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.

One can view intentions as future paths that the agent *chooses to follow*. The intention-accessibility relation $\mathcal{I}$ will be used to map the agent's current situation to all her intention-accessible worlds. We shall say that the agent intends a formula at a certain time if and only if it is true in all the agent's intention-accessible worlds of that time.

We saw above that the goal-accessible worlds of the agent can be viewed as the sub-worlds of the belief-accessible worlds in which the agent desires to be. Similarly, one can view intention-accessible worlds as sub-worlds of the goal-accessible worlds that the agent chooses to follow (i.e., to act upon). Thus, one moves from a belief-accessible world to a goal-accessible world by *desiring* future paths, and from a goal-accessible world to an intention-accessible world by *committing* to certain desired future paths.

The semantics for beliefs, goals, and intentions can be defined formally as follows:

$M, v, w_t \models \mathsf{BEL}(\phi)$ iff $\forall w' \in \mathcal{B}_t^w\ M, v, w'_t \models \phi$.
$M, v, w_t \models \mathsf{GOAL}(\phi)$ iff $\forall w' \in \mathcal{G}_t^w\ M, v, w'_t \models \phi$.
$M, v, w_t \models \mathsf{INTEND}(\phi)$ iff $\forall w' \in \mathcal{I}_t^w\ M, v, w'_t \models \phi$.

We allow intentions over any well-formed formula, which means that one can have intentions about intentions, intentions about goals, intentions about beliefs, and intentions to do certain actions. Some might consider only the last type of intention to correspond with natural usage. While this is arguable, in our formalism the agent might have any type of intention but will only act on the last type of intention.

As an illustration of these ideas, Figure 3 shows one world $b1$ that is belief-accessible from the current situation, say $w0$ at t_0, two worlds $g1$ and $g2$ that are goal-accessible from $w0$ at t_0, and two worlds $i1$ and $i2$ that are intention-accessible from $w0$ at t_0. It is clear from the figure that $i1 \sqsubset g1 \sqsubset b1$ and $i2 \sqsubset g2$. One of the formulas that is true at t_0 in all the intended worlds is $succeeds(d_1)$. Thus the agent intends $succeeds(d_1)$. The sub-world relationship forces the agent to believe, as well as have the goal that $succeeds(d_1)$. The agent intends to succeed and hence intends to carry out the action d_1, but the agent cannot guarantee the ultimate success of her actions—that will be determined by the environment in which the agent is embedded. Thus, even though the above formula is true, it is not necessary, in the actual world, that the formula $succeeded(d_1)$ be true.

From the figure it is clear that at t_0, one of the goal formulas true in all goal accessible worlds is $inevitable(\Diamond f)$. This also implies that the agent believes that this goal is achievable; in other words, $\mathsf{BEL}(optional(\Diamond f))$. From the beliefs, goals, and intentions of the agent, one can say that the agent believes that, if she succeeds in doing d_1, she will achieve the goal f.

3.3 AXIOMATIZATION AND SEMANTIC CONDITIONS

So far, we have not provided any axioms or semantic conditions to capture the desired interrelationships among an agent's beliefs, goals, and intentions. We examine some of these below; additional constraints are discussed elsewhere [Rao and Georgeff, 1990a].

The axiomatization for beliefs is the standard weak-S5 (or KD45) modal system [Hughes and Cresswell, 1984]. We adopt the D and K axioms for goals and intentions; i.e., goals and intentions have to be closed under implication and have to be consistent.

We also have the inference rule of necessitation [Hughes and Cresswell, 1984] for beliefs, goals, and intentions. In other words, the agent believes all valid formulas, intends all valid formulas, and has them as a goal. Hence, like most possible-worlds formalisms, our logic also suffers from the logical omniscience problem [Vardi, 1986]. This problem can be partly alleviated by adopting the *minimal-model* semantics of Chellas [Chellas, 1980] and giving up the inference rule of necessitation and the K-axiom for beliefs, goals, and intentions. However, in this paper we adopt the more traditional modal-logic semantics.

Belief-Goal Compatibility:

The *Axiom of belief-goal compatibility* states that if the agent adopts an O-formula α as a goal, the agent believes that formula.

(AI1) $\mathsf{GOAL}(\alpha) \supset \mathsf{BEL}(\alpha)$.

The above axiom essentially states that, if the agent has the goal that *optional*(ψ) is true, she also believes it; i.e., there is at least one path in all the belief-accessible worlds in which ψ is true.

Consider, for example, the case where the formula ψ above is $\Diamond p$. The axiom then states that, if in all the goal-accessible worlds of the agent there is at least one path where eventually p becomes true, it must be the case that in all the belief-accessible worlds of the agent there is at least one path where eventually p is true. But note that, because of the branching nature of time, the agent need not believe she will ever reach the time point where p is true.

The notion of strong realism as described in Section 2 is captured by imposing the restriction that, for each and every belief-accessible world, there is a corresponding goal-accessible world such that the goal-world is a sub-world of the belief-world. This leads to the following semantic condition:

(CI1) $\forall w' \in \mathcal{B}_t^w \; \exists w'' \in \mathcal{G}_t^w$ such that $w'' \sqsubseteq w'$.

We shall use $\mathcal{B}_t^w \subseteq_{super} \mathcal{G}_t^w$ as a succinct notation for CI1. Such a relationship is shown in Figure 3, where $g1 \sqsubset b1$.

As both beliefs and goals are consistent, the relations $\mathcal{B}$ and $\mathcal{G}$ have to be *serial* (i.e., for any situation there is at least one belief-accessible world and at least one goal-accessible world). This ensures that, in the above semantic condition, we can find at least one belief-accessible world for which there is a goal-accessible world.

To capture the notion of realism, Cohen and Levesque require, instead, that the goal relation $\mathcal{G}$ be a subset of the belief relation $\mathcal{B}$; i.e., $\mathcal{G} \subseteq \mathcal{B}$. As each possible world in their formalism is a time line, this imposes the condition that the chosen (or goal-accessible) worlds are compatible with the agent's belief-accessible worlds. In other words, $\mathsf{BEL}(\phi) \supset \mathsf{GOAL}(\phi)$ is an axiom in their formalism. This axiom forces the agent to adopt as goals certain inevitable facts about the world. As we shall see later, the different semantic condition used in our approach helps us avoid this problem of overcommitment.

