The Semantics of Intention Maintenance for Rational Agents

Michael P. Georgeffand Anand S. Rao
Australian Artificial Intelligence Institute
Level 6, 171 La Trobe Street, Melbourne
Victoria 3000, Australia

Abstract

The specification, design, and verification of agent-oriented systems depends upon having clear, intuitive formalisms for reasoning about the properties of such systems. In this paper, we consider agents whose state comprises the three mental attitudes of belief, desire, and intention. While the static relationships among these entities has had considerable attention, the manner in which these entities change over time has not been formalized rigourously. By considering some simple examples, we show that the formalization of some of these intuitions is problematic. We then examine these intuitions from a possible-worlds perspective and formally describe the dynamics of intention maintenance in the context of changing beliefs and desires. To solve the problems identified in the examples, and to properly capture our semantic intuitions about intention maintenance, we extend the standard logics by introducing forms for *only* modalities of belief, desire, and intention, along the lines of Levesque's only believe operator. This allows us to formalize the process of intention maintenance. We conclude by comparing our work with other related work.

1 Introduction

Agent-oriented systems are finding increasing application in the commercial world. One of the most successful of agent architectures is that based around the notions of *belief, desire,* and *intention* (BDI) [2; 3; 4; 8; 14], representing respectively the informative, motivational, and decision components of the agent. Such systems have been applied to a wide range of large-scale applications, including air traffic control, telecommunications network management, business process management, and simulation.

Within such systems, intentions play an essential role. First, prior intentions pose problems for further deliberation; in AI terms, they specify the goals (ends) for further means-ends analysis. Second, prior intentions constrain the deliberation process because they rule out options that conflict with existing intentions. Under this view, the deliberation process is a continuous resource-bounded activity rather than a one-off exhaustive decision-theoretic analysis [2].

The critical element in this view of practical reasoning is that the adoption of an intention entails some form of commitment to that intention. That is, intentions only have value if they are *maintained* from one time point to the next—if they are not so maintained, they can establish neither the goals for further deliberation nor the basis for ruling out conflicting options.

However, the specification, design, and verification of such systems depends on being able to semantically model these agents and formally describe the process of intention maintenance and the resulting agent behaviour.

A number of formalisms that provide the semantics of intention and its relation to the other attitudes, such as belief and desire, have been proposed in 'the literature. In providing these formalisms, various possible static relationships among belief, desire and intention have been considered. In essence, most of these reflect the intuition that one only adopts an intention to an action or proposition that is (i) desired and (ii) believed to be possible. The variations on this basic intuition concern certain special cases that one may or may not consider important, dependent on the purpose of the formalization.

In addition, some authors have proposed certain axioms to capture the dynamic relationships among these attitudes, particularly those concerning the maintenance of intentions. The intuition here is that an intention should be maintained as long as the object of the intention (i) is continued to be desired, (ii) is continued to be believed possible, and (iii) has not yet been achieved.[1]

Unfortunately, the translation of this condition into formal axioms of intention maintenance is more problematic than it first appears. In fact, it turns out that to express these dynamic properties of intention maintenance requires a more expressive logic than has been considered in the literature so far.

In this paper, we base our approach on a possible-worlds model developed previously [10; 12; 13]. However, the results we obtain apply more generally.

*In this paper we make the simplification that the object of the intention, once achieved, is no longer desired.

2 The Problem

For simplicity, let us consider the relationship between intentions and beliefs only. From the discussion above, one would expect the formalization of the maintenance condition for intentions to take something like the following form:

$\mathsf{INTEND}(X(\phi)) \wedge X(\mathsf{BEL}(\phi)) \supset X(\mathsf{INTEND}(\phi))$ where ϕ is a temporal formula and $X(\psi)$ states that ψ is true at the next time point.

Consider a situation in which John intends to go to the beach. From the above axiom, John will maintain this intention as long as he believes it to be achievable.[2] If, or when, John discovers that it is not possible to go to the beach, this intention can be dropped (and, indeed, the static constraints would force it to be dropped). This is just what we want.

However, let's assume that John also believes it is possible to fly to London, but has no intention of doing so. Because we have *\NTEND(X(go-to-bcach))* we also have (under a possible worlds model) the disjunctive intention I*NTEND(X(go-to-beach* V *go-to-London)).* Now, when it turns out that visiting the beach is impossible, the intention towards visiting the beach will be duly dropped. But, unfortunately, the intention towards the disjunction *(go-to-beach* V *go-to-London)* will be maintained (as the disjunct remains a possibility). From application of the static constraints, it can then be deduced that John, at the next time point, will intend to fly to London. In other words, John will be forced to adopt as new intentions any beliefs about the future he still holds!

