true forever. Well-formed formulas that contain no positive occurrences of *inevitable* (or negative occurrences of *optional*) outside the scope of the modal operators BEL, GOAL, or INTEND will be called O-formulas and denoted by α. Conversely, we define I-formulas, denoted by β, to contain no positive occurrences of *optional*.

3.2.1 Semantics of Events

Event types transform one time point into another. The various aspects involved in this transformation are called the *dynamics* of a system [Gardenfors, 1988; Rao and Foo, 1989]. Just as one can define the *truth* or *falsity* of formulas at a time point, we need mechanisms for defining the *success* or *failure* of events in transforming one time point to another.

We use the formula *succeeded*(e) to denote the successful execution of event e by the agent, and *failed*(e) to denote its failure. Note that event e *not occurring* is not the same as the event e *failing*. Failure of event types alter the world irrevocably, possibly forcing the agent to replan or revise her plans. This aspect is crucial in capturing the dynamics of any system. For example, the consequences of a thief successfully robbing a bank is quite different from the thief failing in his attempt to rob the bank, which is again different from the thief not attempting to rob the bank. All three are distinct behaviors and have to be distinguished accordingly.

We say that the agent has *done*(e) if she has either *succeeded* or *failed* in doing the event. The notions *succeeds*, *failed*, and *does* are similarly defined, but require the event to occur on all paths emanating from the time point at which the formula is evaluated.

[4] For the sake of simplicity, we shall assume that the variable assignment of event terms are events denoted by the same letter, i.e., $v(e) = e$ for any event term e.

Figure 4: Worlds as time trees

More formally, we have:

$M, v, w_{t_1} \models succeeded(e)$ iff there exists t_0 such that $\mathcal{S}_w(t_0, t_1) = e$.
$M, v, w_{t_1} \models failed(e)$ iff there exists t_0 such that $\mathcal{F}_w(t_0, t_1) = e$.

The formula $done(e)$ is defined as $succeeded(e) \vee failed(e)$; $succeeds(e)$ is defined as $inevitable\bigcirc(succeeded(e))$; $fails(e)$ is defined as $inevitable\bigcirc(failed(e))$; $does(e)$ is defined as $inevitable\bigcirc(done(e))$;

In this paper, we have considered only single-agent, non-parallel actions. If parallel actions among multiple agents are to be allowed, the functions $\mathcal{S}_w$ and $\mathcal{F}_w$ must be extended to map to a set of event-agent pairs, signifying which events are performed by which agents.

3.2.2 Semantics of Beliefs, Goals, and Intentions

The traditional possible-worlds semantics of beliefs considers each world to be a collection of propositions and models belief by a belief-accessibility relation $\mathcal{B}$ linking these worlds. A formula is said to be believed in a world if and only if it is true in all its belief-accessible worlds [Halpern and Moses, 1985].

Cohen and Levesque [Cohen and Levesque, 1990] treat each possible world as a *time-line* representing a sequence of events, temporally extended infinitely into the past and the future. As discussed in Section 2, we instead consider each possible world to be a *time tree*. Each time tree denotes the optional courses of events choosable by an agent in a particular world. The belief relation maps a possible world at a time point to other possible worlds. We say that an agent has a belief ϕ, denoted $\mathsf{BEL}(\phi)$, at time point t if and only if ϕ is true in all the belief-accessible worlds of the agent at time t.

Figure 4 shows how the belief relation $\mathcal{B}$ maps the world $w0$ at time t_1 to the worlds $b1$ and $b2$. Let us assume that the formulas that are true at t_1 in $b1$ are ϕ_1 and ϕ_2, while the formulas that are true at t_1 in $b2$ are ϕ_1 and $\neg\phi_2$. From this it is easy to conclude that $\mathsf{BEL}(\phi_1)$ and $\neg\mathsf{BEL}(\phi_2)$ are true at t_1 in $w0$. As discussed earlier, ϕ_1 and ϕ_2 could be any state formulas; in particular, ones involving the future options available to the agent.

As the belief relation is time-dependent, the mapping of $\mathcal{B}$ at some other time point, say t_2, may be different from the one at t_1. Thus the agent can change her beliefs about the options available to her.

The semantics of the modal operator GOAL is given in terms of a goal-accessible relation $\mathcal{G}$ which is similar to that of the $\mathcal{B}$ relation. The goal-accessibility relation specifies situations that the agent *desires* to be in. Thus, in the same way that we treat belief, we say that the agent has a goal ϕ at time t if and only if ϕ is true in all the goal-accessible worlds of the agent at time t.