Strong realism and realism are not the only ways of capturing the relationship between beliefs and goals. Elsewhere [Rao and Georgeff, 1991] we provide a different semantic relation between beliefs and goals that is suited to realizing other properties of these attitudes.

Goal-Intention Compatibility:

The *Axiom of goal-intention compatibility* states that, if the agent adopts an O-formula α as an intention, the agent should have adopted that formula as a goal to be achieved.

(AI2) $\mathsf{INTEND}(\alpha) \supset \mathsf{GOAL}(\alpha)$.

From the above axioms we have that, if the agent intends α, she believes in α as well. For example, if the agent intends to do an event e, she has the goal to (optionally) do e and also believes that she will (optionally) do e. Nested intentions lead to some interesting consequences. If the formula $\mathsf{INTEND}(\mathit{inevitable}(\Diamond \mathsf{INTEND}(\mathit{does}(e))))$ is true, then $\mathsf{BEL}(\mathit{optional}(\Diamond \mathsf{INTEND}(\mathit{does}(e))))$ is true and also $\mathsf{BEL}(\mathit{optional}(\Diamond \mathsf{BEL}(\mathit{does}(e))))$ is true.

Analogous to the semantic condition CI1 we have the semantic condition CI2, which imposes the restriction that for each and every goal-accessible world there is a corresponding intention-accessible world such that the intention-world is a sub-world of the goal-world.

(CI2) $\forall w' \in \mathcal{G}_t^w \; \exists w'' \in \mathcal{I}_t^w$ such that $w'' \sqsubseteq w'$.

We shall use $\mathcal{G}_t^w \subseteq_{super} \mathcal{I}_t^w$ as a succinct notation for CI2. Figure 3 illustrates the above semantic condition, where $i1 \sqsubset g1$ and $i2 \sqsubset g2$.

As discussed earlier, for each situation there is at least one goal-accessible world and at least one intention-accessible world.

Intentions leading to Actions:

The *Axiom of intention to action* (AI3) captures volitional commitment [Bratman, 1987] by stating that the agent will act if she has an intention towards a single primitive action e. Note that we have not said that the event e will occur successfully, just that the agent is committed to *trying* it. Whether the agent is successful or not depends on the environment in which she is embedded.

(AI3) $\mathsf{INTEND}(does(e)) \supset does(e)$.

Thus, whenever an agent has an intention to do a particular primitive action, she will do that action. However, the axiom does not prevent the agent from doing actions that are not intended. Nor does it say anything about non-primitive actions or other forms of nested intentions.

Note that, if the agent has a *choice* of actions at the current time point, she would be incapable of acting *intentionally* until she deliberates and chooses one of them. One way of modeling this deliberation is to treat the process of deliberation itself as an action to be chosen by the agent [Russell and Wefald, 1989]. An alternative approach would be to modify Axiom AI3 so that the agent arbitrarily chooses one of her intended actions and does that action.

Beliefs about Intentions:

If an agent has an intention, she believes that she has such an intention. The following axiom and semantic condition capture this notion.

(AI4) $\mathsf{INTEND}(\phi) \supset \mathsf{BEL}(\mathsf{INTEND}(\phi))$.
(CI4) $\forall w' \in \mathcal{B}_t^w$ and $\forall w'' \in \mathcal{I}_t^w$ we have $w'' \in \mathcal{B}_t^{w'}$.

In Figure 3, this requires that $b1$ be $\mathcal{I}$-related to $i1$ and $i2$.

Beliefs about Goals:

If the agent has a goal to achieve ϕ, the agent believes that she has such a goal. This intuition can be captured by the following axiom and its corresponding semantic condition.

(AI5) $\mathsf{GOAL}(\phi) \supset \mathsf{BEL}(\mathsf{GOAL}(\phi))$.
(CI5) $\forall w' \in \mathcal{B}_t^w$ and $\forall w'' \in \mathcal{G}_t^w$ we have $w'' \in \mathcal{B}_t^{w'}$.

In Figure 3, this requires that $b1$ be $\mathcal{G}$-related to $g1$ and $g2$.

Goals about Intentions:

If an agent intends to achieve ϕ, the agent must have the goal to intend ϕ. This requires the following axiom and semantic condition.

(AI6) $\mathsf{INTEND}(\phi) \supset \mathsf{GOAL}(\mathsf{INTEND}(\phi))$.
(CI6) $\forall w' \in \mathcal{G}_t^w$ and $\forall w'' \in \mathcal{I}_t^w$ we have $w'' \in \mathcal{G}_t^{w'}$.

In Figure 3, this requires that $g1$ be $\mathcal{I}$-related to $i1$ and $i2$ and $g2$ be $\mathcal{I}$-related to $i1$ and $i2$.

One can strengthen Axioms AI4–AI6 by replacing each implications by an equivalence. This would result in $\mathsf{INTEND}(\phi) \equiv \mathsf{BEL}(\mathsf{INTEND}(\phi)) \equiv \mathsf{GOAL}(\mathsf{INTEND}(\phi))$ and similarly $\mathsf{GOAL}(\phi) \equiv \mathsf{BEL}(\mathsf{GOAL}(\phi))$, which has the effect of collapsing mixed, nested modalities to their simpler non-nested forms.

Awareness of Primitive Events

The next axiom requires the agent to be aware of all primitive events occurring in the world. Once again, we require only that the agent believe a primitive action has been done, not necessarily whether or not it was done successfully.

(AI7) $\mathit{done}(e) \supset \mathsf{BEL}(\mathit{done}(e))$.

No Infinite Deferral

Finally, we require the agent not to procrastinate with respect to her intentions. In other words, if an agent forms an intention, then some time in the future she will give up that intention. This axiom is similar to the one adopted by Cohen and Levesque [Cohen and Levesque, 1990], which requires that there be no infinite deferral of achievement goals.

(AI8) $\mathsf{INTEND}(\phi) \supset \mathit{inevitable}\Diamond(\neg\mathsf{INTEND}(\phi))$.

The above axiom assumes that the intentions corresponding to maintenance goals are also dropped eventually. This could, if necessary, be avoided by restricting the formula ϕ in Axiom AI8 to be an action formula.

We shall refer to this set of eight axioms, AI1 – AI8, together with the standard axioms for beliefs and goals, as the *basic I-system*.

4 COMMITMENT AS AXIOMS OF CHANGE

So far we have treated intentions as a commitment to the performance of current actions. However, we have not formalized how these intentions guide or determine the agent's future commitment to her actions. In other words, we have not discussed how the agent's current intentions relate to her future intentions.