A similar problem arises in the the following situation. John intends to obtain milk from the milk bar and cereal from the supermarket. He goes to the milk bar, sees that it is closed, and thus abandons the intention of obtaining milk. As a result John also gives up his intention to have milk and cereal. However, if intentions are closed under conjunction—as they are in a possible worlds model—intending to have milk and cereal implies an intention to have milk and an intention to have cereal. While the former two can no longer hold, using the above axiom of intention maintenance, the intention to have cereal would be (incorrectly) maintained.

Noting similar problems, Konolige and Pollack also considered closure under conjunction to be a problem for intentions, although in relation to static rather than dynamic properties. Their solution involves representationalist approach to the modelling of intentions [8].

But what is the real problem here? Is it simply that we do not want closure under conjunction, or is our simple axiomatization just not properly capturing our intuitions? While some of the undesirable symptoms of the problem are clear, the cause is not.

The approach we adopt here is to go back to our semantic model and understand what was really intended by the conditions of intention maintenance, and to develop axioms that properly reflect our semantic intuitions.

3 Informal Model

The formal and informal models of our BDI agents have been discussed elsewhere [10; 12; 13]. In this section, we briefly describe our model and then motivate the static and dynamic relationships between different entities within the model.

Our semantic model consists of sets of possible worlds where each possible world is a branching tree structure with a single past. A particular index within a possible world is called a time point or situation. The branches within a tree structure represent different courses of action or execution paths. We model the beliefs of the agent by a set of such possible worlds, called the belief-accessible worlds of the agent. Similarly, we model the desires and intentions of the agent by a set of desire-accessible worlds and a set of intention-accessible worlds, respectively.

The different belief-accessible worlds represent the agent's lack of knowledge or *chance* inherent in the environment; that is, as far as the agent knows, the actual world could be any one of the belief-accessible worlds. Within each belief world, each path represents the options or *choice* of action available to the agent.

Corresponding to each belief-accessible world is a desire-accessible world and an intention-accessible world.[3] These represent, respectively, the desires and intentions of the agent with respect to that particular belief world (that is, the desires and intentions the agent would have if that world was known to be the actual world). Each path within the desire-accessible world represents an execution path that the agent wants to achieve (or is content to achieve), and each path within an intention-accessible world represents an execution path that the agent has decided upon (one of which, in the context of our earlier discussion, the agent is committed to bringing about).

Now consider the static structural relationships among such a triple of belief-desire-intention worlds. While, for any such triple, we place no constraints on the relationship between the paths believed possible and the desired paths, we require that the intention paths be a subset of both (see Figure 1). This reflects the intuition that one will only intend a course of action that is both believed possible and desired.[4]

But what happens now as we move from one time point to the next (from *t* to *v* in world *w* as shown in Figure 1)? The basic intuition is that, provided the agent's beliefs and desires are not significantly changed, the agent's intentions will be maintained. More specifically, for any triple of belief-desire-intention worlds, we would like to retain any existing intention path provided it was still both believed possible and desired. Any intention path that was no longer believed possible, or was no longer desired, would be pruned off the intention structure. Any new belief paths, i.e., new opportunities (shown as a dotted path with r true in the future in

[2] Clearly, we need to add additional conditions to account for changing desires and the successful achievement of John's intentions. However, for simplicity, we do not consider these situations here.

[3] We elsewhere [10] consider the more general case where we relax the requirement for such a one-to-one correspondence.

[4] For discussion of this point, see our earlier work [10; 12].

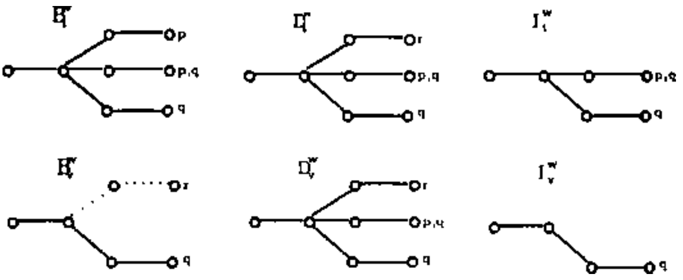

Figure 1: An example of Belief, Desire, and Intention Revision

Figure 1) are not considered (See Section 5.1 for more discussion on this). That is, we maintain as many old intention paths as possible consistent with satisfying the static structural constraints at the next time point.

