

Model Theory for PRS-Like Agents: Modelling Belief Update and Action Attempts

Wayne Wobcke

School of Computer Science and Engineering
University of New South Wales
Sydney NSW 2052, Australia
wobcke@cse.unsw.edu.au

Abstract. In this paper, we extend our earlier work on modelling the mental states of PRS-like agents by considering the dynamics of belief and modelling of action attempts. The major constraint on our theory is that belief update is modelled within the theory of action as part of the logic ADL, a logic of belief, desire and intention that incorporates propositional dynamic logic. Some logical properties of belief update and attempts are given. The account provides a more complete modelling of both the statics and dynamics of agent programs based on the PRS-like architecture, and thus is a suitable foundation for developing model checking algorithms for this class of agents.

1 Introduction

In previous work, we developed a modelling of the mental states of a class of BDI agents based on a new logic called Agent Dynamic Logic (ADL) that combines elements from Emerson and Clarke's Computation Tree Logic [3], Pratt's Propositional Dynamic Logic [6] and Rao and Georgeff's BDI Logic [7]. The motivation of that work was to develop a logical framework that is closely aligned to the operational behaviour of a range of BDI agent architectures – those based on the PRS system – which we called *PRS-like* architectures. We take the PRS-like family to include PRS itself, Georgeff and Lansky [4], as well as derivative architectures such as UM-PRS, C-PRS, AgentSpeak(L), JACK Intelligent AgentsTM and JAM.

The formalism in that paper was explicitly limited to modelling an agent's mental state at single points in time, and did not address the issue of the relationships between an agent's mental states at different time points, i.e. the *dynamics* of mental states. Subtly, even though the formalism is based on temporal logic and thus includes aspects of the past and future, especially in that an intention is characterized as some action the agent believes it will eventually successfully perform, this is not enough to capture the commitment through time to those intentions as the future unfolds, and as it becomes necessary to update beliefs and make choices to commit to particular intentions.

The main objective of this paper is to extend our earlier modelling to incorporate belief update in a purely semantical way. This is done by further restricting

the class of PRS interpretations to those in which belief update conforms to the standard operational definition used in the PRS-like architecture. The major technical constraint that we adhere to in this paper is that the belief update of an agent must be modelled as a by-product of action performance within possible worlds models that correspond to execution structures of the agent. This enables belief update to be formalized within Agent Dynamic Logic as part of the semantics of actions; this in turn being possible only because of the restricted belief update functions used by PRS-like systems.

The issues involved in developing a purely semantical account of PRS-like agents turn out to be quite complex, and relate to other fundamental issues such as the notion of “possibility” underlying possible worlds models as used in BDI logic, Bratman’s two faces of intention (future and present directed) [1] and his idea of acting with an intention (doing one thing in order to achieve some intended effect), the treatment of action success and failure in BDI models, the modelling of attempted actions (the agent intends an action but can plan only to attempt the action) and the relation between intention and control (the agent should attempt only actions somehow under its direct control). Of course, these issues are themselves interrelated. Accordingly, we begin the paper with a summary of how these issues relate to the semantics of PRS-like architectures, then proceed to the technical definitions of Agent Dynamic Logic.

2 Motivation

One key aspect of the present approach to modelling PRS-like agents is to characterize beliefs, desires and intentions using the dynamic logic of action, where the semantics of an agent program is defined purely in terms of its execution structures. This is in line with work in Computer Science where the semantics of programs is given in terms of state transition relations, enabling model checking algorithms that construct such execution structures to be used for verifying properties of programs. However, the semantics of BDI agent programs is often given in terms of possible worlds models, e.g. Cohen and Levesque [2], Rao and Georgeff [7], Wooldridge [10], where the connection (if any) of the possible worlds models to execution structures is at best indirect.

A core issue to be resolved is what “possible” means in such models – far from being a peripheral question, this turns out to be central to the correct modelling of belief update functions using execution structures. In our framework, basic *execution structures* are analogous to computation trees where the transitions between states are all generated by transition relations defining the semantics of actions; the beliefs of an agent are modelled as a set of “epistemic alternatives”. The natural way to model belief update is to derive a set (or sets) of epistemic alternatives for the belief state(s) after the agent executes an action from the initial set of epistemic alternatives and the semantics of the action performed. Putting these ideas together gives a significant restriction: the execution structures of the epistemic alternatives of the agent must all be somehow realizable (generable from transition relations). The agent is not free to imagine (believe possible) execution structures that are not realizable.

A simple example illustrates the nature of this restriction. Suppose the agent has a plan for starting a car that consists of the action of turning the key in the ignition. The semantics of “turning the key” is that when the battery is charged, the car starts, and when the battery is dead, the car does not start. If the agent does not have any belief about the battery, its belief state is characterized by two epistemic alternatives. Suppose the battery is in fact dead and the agent turns the key. In that alternative where the battery is dead (the “actual world”), the car doesn’t start and in the resulting state the battery is still dead. However, in the other *epistemic* alternative (where the battery is charged), executing the action does *not* result in a state in which the car starts, because whether the car starts is determined by which world is the actual world (which world the agent is “in”), not by what the agent believes (even more, the agent may observe that the car doesn’t start, and may come to believe the battery is dead because of this). This scenario is different from that where there is an alternative actual world in which the battery is actually charged, where the car starts in both epistemic alternatives of the agent, assuming successful key turning, etc.