An alternative, proof-theoretic way of viewing the relationship between current and future intentions is as a process of intention maintenance and revision, or what we could intuitively think of as a commitment strategy. Different types of agent will have different commitment strategies. In what follows, we describe three different commitment strategies: *blind*, *single minded*, and *open minded*.

We define a *blindly* committed agent to be one who maintains her intentions until she *actually* believes that she has achieved them. Formally, the axiom of blind commitment states that, if an agent intends that inevitably ϕ be eventually true, then the agent will inevitably maintain her intentions until she believes ϕ.

(AI9a) $\mathsf{INTEND}(\mathit{inevitable}\Diamond\phi) \supset \mathit{inevitable}(\mathsf{INTEND}(\mathit{inevitable}\Diamond\phi)\ \mathsf{U}\ \mathsf{BEL}(\phi))$.

Depending on whether the formula ϕ is an event formula or not, we can capture commitment to actions (i.e., means) or to conditions that have to be true in the future (i.e., ends). Note also that the axiom is defined only for I-formulas (i.e., for intentions towards actions or conditions that are true of *all* paths in the agent's intention-accessible worlds); we do not say anything about the commitment of agents to *optionally* achieve particular means or ends.

A blind-commitment strategy is clearly very strong: the agent will eventually come to believe she has achieved her intentions or keep them forever. Relaxing this requirement, one can define *single-minded* commitment, in which the agent maintains her intentions as long as she believes that they are still options. More formally, we have the following axiom of single-minded commitment:

(AI9b) $\mathsf{INTEND}(\mathit{inevitable}\Diamond\phi) \supset$
$\mathit{inevitable}(\mathsf{INTEND}(\mathit{inevitable}\Diamond\phi)\ \mathsf{U}\ (\mathsf{BEL}(\phi) \vee \neg\mathsf{BEL}(\mathit{optional}\Diamond\phi)))$.

As long as she believes her intentions to be achievable, a single-minded agent will not drop her intentions and thus is committed to her goals. This requirement can also be relaxed. We define an *open-minded* agent to be one who maintains her intentions as long as these intentions are still her goals. In other words, the axiom of open-minded commitment can be stated as follows:

(AI9c) $\mathsf{INTEND}(\mathit{inevitable}\Diamond\phi) \supset$
$\mathit{inevitable}(\mathsf{INTEND}(\mathit{inevitable}\Diamond\phi)\ \mathsf{U}\ (\mathsf{BEL}(\phi) \vee \neg\mathsf{GOAL}(\mathit{optional}\Diamond\phi)))$.

We are now in a position to analyze the properties of different types of agent who adopt the basic I-system, together with one of the above axioms of commitment. Such an agent will be called a *basic agent.*

A basic agent blindly committed to her means (or ends) will inevitably eventually *believe* that she has achieved her means (or ends). This is because Axiom AI9a only allows future paths in which either the object of the intention is eventually believed or the intention is maintained forever. However, by Axiom AI8, the latter paths are not allowed, leading the agent to eventually believe that she has accomplished her intentions.

A basic single-minded agent reaches an identical conclusion only if she continues to believe, until the time she believes she has realized her intentions, that the intended means (or ends) remains an option. Similarly, a basic open-minded agent will eventually believe she has achieved her intentions provided she maintains these intentions as goals until they are believed to have been achieved.

More formally, we have the following theorem for basic agents.

Theorem 1 *:*

(a) A basic, blindly committed agent, with the basic I-system and Axiom AI9a, satisfies the following property:

$\mathsf{INTEND}(\mathit{inevitable}(\Diamond\phi)) \supset \mathit{inevitable}(\Diamond\mathsf{BEL}(\phi))$.

(b) A basic single-minded agent, with the basic I-system and Axiom AI9b, satisfies the following property:

$\mathsf{INTEND}(\mathit{inevitable}(\Diamond\phi)) \wedge \mathit{inevitable}(\mathsf{BEL}(\mathit{optional}(\Diamond\phi))\ \mathsf{U}\ \mathsf{BEL}(\phi)) \supset$
$\mathit{inevitable}(\Diamond\mathsf{BEL}(\phi))$.

(c) A basic open-minded agent, with the basic I-system and Axiom AI9c, satisfies the following property:

$\mathsf{INTEND}(\mathit{inevitable}(\Diamond\phi)) \wedge \mathit{inevitable}(\mathsf{GOAL}(\mathit{optional}(\Diamond\phi))\ \mathsf{U}\ \mathsf{BEL}(\phi)) \supset$
$\mathit{inevitable}(\Diamond\mathsf{BEL}(\phi))$.

Proof:

(a) Assume the premise INTEND (*inevitable* ($\Diamond\phi$)). By Axiom AI9a we can conclude *inevitable* (INTEND (*inevitable* $\Diamond\phi$) U BEL (ϕ)). By Axiom AI8 and the definition of weak until we can conclude *inevitable* ($\Diamond$ BEL (ϕ)). Cases (b) and (c) follow a similar line of reasoning. ♣

Consider now a *competent agent* [Cohen and Levesque, 1990] who satisfies the *Axiom of True Beliefs*, namely $\mathsf{BEL}(\phi) \supset \phi$ (AI10). Under each of the different commitment strategies AI9a, AI9b, and AI9c, the competent agent *will* actually achieve her means (or ends), rather than just believe so. However, AI10 is often difficult for real agents to live up to, as it requires an agent to have true beliefs about the future realization of her intentions. By restricting Axiom AI10 to current beliefs only or to beliefs about primitive action formulas, we can define a less omniscient class of agents who will also inevitably eventually achieve their intentions.

Theorem 2 *: Under the same conditions as Theorem 1, competent agents yield the conclusion* inevitable*($\Diamond\phi$) for all three types of commitment.*

Proof: Follows from the proofs of Theorem 1 followed by the use of Axiom AI10. ♣

The above theorems, however, are not as useful as one would like. First, they do not make any use of Axiom AI3. This means that the same result is achieved independent of whether or not the agent acts intentionally. Moreover, the second conjunct of the premises of (b) and (c) are conditions that have to be true in the real world and which are impossible for a situated agent to enforce; i.e., the agent cannot control these conditions. As a result, the above theorems, although interesting, do not provide a sufficient basis for a situated agent to reason about her intentions and actions.