In the following sections, we specify this model more formally and then develop axioms to reflect both the static and dynamic relationships. As we will see, this enterprise turns out to be less straightforward than the above semantic story would have us expect.

4 BDI Logic

To model agents, we use the language $\mathrm{BDI_{CTL^*}}$ [13], a propositional modal logic based on the branching temporal logic CTL* [6]. The primitives of $\mathrm{BDI_{CTL^*}}$ include a non-empty set Φ of *primitive propositions*; propositional *connectives* $\vee$ (or) and $\neg$ (not); *modal operators* BEL (agent believes), DES (agent desires), and INTEND (agent intends); and *temporal operators* X (next), U (until), F (sometime in the future or eventually), and E (some path in the future or optionally). Other connectives and operators such as $\wedge$, $\supset$, $\equiv$, G (all times in the future or always), A (all paths in the future or inevitably) can be defined in terms of the above primitives. For example, the expression BEL(EF(*at-beach*)) states that the agent believes that he has the option (dependent on what course of action he undertakes) to eventually be at the beach; and INTEND(AF(*at-beach*)) states that the agent intends (in all execution paths) to eventually be at the beach.

There are two types of well-formed formulas in the language: *state formulas* (which are true in a particular world at a particular time point) and *path formulas* (which are true in a particular world along a certain path). State formulas are defined in the standard way as propositional formulas, modal formulas, and their conjunctions and negations. The objects of E and A are path formulas.

The semantics for the above logic is given with respect to a possible-worlds model. A structure M is a tuple M = $\langle$W, $\{S_w\}$, $\{\mathrm{R}_w\}$, $\{\mathcal{B}_t^w\}$, $\{\mathcal{D}_t^w\}$, $\{\mathcal{I}_t^w\}$, $\{f_t^w\}$, $\{g_t^w\}$, $\{h_t^w\}$, $L\rangle$, where w and t range over W and S_w, respectively; W is the set of worlds; S_w is a set of time points in world w; $\mathrm{R}_w \subseteq S_w \times S_w$ is a total binary *temporal accessibility* relation; $\mathcal{B}_t^w$ is a set of belief-accessible worlds at world w and time t, i.e., $\mathcal{B}_t^w \subseteq$ W; $\mathcal{D}_t^w$ and $\mathcal{I}_t^w$ are *desire* and *intention accessible* worlds, respectively, that are defined in the same way as $\mathcal{B}_t^w$; the functions f_t^w, g_t^w, and h_t^w, are mappings from $\mathcal{B}_t^w$ to $\mathcal{D}_t^w$, $\mathcal{D}_t^w$ to $\mathcal{I}_t^w$, and $\mathcal{B}_t^w$ to $\mathcal{I}_t^w$, respectively; Φ is the set of primitive propositions; and L is a truth assignment function that assigns to each time point in a world, the set of propositions true at that time point.

The semantics of propositional formulas and temporal formulas are as given by us elsewhere [13]. Intentions are defined in a straightforward manner similar to any normal modal operator, i.e., an agent intends ϕ if ϕ is true in all intention-accessible worlds. More formally, $M, w_t \models$ INTEND(ϕ) iff $\forall$ $i \in$ W, if $i \in \mathcal{I}_t^w$ then $M, i_t \models \phi$. Similar definitions hold for beliefs and desires.

We adopt a weak S5 modal system for BEL, a K system for DES and INTEND, and the standard CTL axioms [6]. Following Emerson [6], we additionally define a *fullpath* for a world w to be an infinite sequence of situations starting from an initial situation t_0. The set paths(w) denotes the set of fullpaths for world w. We also assume total 1-1 mappings among belief-, desire-, and intention-accessible worlds.[5]

Consider now the structural relationships among belief-, desire-, and intention-accessible worlds. As discussed above, the most intuitive characterization is one in which every intention-accessible world is a sub-world of both its corresponding belief-accessible world and its corresponding desire-accessible world and there is at least one path common to each belief-accessible world and its corresponding desire-accessible world.