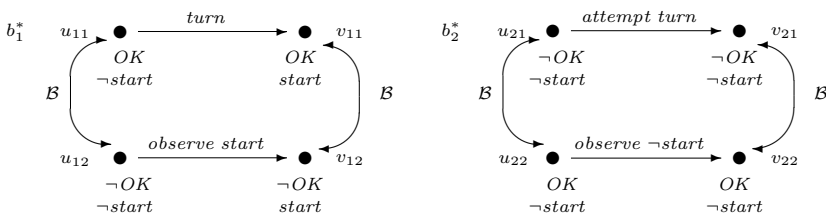


Fig. 1. Execution structures for differing actual worlds b_1^* and b_2^*

The scenarios are as illustrated in Figure 1. There are two execution structures, differing according to which world corresponds to the “actual world”, denoted b_1^* and b_2^* in the figure. The transitions corresponding to turning the key in the actual worlds are assumed to be as described above; the transitions in the epistemic alternatives correspond to updates based on observations made of the actual world, e.g. the transition from u_{12} to v_{12} is a transition allowed by the action *observe start*, assuming here successful execution of the action and that the agent observes (only) whether the car starts. The sets of resulting states $\{v_{11}, v_{12}\}$ and $\{v_{21}, v_{22}\}$ characterize the updated belief states of the agent after executing the action in the respective actual worlds (b_1^* and b_2^*). In our semantics, there are thus two grades of “epistemic alternatives” – that arising from the agent’s ignorance of which world it is inhabiting (which we call alternative actual worlds), and within some actual world, ignorance of what the state is in that world (which we call epistemic alternatives).

Belief update is modelled as defined in the standard operational definitions of PRS-like agent architectures. However, we make some simplifying assumptions about what the agent observes at any given point in time. First, we assume that the agent’s observations are veridical (the agent only observes what is true). This

simplifies the way that the set of states modelling the agent's beliefs after performing an action are related to the states resulting from the transition relation modelling the action performed (specifically, when the initial state in the actual world is one of the agent's epistemic alternatives and the agent observes all relevant changes, the resulting state after action execution is also one of the agent's epistemic alternatives). Second, we assume the agent observes whether the post-condition of the intended action holds in the resulting state. Together with the first assumption, it follows that the agent knows (after the event) whether the actions it attempts are successful, and that this is reflected in its beliefs.

3 Agent Dynamic Logic

In this section, we summarize our extended approach to defining the semantics of PRS-like agents' mental state dynamics; the initial theory without reference to dynamics was given in Wobcke [9]. The framework is based on a new logic called Agent Dynamic Logic (ADL) that combines aspects of Computation Tree Logic (CTL), Propositional Dynamic Logic (PDL) and the BDI Logic of Rao and Georgeff (called here BDI-CTL). In combining modal and dynamic logics, ADL is related to the logic of van der Hoek *et al.* [8], and similar to their approach, we use actions *observe p* to model belief update resulting from observations of basic propositions *p*. The term *observe* is chosen to emphasize that the agent is actively involved in making the observation, and does not simply receive information in a passive process of belief update.

The language ADL (Agent Dynamic Logic) is based on both BDI-CTL, which extends CTL with modal operators for modelling beliefs, desires (goals) and intentions, and PDL, which includes modal operators corresponding to program terms. Our definitions of BDI-CTL are modifications of Rao and Georgeff's in that, though there are three modal operators, **B** (belief), **G** (goal) and **I** (intention) in the language, the operators **G** and **I** are defined in terms of other primitives. We assume there is a base propositional language \mathcal{L} for expressing time-independent properties of states. The language of ADL includes the formulae of BDI-CTL (which includes the CTL state formulae) plus formulae built using the PDL constructs, as defined more precisely below. We assume knowledge of the basic CTL definitions.

Definition 1. A BDI interpretation is a tuple $\langle \mathcal{T}, \prec, \mathcal{S}, \mathcal{W}, \mathcal{A}, \mathcal{B}, \mathcal{I} \rangle$ where $\langle \mathcal{T}, \prec \rangle$ is a time tree, \mathcal{S} is a nonempty set of states, \mathcal{W} is a nonempty set of worlds based on \mathcal{S} with each world over a subtree of $\langle \mathcal{T}, \prec \rangle$, \mathcal{A} and \mathcal{B} are subsets of $\mathcal{W} \times \mathcal{T} \times \mathcal{W} \times \mathcal{T}$ defined only for tuples (w, t, w', t') for which t (t') is a time point in w (w') and w_t and w'_t share a common history (defined formally below), and \mathcal{I} is a function $\mathcal{W} \times \mathcal{T} \rightarrow \mathcal{W}$ mapping each time point t in a world w to a subworld of w containing t .

Definition 2. A subworld of a world w over a time tree $\langle \mathcal{T}, \prec \rangle$ based on a set of states \mathcal{S} is the world w restricted to a subtree of $\langle \mathcal{T}, \prec \rangle$ whose root is the root of w .