Consider now an agent who always performs only intentional actions. This can be enforced by requiring the agent to intend a single primitive action at each and every time point. It is reasonable to expect that, in a world free of surprises, such an agent would maintain her beliefs after doing each intended action; i.e., she would not forget previously held beliefs.

More formally, we can state that an agent *preserves a belief* γ *over an intentional action* x if and only if (a) she intends to do x and (b) if she believes γ will hold after doing x, then after doing x, she does indeed believe γ:

$\mathsf{INTEND}(\mathit{does}(x)) \wedge (\mathsf{BEL}(\mathit{optional}\bigcirc(\mathit{done}(x) \wedge \gamma)) \supset \mathit{optional}\bigcirc\mathsf{BEL}(\mathit{done}(x) \wedge \gamma)$.

A single-minded agent who intends inevitably that ϕ is true in the future will inevitably come to believe ϕ provided that she carry out only intentional actions and that she preserve her beliefs about ϕ over these actions. If she were also competent, she would actually come to achieve ϕ.

Theorem 3 *:*

(a) A basic single-minded agent, with the basic I-system and Axiom AI9b, satisfies the following property:

$\mathsf{INTEND}(\mathit{inevitable}(\Diamond\phi)) \wedge$
$\quad \mathit{inevitable}\Box(\exists x(\mathsf{INTEND}(\mathit{does}(x)) \wedge$
$\quad\quad (\mathsf{BEL}(\mathit{optional}\bigcirc(\mathit{done}(x) \wedge (\Diamond\phi))) \supset \mathit{optional}\bigcirc\mathsf{BEL}(\mathit{done}(x) \wedge (\Diamond\phi)))$
$\supset \mathit{inevitable}(\Diamond\mathsf{BEL}(\phi))$.

(b) A competent single-minded agent, with the basic I-system, Axiom AI9b, Axiom AI10, and Axiom AI11 satisfies the following property:

$\mathsf{INTEND}(\mathit{inevitable}(\Diamond\phi)) \wedge$
$\quad \mathit{inevitable}\Box(\exists x(\mathsf{INTEND}(\mathit{does}(x)) \wedge$
$\quad\quad (\mathsf{BEL}(\mathit{optional}\bigcirc(\mathit{done}(x) \wedge (\Diamond\phi))) \supset \mathit{optional}\bigcirc\mathsf{BEL}(\mathit{done}(x) \wedge (\Diamond\phi)))$
$\supset \mathit{inevitable}(\Diamond(\phi)$.

where the event variable x maps to a primitive event type.

Proof: (a) Assume the premise (i) INTEND ($\mathit{inevitable}$ ($\Diamond\phi$)) and (ii) $\mathit{inevitable}$ $\Box$ ($\exists x$ (INTEND ($\mathit{does}(x)$) $\wedge$ (BEL ($\mathit{optional}\bigcirc$ ($\mathit{done}(x) \wedge (\Diamond\phi)$)) $\supset$ $\mathit{optional}\bigcirc$ BEL ($\mathit{done}(x) \wedge (\Diamond\phi)$))).

From (i) and Axioms AI2 and AI3 we have BEL ($\mathit{optional}(\Diamond\phi)$). From this conclusion and (ii), we have the conclusion $\mathit{inevitable}(\mathsf{BEL}(\mathit{optional}(\Diamond\phi))$ U $\mathsf{BEL}(\phi))$. Now we can use Theorem 1 to draw the desired conclusion.

Case (b) is identical to the above proof followed by the application of Axiom AI10. ♣

The second conjunct of the premises of both (a) and (b) can be weakened in several ways. First, Axiom AI3 allows us to drop the $\mathit{done}(x)$ formula in the real world. Second, the agent needs to act according to her intentions only until the moment that she achieves her intentions. In other words, $\mathit{inevitable}\Box$ can be replaced by the until operator.

Given that the agent will also believe the above theorem (by the inference rule of necessitation for beliefs) she will believe that, if she does only intentional actions and preserves her beliefs while doing so, she would ultimately achieve her goals. However, at the same time she can also reason that, if she is forced to do unintentional actions or does not maintain her beliefs, she may not be able to achieve her goals. Therefore, the "Little Nell" problem [McDermott, 1982], in which an agent drops an intention precisely because he believes that its fulfillment will achieve his goal, does not arise.

Similar to the property of preservation of beliefs over intentional actions, one can introduce an analogous property of preservation of goals. This would allow open-minded agents to similarly achieve the object of their intentions.

We can also define other types of agent with mixed commitment strategies. For example, a particularly interesting commitment strategy is one in which the agent is open-minded with respect to ends but single-minded with respect to the means towards those ends. Such an agent is free to change the ends to which she aspires but, once committed to a means for realizing those ends, will not reconsider those means.

A *fanatically committed* agent [Cohen and Levesque, 1990] corresponds to a competent single-minded agent. Similarly, an agent with a *relativized commitment* is competent and open-minded with respect both to means and ends.

We are not suggesting that the above categorization is exhaustive or sufficient for describing realistic rational agents. Our aim in providing the above categorization has simply been to show that the formalism presented here provides a good basis for defining different types of agents and investigating their behavioral properties. It also lays the foundation for a more detailed analysis of reconsideration of intentions [Bratman, 1987].

5 PROPERTIES OF THE LOGIC

There are two important aspects of belief-goal-intention interaction that have received attention in the literature [Bratman, 1987; Cohen and Levesque, 1987]. First, if an agent believes that a formula ϕ is *inevitably always* true (i.e., ϕ is true at all time points in all the future paths of all belief-accessible worlds), then the agent should *not* be forced to adopt this proposition as a goal nor intend it. For example, the belief that, in every possible future state, the earth is round, should not entail that this also be a goal or intention of the agent. The same requirement holds for a slightly weaker form of belief; namely, the belief that a formula ϕ is *inevitably eventually* true (such as the belief that the sun will rise). Moreover, this requirement should hold no matter how persistent are the agent's beliefs. In particular, it should hold even if the agent *inevitably always* believes that a formula ϕ is *inevitably always* true.

Second, if an agent believes that a formula $\phi \supset \gamma$ is *inevitably always* true, and the agent intends ϕ (or has the goal ϕ), then the agent should *not* be forced to intend γ (or have the goal γ). In other words, an agent who intends to do a certain action should not be forced to intend all the *side-effects* of such an action. For example, an agent who intends to go to the dentist to have her tooth removed, but believes *inevitably* that going to the dentist will *always* cause her pain as a side-effect, should not be forced to intend herself pain [Cohen and Levesque, 1987]. As before, the above requirement also applies to the weaker form of belief and to persistent beliefs.