More formally, these constraints and their corresponding axioms can be expressed as follows:

(SC1) $\forall w$, $\forall t$, f_t^w, g_t^w, and h_t^w are total 1-1 mappings.
(SC2) $\forall b \in \mathcal{B}_t^w$ paths($h_t^w(b)$) $\subseteq$ paths(b).
(SC3) $\forall d \in \mathcal{D}_t^w$ paths($g_t^w(d)$) $\subseteq$ paths(d).
(SC4) $\forall b \in \mathcal{B}_t^w$, paths($b$) $\cap$ paths($f_t^w(b)$) $\neq \emptyset$.

(A1) INTEND(α) $\supset$ BEL(α).
(A2) BEL(β) $\supset$ INTEND(β).
(A3) INTEND(α) $\supset$ DES(α).
(A4) DES(β) $\supset$ INTEND(β).
(A5) BEL(α) $\supset$ $\neg$DES($\neg\alpha$).

The formula α is an optional formula and β is an inevitable formula [12]. Axioms A1 and A2 derive from

[5] We consider partial mappings elsewhere [10; 12].

the semantic constraints SCI and SC2. Axiom AI states that any intended execution path must be believed to be possible (that is, must be believed to be an option for the agent). Axiom A2 states that any inevitable belief will be intended.[6] Axioms A3 and A4, resulting from the constraints SCI and SC3, state that any path that is intended must be desired and any inevitable desire will be intended. Axiom A5, resulting from constraints SCI and SC4, states that at least one of the desired execution paths is believed achievable.

To preserve the mapping to decision trees [II], we make the following *deterministic world* assumption. This assumption requires for a given model, and all world time point pairs, that $\forall b, b' \in \mathcal{B}_t^w$, if $L(b, t) = L(b', t)$ then 6 = b', where *L* is the truth assignment function. Intuitively, this means that there is no additional non-determinism beyond that represented by different belief worlds. In other words, the real world is deterministic; any perceived non-determinism results from an agent's lack of knowledge of the world. Similar assumptions hold for desire- and intention-accessible worlds.[7]

We refer to the above axiomatization together with the axioms relating intention and action (see our earlier work[I2] for details) as the *BDI-modal system.* Other variations to this axiomatization can be obtained by allowing the total 1-1 mappings /, *g* and *h* to be partial, which account for the cases that have been referred to as realism [4], weak-realism [10], and strong-realism [10]. Different structural relationships can also be adopted among B-, *V-,* and Z-accessible worlds to obtain further variations in the axiomatizations.

It turns out, however, that under all these variants we need some additional expressive power to capture the notion of intention maintenance discussed above. To achieve this, we now extend the language BDICTL* by introducing *only* forms of the modalities for beliefs, desires, and intentions. Intuitively, if an agent *only intends* a formula ϕ then ϕ is true in all the intention-accessible worlds and the set of intention-accessible worlds includes all worlds where ϕ is true.

$$M, w_t \models \mathrm{OINTEND}(\phi) \text{ iff } \forall i \in W,$$
$$i \in \mathcal{I}_t^w \text{ iff } M, i_t \models \phi.$$

The definition of INTEND(ϕ) includes only the if part of the definition above. It is important to note that, whereas the operator INTEND is closed under conjunction, OINTEND is not. That is, we have the following theorem:

Theorem 1 *The following statements are true of the* OINTEND *operator.*

- $\not\models \mathrm{OINTEND}(\phi \wedge \psi) \wedge (\phi \not\equiv \psi) \equiv \mathrm{OINTEND}(\phi) \wedge \mathrm{OINTEND}(\psi)$;
- $\not\models (\mathrm{OINTEND}(\phi) \vee \mathrm{OINTEND}(\psi)) \wedge (\phi \not\equiv \psi) \supset \mathrm{OINTEND}(\phi \vee \psi)$;

The proof is straightforward [7]. For example, the above properties of the only intend modality entail that, if John only intends having milk and cereal for breakfast, he will not necessarily also only intend having milk and only intend having cereal. Similarly, if John only intends to go to the beach, he will not necessarily also only intend to go to the beach or only to go to London.

5 Maintenance of Intentions

Now let us consider the problem of an agent maintaining its intentions as the environment changes. Our aim is to specify semantic constraints on our models that will determine how the model changes from one time point to another. In so doing, we will treat the processes of belief and desire revision as given and consider how these processes determine intention revision.

Let us assume that the agent revises its beliefs using some well-known belief revision or update procedure [I]. For the purpose of this paper, we assume that the non-determinism (chance) inherent in the beliefs of the agent remains constant over time. Intuitively, this corresponds to an agent believing it is in one of a number of possible worlds, its beliefs about which can change over time, but about which it can never get sufficient information to eliminate any from consideration. It may, for example, discover that, for any particular possible world, it has different options than previously believed, but will not be able to reduce the uncertainty concerning which possible world it is actually in.

Under this assumption, at the semantic level the belief revision function is a total 1-1 mapping; that is, the belief revision process maps each old belief world into a corresponding new belief world. The propositions that hold in that new belief world may be quite different from those that held in the previous belief world, but no new belief worlds are introduced nor old ones deleted.

Although this seems restrictive, the assumption can be relaxed without too much difficulty by removing the semantic constraint SCI on the functions /, *g,* and *h.* However, for the purposes of this paper, this unnecessarily complicates the picture.

We therefore postulate a *belief revision functic* $\mathcal{BR}_t^w$ which maps each belief-accessible world to its revised belief-accessible world. More formally, we have:

Definition 1 *For each world w and time t the belief revision funct* $\mathcal{BR}_t^w$ *is a mapping from the set of belief-accessible worlds at t to the set of belief-accessible worlds at the next instant v. Formally* $\mathcal{BR}_t^w : \mathcal{B}_t^w \rightarrow \mathcal{B}_v^w$.

We postulate similar desire revision and intention revision functions for each world *w* and time point *t,* denoted by $\mathcal{DR}_t^w$ and $\mathcal{IR}_t^w$, respectively. Figure 2 shows the various functions involved in the revision process. Each solid circle represents a world which is a branching tree structure. The set of belief-accessible worlds at world *w* and time *t* has a total 1-1 mapping to its corresponding desire-accessible (denoted by ,f_t^w)l and intention-accessible worlds (denoted by h_t^w). The belief revision function maps each world in $\mathcal{B}_t^w$ to its corresponding world in $\mathcal{B}_v^w$ and similarly for the desire and intention revision functions. The functions ,f_v^w, g_v^w, and

[6] As discussed in our previous work [10] these axioms can be weakened by adopting alternative semantics constraints to that of SCI and SC2.

[7] The mappings /, *g,* and *h* are uniquely determined by the truth function assignment L, given the assumption of a deterministic world.

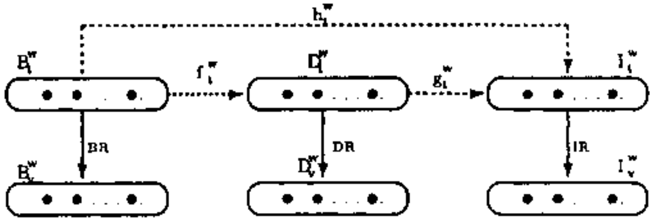

Figure 2: Schematic of Belief, Desire, and Intention Revision

h_v^w (not shown in the figure) provide similar mappings, but with respect to the revised worlds.

5.1 Belief Revision

Now we discuss how the intention revision function $\mathcal{IR}_t^w$ is related to the belief revision function and the previous intentions of the agent.

Intuitively, we want the intentions of the agent to be derived from its old intentions, but yet be compatible with the new beliefs adopted by the agent. We can model this semantically by treating each new belief-accessible world as a filter through which its corresponding old intention-accessible world is passed to derive the new intention-accessible world. Figure 1 gives a pictorial representation of the process. The belief accessible world at time t, has three paths, each ending with the propositions $\{p, q\}$, $\{p\}$, and $\{q\}$, respectively. Belief revision has resulted in all paths leading to p being removed, resulting in a new belief-accessible world at time t' with just one path leading to q. The intention-accessible world is a sub-world of the belief-accessible world at time t containing the two paths leading to q. Now we filter this world through the belief-accessible world at time t'. The only path of the intention-accessible world which passes through this filter is the path ending with the proposition $\{q\}$.

The same process can be applied to all the belief-accessible worlds of an agent (and their corresponding intention-accessible worlds) at any given time point. This reflects our intuition that the background intentions and beliefs help determine the future intentions of an agent [2].

More formally, we can write the belief filtration constraint as follows:

Definition 2 *(BFC1)* $\forall b \in \mathcal{B}_t^w$, *paths*$(\mathcal{IR}_t^w(h_t^w(b)))$ = *paths*$(\mathcal{BR}_t^w(b)) \cap$ *paths*$(h_t^w(b)) \neq \emptyset$.