The above requirements are met by our formalism. 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). In this way, an agent believing that it is inevitable that pain always accompanies having a tooth filled, 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 shown in Figure 3.

Let us define a binary relation $<_{strong}$ on the modal operators such that $\mathsf{BEL} <_{strong} \mathsf{GOAL} <_{strong} \mathsf{INTEND}$. A modal formula $R_2(\phi)$ is said to be stronger than $R_1(\phi)$ if and only if $R_1 <_{strong} R_2$. We then have the following two propositions:

Proposition 1 *: A modal formula does not imply a stronger modal formula. For example, if the agent believes (or inevitably always believes) that ϕ is true, she need not adopt ϕ as a goal. In other words, the following formulas are satisfiable:*

(a) $\mathsf{BEL}(\phi) \wedge \neg\mathsf{GOAL}(\phi)$;

(b) $\mathit{inevitable}(\Box\mathsf{BEL}(\phi)) \wedge \neg\mathsf{GOAL}(\phi)$.

General case*: In general, the above results hold if* BEL *is substituted by* R_1 *and* GOAL *by* R_2*, where* $R_1 <_{strong} R_2$.

Proof: If ϕ is of the form *optional*(ψ), then Case (a) is trivially satisfied. Each goal-accessible world is a sub-world of its corresponding belief-accessible world and therefore need not have any path in which ψ is true. For other cases, where ϕ is a first-order formula or a formula of the form *inevitable*(ψ), we make use of the fact that there can be goal-accessible worlds that are not belief-accessible. Thus, if in one such world the formula ϕ is not true, the agent will not have ϕ as a goal. This shows the satisfiability of Case (a). The satisfiability of Case (b) follows a similar pattern. ♣

Proposition 2 *: A modal operator is not closed under implication with respect to a weaker modality. For example, the following formulas are satisfiable:*

(a) $\mathsf{GOAL}(\phi) \wedge \mathsf{BEL}(\mathit{inevitable}(\Box(\phi \supset \gamma))) \wedge \neg\mathsf{GOAL}(\gamma)$;

(b) $\mathsf{GOAL}(\phi) \wedge \mathit{inevitable}(\Box\mathsf{BEL}\ (\mathit{inevitable}(\Box(\phi \supset \gamma))) \wedge \neg\mathsf{GOAL}(\gamma)$.

General case*: In general, the above results hold if* BEL *is substituted by* R_1 *and* GOAL *by* R_2*, where* $R_1 <_{strong} R_2$.

Proof: For every belief-accessible world there has to be a goal-accessible world. But the goal relation $\mathcal{G}$ can map to worlds that do not correspond to any belief-accessible world. Thus, if in one such world the formula $\phi \supset \gamma$ is not true, the agent will not have γ as a goal. This shows the satisfiability of Case (a).

The satisfiability of Case (b) follows a similar pattern. ♣

Both the above propositions deal with the stronger form of beliefs; namely, that it is inevitable that always the agent believes in ϕ. They can be suitably modified for the weaker form as well.

Note that although we have the above propositions, the agent's goals and intentions are closed under implication. In other words, the following formulas are valid formulas in our system:

$\mathsf{INTEND}(\phi) \wedge \mathsf{INTEND}(\phi \supset \gamma) \supset \mathsf{INTEND}(\gamma)$.
$\mathsf{GOAL}(\phi) \wedge \mathsf{GOAL}(\phi \supset \gamma) \supset \mathsf{GOAL}(\gamma)$.

Moreover, although an agent need not have as a goal or intention such inevitable facts, this does not prevent her from reasoning about them. Thus, for example, on adopting the intention to go to the dentist, the agent could still use her beliefs about the certainty of accompanying pain in deciding to take along a strong analgesic.

Cohen and Levesque define a notion of persistent goal which appears to have some of the above properties. However, these properties are obtained by appealing to the temporal nature of persistent goals, rather than any intrinsic properties of beliefs, goals, or intentions [Allen, 1990]. For example, Cohen and Levesque can only avoid intending the side effects of any intended action if, at *some* time point in the future, the agent does not believe that the side effect will result from performance of the intended action. The problem remains, however, for cases in which the agent, for example, always believes that the side effect will occur. In other words, Case (b) is not satisfiable for the above propositions for the non-trivial cases. In contrast, in our formalism the agent is not forced to adopt unwanted goals or intentions on account of her beliefs, no matter how strong or persistent these beliefs are.

6 COMPARISON AND CONCLUSION

Bratman [Bratman, 1987] argues against the reducibility of intentions to an agent's beliefs and desires and treats intentions as partial plans of action to which the agent is committed. He then goes on to show how the agent's existing beliefs, desires, and intentions form a background for future deliberation. Following this line of argument, we have introduced a logical formalism which accords a primary status to intentions. Further, by adopting certain axioms of change, we have shown how the present beliefs, goals, and intentions of an agent constrain her future attitudes.

Philosophically, our approach differs from that of Cohen and Levesque [Cohen and Levesque, 1990] in that it treats intention as a basic attitude and shifts the emphasis of

future commitment from the definition of intention to the process of intention revision. Semantically, our approach differs in that we distinguish between the choice available to the agent in choosing her actions and her beliefs about which worlds are possible. In addition, we specify an interrelationship between beliefs, goals, and intentions that allows us to avoid all variations of the problem of unwanted side effects.

We have only considered constraints on the maintenance of intentions. Important aspects of rational behavior that concerns intention formation by *deliberation* and intention modification in the light of changing circumstances or *reconsideration* [Bratman, 1987] have not been dealt with in this paper. These are separate topics which will be considered elsewhere; the sub-world relationship between beliefs, goals, and intentions provides useful techniques for analyzing these issues.

In summary, we have presented a theory of intention that treats intentions on a par with the agent's beliefs and goals. By introducing various axioms of change, we were able to categorize a variety of rational agents and their commitment strategies. We also captured, for the first time, the process of belief, goal, and intention revision, which is crucial for understanding rational behavior [Allen, 1990]. Although there are many aspects of a theory of rational agency that we have not addressed, we believe that we have presented a formalism that provides a foundation upon which such a theory can be constructed.

Acknowledgements

The authors would like to thank Phil Cohen, Martha Pollack, Liz Sonenberg, David Israel, Kurt Konolige, Douglas Appelt, and Félix Ingrand for valuable discussions and comments on the contents of this paper.