The above constraint states that for all belief-accessible worlds at world w and time t, the set of paths of each revised intention-accessible world corresponding to each belief accessible world are equal to the intersection of the paths of the revised belief-accessible world and the previous intention-accessible world.[8]

Having formally specified the semantics of intention maintenance, the aim now is to construct an axiom (or axioms) that captures this semantic constraint. We first introduce the following Lemma.

Lemma 1 *If* OBEL(Aϕ) *and* $M, (p_t, p_{t+1}, \ldots) \models$ Aϕ, *then there exists a world* $w' \in \mathcal{B}_t^w$ *such that* $p \in$ *paths*(w'). *Moreover, the truth assignment function* L *uniquely determines this world.*

Proof: Consider a path p such that $M, (p_t, p_{t+1}, \ldots \models \phi$ and a world w' such that $M, w'_t \models$ Aϕ and $L(w', t) = L(p, t)$. If p is not in *paths*(w'), then we can create a new world w'', such that $w' \neq w''$, containing p as well as the paths in w'. Clearly, we also have, $M, w''_t \models$ Aϕ. Thus, from the definition of OBEL, $w' \in \mathcal{B}_t^w$ and $w'' \in \mathcal{B}_t^w$. From the assumption of a deterministic world, we have $w' = w''$ contradicting our original assumption. Hence, p is in *paths*(w'). The same result holds for OINTEND and ODES.

Now, it follows from the lemma that, if there exists a path in the old intention-accessible worlds that satisfies a formula that is now *only believed*, then that path must be in the intersection of the old intention-accessible worlds and the new belief-accessible worlds. In other words, from BFC1, the path must be included in the new intention-accessible worlds.

Theorem 2 *The following formula is valid in the class of models satisfying the belief filtration constraint (BFC1):*

(A6) INTEND(AXEξ) $\wedge$ AX(OBEL(Aξ)) $\supset$
AX(INTEND(Eξ)).

Proof:

1. *From* $M, w_t \models$ AXOBEL(Aξ) *applying the various definitions we have* $M, (p_x, p_{x+1}, \ldots) \models \xi$, $\forall x \in$ *succ*(w,t), $\forall b' \in \mathcal{B}_x^w$, $\forall p \in$ *paths*(b'), *where* $b' = \mathcal{BR}_t^w(b)$.
2. *From* $M, w_t \models$ INTEND(AXEξ) *applying the various definitions we have* $M, (q_v, q_{v+1}, \ldots) \models \xi$, $\forall i \in \mathcal{I}_t^w$, $\forall v \in$ *succ*(i,t), $\exists q \in$ *paths*(i).
3. $q \in$ *paths*(b') *{Lemma, 1 and 2 above}*
4. $q \in$ *paths*(i'), *where* $i' = \mathcal{IR}_t^w(h_t^w(b))$ *{From Constraint BFC1, the uniqueness of* h_t^w *and 3}*
5. $M, (q_x, q_{x+1}, \ldots) \models \xi$, $\forall i' \in \mathcal{I}_x^w$, $\forall x \in$ *succ*(w,t), $\exists q \in$ *paths*(i') *{From 2 and 4 above}*

[8] As with the sub-world relationship [12] we require the intersection to satisfy the truth assignment conditions at all the states of the paths.

6. $M, w_1 \models \mathsf{AXINTEND}(\mathsf{E}\xi)$ {*From definitions of* A, X, INTEND, *and* E}

By symmetry, we also have the following theorem.

Theorem 3 *The following formula is valid in the class of models satisfying the belief filtration constraint (BFC1):*

(A7) $\mathsf{OINTEND}(\mathsf{AXA}\xi) \wedge \mathsf{AX}(\mathsf{BEL}(\mathsf{E}\xi)) \supset \mathsf{AX}(\mathsf{INTEND}(\mathsf{E}\xi))$.

The proof is similar to that given above [7].

We have not proved completeness, and indeed it is unlikely that the above two maintenance axioms are complete with respect to the above belief-filtration constraint. However, the axioms appear sufficiently powerful to allow the kind of deductions needed for practical agent-oriented systems. For example, we can prove the following theorem which removes the problems associated with the two examples discussed earlier.