References

[Allen, 1990] J. Allen. Two views of intention: Comments on Bratman and on Cohen and Levesque. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, Ma., 1990.

[Bratman *et al.*, 1988] M. E. Bratman, D. Israel, and M. E. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.

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

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

[Cohen and Levesque, 1987] P. R. Cohen and H. J. Levesque. Persistence, intention and commitment. In M. P. Georgeff and A. L. Lansky, editors, *Proceedings of the 1986 workshop on Reasoning about Actions and Plans*, pages 297–340. Morgan Kaufmann Publishers, San Mateo, CA, 1987.

[Cohen and Levesque, 1990] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(3), 1990.

[Emerson and Srinivasan, 1989] E. A. Emerson and J. Srinivasan. Branching time temporal logic. In J. W. de Bakker, W.-P. de Roever, and G. Rozenberg, editors, *Linear Time, Branching Time and Partial Order in Logics and Models for Concurrency*, pages 123–172. Springer-Verlag, Berlin, 1989.

[Gardenfors, 1988] P. Gardenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. Bradford Book, MIT Press, Cambridge, MA., 1988.

[Georgeff and Ingrand, 1989] M.P. Georgeff and F.F. Ingrand. Decision-making in an embedded reasoning system. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit, MI, 1989.

[Halpern and Moses, 1985] J. Y. Halpern and Y. O. Moses. A guide to the modal logics of knowledge and belief. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence(IJCAI-85)*, Los Angeles, CA, 1985.

[Hughes and Cresswell, 1984] G. E. Hughes and M. J. Cresswell. *A Companion to Modal Logic*. Methuen & Co. Ltd., London, England, 1984.

[McDermott, 1982] D. V. McDermott. A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6:101–155, 1982.

[Rao and Foo, 1989] A. S. Rao and N. Y. Foo. Minimal change and maximal coherence: A basis for belief revision and reasoning about actions. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI-89*, Detroit, MI, 1989.

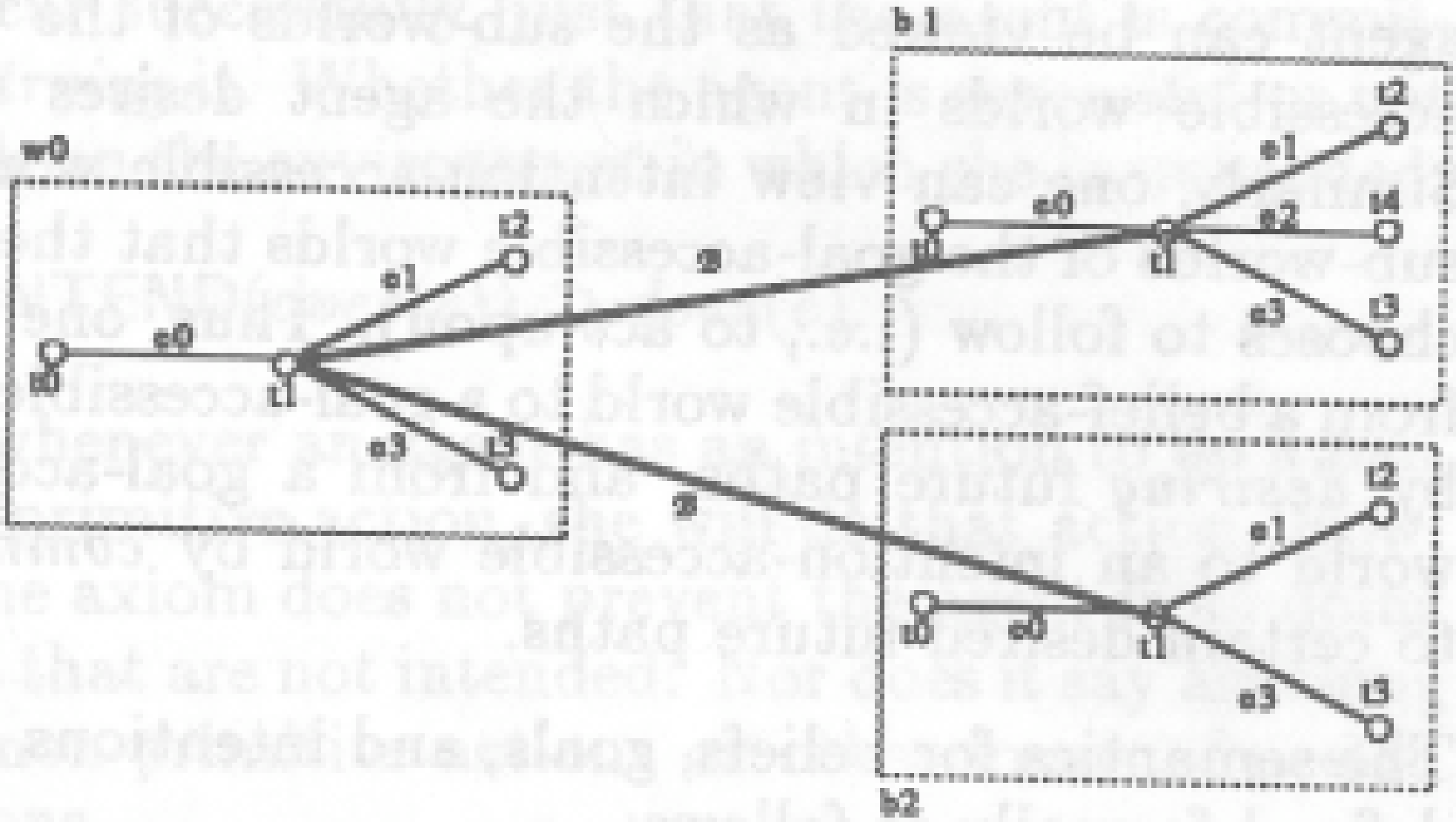
[Rao and Georgeff, 1990a] A. S. Rao and M. P. Georgeff. Deliberation and the formation of intentions. Technical Report 10, Australian Artificial Intelligence Institute, Carlton, Australia, 1990.