Theorem 4 *The following statements are true in the logic:*

$\not\models \mathsf{OINTEND}(\mathsf{AXAF}p) \wedge \mathsf{AX}(\mathsf{BEL}(\mathsf{EF}q)) \supset \mathsf{AX}(\mathsf{INTEND}(\mathsf{E}(\mathsf{F}p \vee \mathsf{F}q)))$;
$\not\models \mathsf{OINTEND}(\mathsf{AXA}(\mathsf{F}p \wedge \mathsf{F}q)) \wedge \mathsf{AX}(\mathsf{BEL}(\mathsf{EF}q)) \supset \mathsf{AX}(\mathsf{INTEND}(\mathsf{EF}q))$;

Proof: *The proofs follow directly from the earlier theorems.. The only intend operator prevents the conclusions of a disjunction* ($\mathsf{E}(\mathsf{F}p \vee \mathsf{F}q)$) *from either one of the disjuncts (in this case,* $\mathsf{EF}(p)$*); and similarly the only intend operator prevents the conclusion of separate conjuncts* ($\mathsf{EF}(p)$, $\mathsf{EF}(q)$) *from a conjunction* ($\mathsf{E}(\mathsf{F}p \wedge \mathsf{F}q)$).

Finally, it is worth considering two variations to the above model of intention maintenance: (a) what happens if the new belief-accessible world contains a new option (e.g., a path ending in the proposition r) that was not present in the previous time point; and (b) what happens if the filtered intention-accessible world has no future options (e.g., if the original intention-accessible world did not have the path ending only in *q).*

In the first case, intention maintenance will ensure the stability of intentions but does not allow the exploitation of new opportunities. As a result, any additional options that are part of the revised belief-accessible world will not be included in the corresponding new intention-accessible world. This is exactly what one wants for intention maintenance. However, this does not mean that new options can never be considered—an agent with sufficient computational resources may reconsider its intentions in the light of new opportunities. This can be modelled as a separate process following the above fdtering process.

In the second case, no intentions will be maintained and the agent has no choice but to reconsider his available options. That is, the agent would have to deliberate anew [II] to derive new intention-accessible worlds from its current belief- and desire-accessible worlds.

Similar results can be expected to hold when we relax the constraints that the revision functions be total 1-1 mappings (together with the semantic constraint SCI). However, this goes beyond the scope of this paper.

5.2 Desire Revision

The same belief filtration principle applies for desires as well. In other words, when an agent revises its intentions it should ensure that the new intentions are compatible with its new desires. Semantically, we therefore impose the constraint that the intention-accessible worlds are filtered through the corresponding revised desire-accessible worlds to obtain new intention-accessible worlds.

Of course, we want, our intentions to be compatible with both beliefs and desires. This results in the new intention paths being the intersection of new believed paths, new desired paths, and old intention paths.

(BDFC1) $\forall b \in \mathcal{B}_t^w$, $\mathrm{paths}(\mathcal{IR}_t^w(h_t^w(b))) = \mathrm{paths}(\mathcal{BR}_t^w(b)) \cap \mathrm{paths}(\mathcal{DR}_t^w(f_t^w(b))) \cap \mathrm{paths}(h_t^w(b)) \neq \emptyset$.

The theorem corresponding to the above constraint can be given as follows:

Theorem 5 *The following formulas are valid in the class of models that satisfy the filtration constraint (BDFC1):*

(A8) $\mathsf{INTEND}(\mathsf{AXE}\xi) \wedge \mathsf{AX}(\mathsf{OBEL}(\mathsf{A}\xi)) \wedge \mathsf{AX}(\mathsf{ODES}(\mathsf{A}\xi)) \supset \mathsf{AX}(\mathsf{INTEND}(\mathsf{E}\xi))$.
(A9) $\mathsf{OINTEND}(\mathsf{AXA}\xi) \wedge \mathsf{AX}(\mathsf{BEL}(\mathsf{E}\xi)) \wedge \mathsf{AX}(\mathsf{DES}(\mathsf{E}\xi)) \supset \mathsf{AX}(\mathsf{INTEND}(\mathsf{E}\xi))$

We refer to the BDI modal system together with the axioms A8 and A9 as the *dynamic BDI modal system.*

6 Comparison and Conclusion

Cohen and Levesque [4] define the notion of intention in terms of the other entities, such as beliefs, goals, persistent goals, and actions. In their formalism, an agent has a *persistent goal* or PGOA(ϕ) if and only if the agent currently believes $\neg\phi$, has the goal to eventually make ϕ true, and maintains this goal until it either comes to believe in ϕ or comes to believe that ϕ is impossible. PGOAL is closed under conjunction except in the special case where the agent already believes that one of the conjuncts is true or when the conjuncts hold at different time points. As neither example given in Section 2 is one of these special cases, the problems identified there are also exhibited in Cohen and Levesque's theory. Similarly, PGOAL is closed under disjunction except in very special circumstances. One could rectify the problems by adopting a similar approach to that used here.