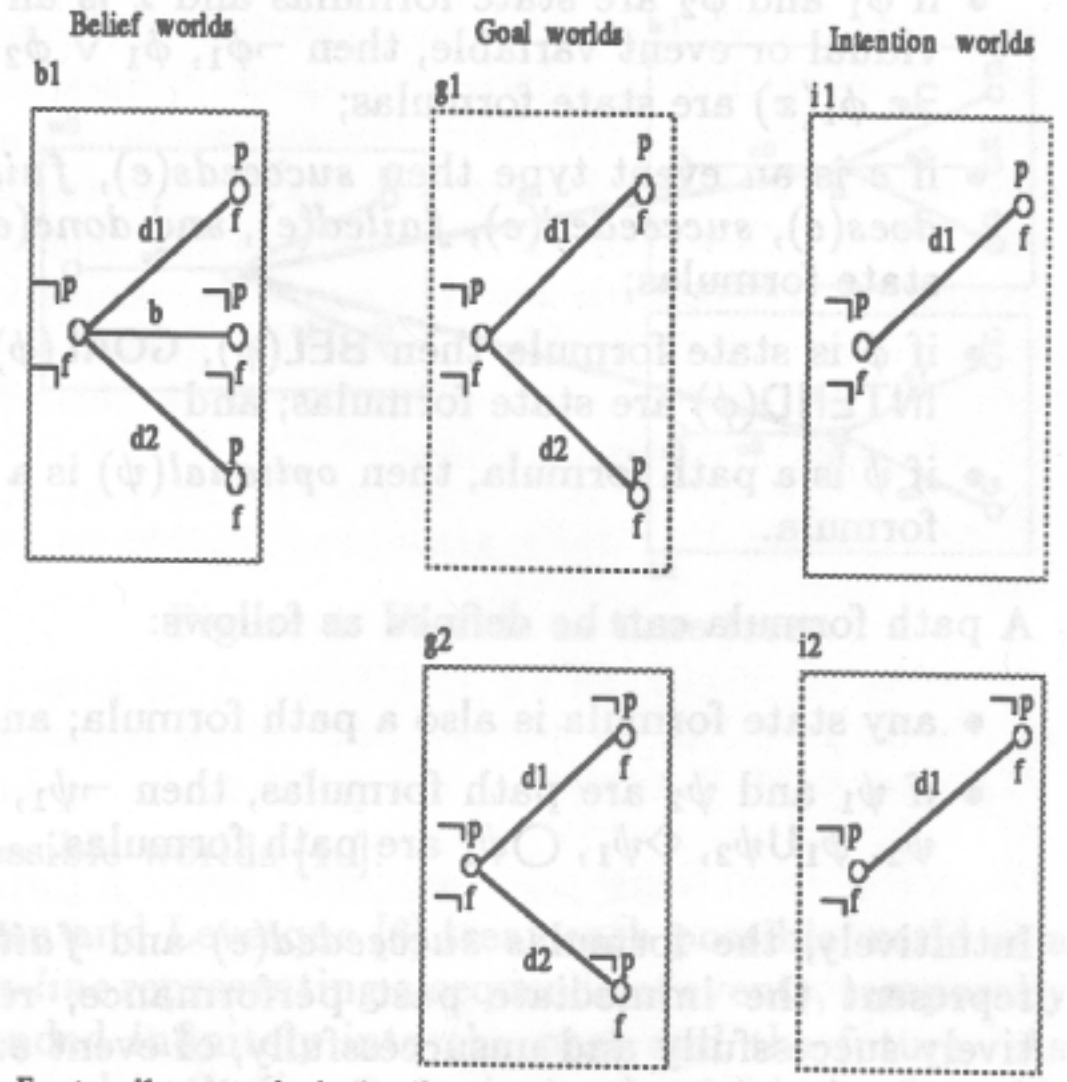
[Rao and Georgeff, 1990b] A. S. Rao and M. P. Georgeff. A formal model of intentions. In *Pacific Rim International Conference on Artificial Intelligence, PRICAI-90*, Nagoya, Japan, November 1990.

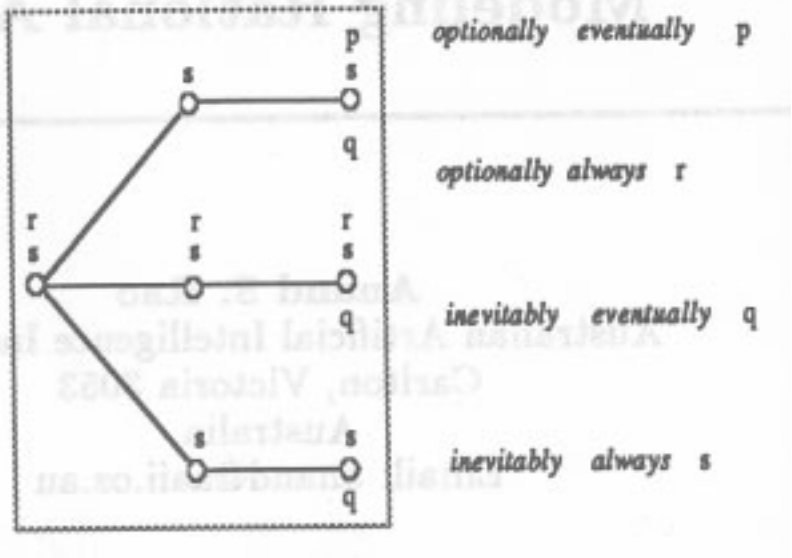
[Rao and Georgeff, 1991] A. S. Rao and M. P. Georgeff. Asymmetry thesis and side-effect problems in linear time and branching time intention logics. Technical Report 13, Australian Artificial Intelligence Institute, Carlton, Australia, 1991.

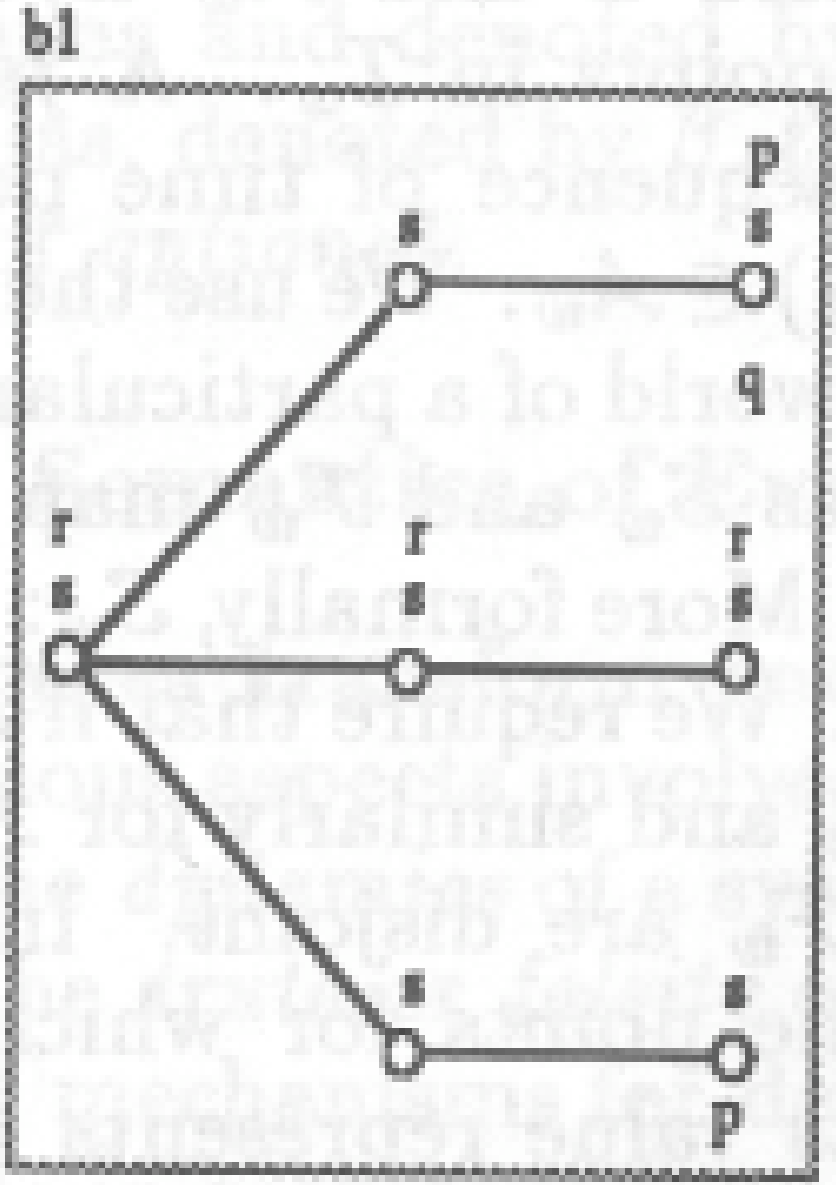
[Russell and Wefald, 1989] S. Russell and E. Wefald. Principles of metareasoning. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, Toronto, 1989.

[Vardi, 1986] M. Y. Vardi. On epistemic logic and logical omniscience. In J. Y. Halpern, editor, *Proceedings of the First Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 293–306, San Mateo, California, 1986. Morgan Kaufmann Publishers.

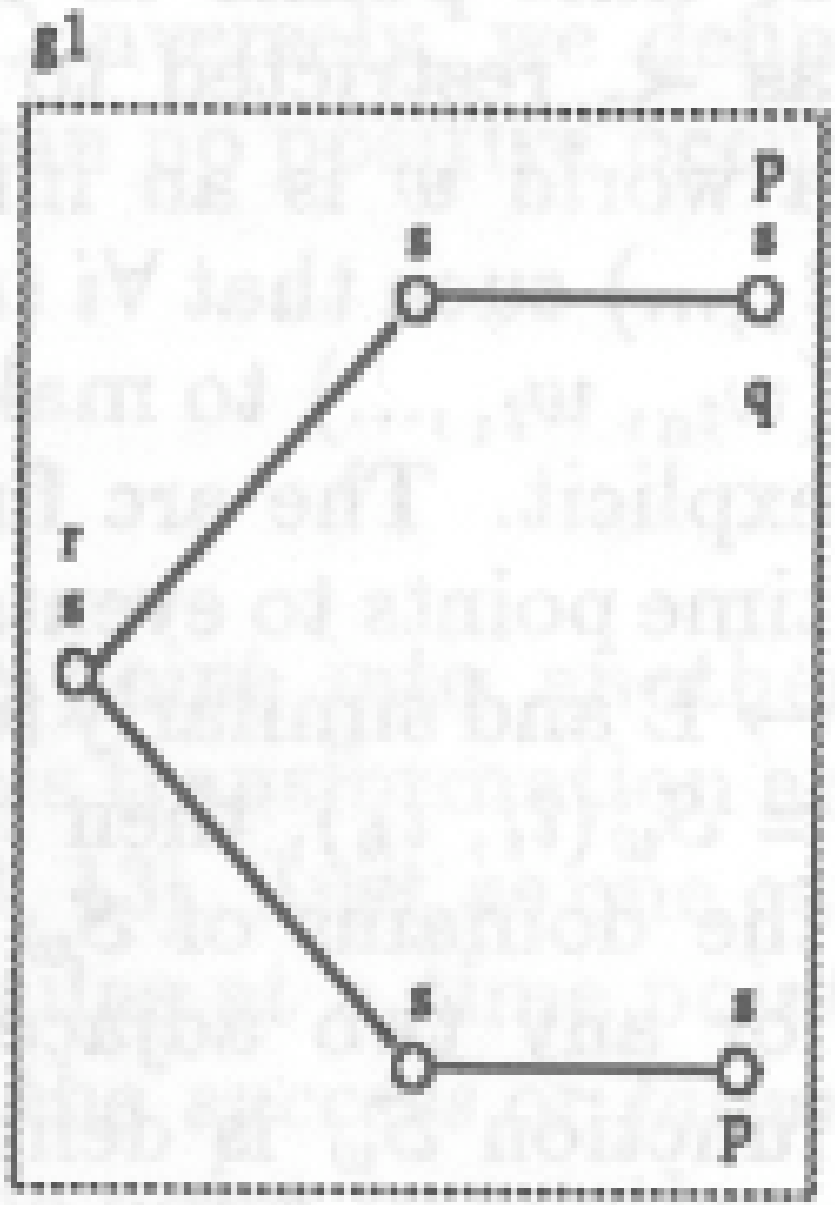

w0
e0
e1
e2
e3
l1
l2
l3
s
b1
e0
e1
e2
e3
l1
l2
l3
s
b2
e0
e1
e3
l1
l3


Belief worlds
Goal worlds
Intention worlds
b1
g1
i1
g2
i2
d1
d2
b
p
f
Events: d1 - go to dentist 1, d2 - go to dentist 2, b - go shopping
Facts: p - pain, f - tooth-filled


p
s
s
q
r
s
r
s
r
s
q
s
s
q
optionally eventually p
optionally always r
inevitably eventually q
inevitably always s



Belief-accessible World



Goal-accessible World