As mentioned earlier, Konolige and Pollack [8] claim that Normal Modal Logics (NML) are not suitable for modelling intentions. They introduce a model of intentions that has two components: "possible worlds that represent possible future courses of events, and *cognitive structures,* a representation of the mental state components of an agent" [8].

They define a *scenario* for a proposition ϕ as the set of worlds in W that make ϕ true, denoted by M_ϕ. An agent intends ϕ iff the set of scenarios for ϕ is identical to the set of scenarios for any intention in the cognitive structure of the agent. This has an interesting correlation with our definition of OINTEND, if one considers each of their intention worlds as a path in our branching tree structures. The primary difference between the two approaches being that Konolige and Pollack follow a syntactic or representationalist approach and we follow a semantic approach. As a consequence, in their approach one has to explicitly conjoin formulas in the set of intentions given by the cognitive structure. Our semantic approach makes this unnecessary. Moreover, and perhaps more importantly, the semantic approach allows us to address the cause of the problem, not its symptoms.

Konolige and Pollack do not address the issue of belief and intention revision but do extend the notion of cognitive structures in terms of the plans of an agent. In this paper, we have explored the role of the only modalities in intention revision, but have remained silent on the important notion of plans [12].

The only modality was introduced by Levesque [9] in the context of beliefs and non-monotonic reasoning to capture the notions of stable sets in autoepistemic logic on a semantic basis. We have used the same concept for all the mental attitudes of the agent to give semantic characterizations of intention revision.

The primary contribution of this paper has been to lay out a semantic story of intention maintenance in the context of changing beliefs and desires. By introducing the only modalities to exactly specify paths of execution, we have also been able to provide a sound axiomatization of the intention maintenance process.

Of course, considerable work remains to be done. The completeness of the axiomatization needs further investigation. In addition, the restrictive conditions on the correspondence functions relating beliefs, desires, and intentions need to be removed and the proofs redone in this context. Finally, we need to show clearly how all this fits equally well within a decision-theoretic framework.

Acknowledgements: This research was supported by the Cooperative Research Centre for Intelligent Decision Systems under the Australian Government's Cooperative Research Centres Program.

References

[1] C. Alchourron, P. Gardenfors, and D. Makinson. On the logic of theory change: Partial meet contraction functions and their associated revision functions. Journal of Symbolic Logic, 50:510-530, 1985.

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

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

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

[5] J. Doyle. A truth maintenance system. Artificial Intelligence, 12:231-272,1979.

[6] E. A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, Handbook of Theoretical Computer Science: Volume B, Formal Models and Semantics, pages 995-1072. Elsevier Science Publishers and MIT Press, Amsterdam and Cambridge, MA, 1990.

[7] M. P. Georgeff and A. S. Rao. The semantics of intention maintenance for rational agents. Technical Report 57, Australian Artificial Intelligence Institute, Melbourne, Australia, 1995.

[8] K. Konolige and M. Pollack. A representationalist theory of intention. In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93), Chamberey, France, 1993.

[9] H. J. Levesque. All I know: A study in autoepistemic logic. Artificial Intelligence, 42(3), 1990.

[10] A. S. Rao and M. P. Georgeff. Asymmetry thesis and side-effect problems in linear time and branching time intention logics. In Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI-9I), Sydney, Australia, 1991.

[11] A. S. Rao and M. P. Georgeff. Deliberation and its role in the formation of intentions. In Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence (UAI-9I). Morgan Kaufmann Publishers, San Mateo, CA, 1991.

[12] A. S. Rao and M. P. Georgeff. Modelling rational agents within a BDI-architecture. In J. Allen, R. Fikes, and E. Sandewall, editors, Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning. Morgan Kaufmann Publishers, San Mateo, CA, 1991.

[13] A. S. Rao and M. P. Georgeff. A model-theoretic approach to the verification of situated reasoning systems. In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAJ-98), Chamberey, France, 1993.

[14] A. S. Rao and M. P. Georgeff. BD1 Agents: From theory to practice. In Proceedings of the International Conference on Multi-Agent Systems (1CMAS-95), San Francisco, USA, June, 1995.