As mentioned above, the agent needs to keep track of two senses of epistemic alternative: alternative actualities and alternative epistemic states with respect to one actual world. The relation \mathcal{A} is for alternative actual worlds, while the relation \mathcal{B} captures ignorance of an agent with respect to one actual world. More formally, the beliefs of the agent in some situation w_t (at a time point t in a world w) are precisely those propositions holding at all epistemic alternatives w' of w at t' , i.e. in those situations w'_t , such that $\mathcal{B}(w, t, w', t')$. The relation \mathcal{A} is used as part of the definition of intentions; no matter how the world is actually, the agent believes it will eventually perform the intended action on all possible futures. Both relations \mathcal{A} and \mathcal{B} on situations are assumed to be serial, transitive and Euclidean, and in addition \mathcal{A} is assumed to be reflexive. Also as noted above, w_t and w'_t are required to share a common history (defined formally below, but intuitively the sequence of prior states and the agent's beliefs at those states must be identical). As consequences, (i) the agent always “knows the time” in that the epistemic alternatives of a situation are always at corresponding time points in the execution structures, and (ii) w_t may have as an alternative another situation in w (though only one at a corresponding time point – the time at a situation can be counted as the number of situations in its history). Finally, the relations \mathcal{B} and \mathcal{I} are assumed to satisfy the condition that $\mathcal{B}(w, t, w', t')$ iff $\mathcal{B}(\mathcal{I}(w, t), t, \mathcal{I}(w', t'), t')$, i.e. the epistemic alternatives of $\mathcal{I}(w, t)$ at t are the intended subworlds of the epistemic alternatives of w at t .

Definition 3. *Let $\langle \mathcal{T}, \prec, \mathcal{S}, \mathcal{W}, \mathcal{A}, \mathcal{B}, \mathcal{I} \rangle$ be a BDI interpretation. Then a world $w \in \mathcal{W}$ satisfies a BDI-CTL formula at a time point t in w as follows.*

$$\begin{aligned}
 w \models_t \mathbf{B}\alpha & \quad \text{if } w' \models_{t'} \alpha \text{ whenever } \mathcal{B}(w, t, w', t') \\
 w \models_t \mathbf{I}\pi & \quad \text{if } \mathcal{I}(w', t') \models_{t'} \mathbf{A}\Diamond do(\pi) \text{ whenever } \mathcal{A}(w, t, w', t') \\
 w \models_t \mathbf{G}\gamma & \quad \text{if } w \models_t \mathbf{I}(\text{achieve } \gamma)
 \end{aligned}$$

We impose two additional constraints: (i) $w \models_t do(\text{attempt } \pi)$ iff $\mathcal{I}(w, t) \models_t do(\pi)$, and (ii) if $\mathcal{B}(w, t, w', t')$ then when $w \models_t do(\text{attempt } \pi)$, for any path p containing t , $\mathcal{B}(w, s_p(t), w', s_{p'}(t'))$ for any path p' containing a successor t' of t for which $(w'_t, w'_{s_{p'}(t')}) \in R_{\text{observe } \alpha}$, and $\mathcal{B}(w, s_p(t), w, s_{p'}(t'))$ for any path p' containing a successor t' of t in w which satisfies α , for some α . The first constraint means that the agent's intended futures are exactly those where the actions attempted by the agent are successfully performed, and that the agent's attempts are based solely on its mental state, not on the state of the world. The second constraint means that the epistemic alternatives of the agent after performing an action are a partition of those derived from the set of prior epistemic alternatives using the semantics of action execution and belief update, as observations vary. The part of the definition relating to epistemic alternatives in the actual world is used to capture ignorance arising from a failure to distinguish the outcomes of a nondeterministic action.

We can now present the definition of Agent Dynamic Logic (ADL). Analogous to PDL, the language of ADL includes modal operators $[\pi]$ and $\langle \pi \rangle$ corresponding to each program π , and the semantics is based on computation trees, as in the approach of Harel [5]. The programs are assumed to consist of a set of atomic

programs that can be combined with a unary operator $*$ (iteration) and binary operators $;$ (sequencing) and \cup (alternation), and corresponding to any formula α of the base language \mathcal{L} is a test statement $\alpha?$ – note, however, that the test is on whether α is a belief of the agent, not whether α holds in the environment. The language of programs also includes special actions *achieve* γ , *attempt* π and *observe* α (where γ and α are formulae of the base language \mathcal{L} with α a conjunction of literals, and π is a program), an “empty” program A and a dummy action *env* that models the changes in the environment that occur over cycles in which the agent tests its beliefs.

The semantics of ADL is based on what we call *PRS interpretations*. Each program is modelled as a set of reduced BDI interpretations that arises by varying the actual world and belief accessibility relation \mathcal{B} ; these are reduced BDI interpretations in that the relations \mathcal{T} and \mathcal{A} play no role and are therefore omitted from the definitions (note that the \mathcal{B} relation is used to define the meaning of the test statements and the semantics of belief update, so cannot be omitted). In any reduced BDI interpretation b modelling a program π , one distinguished world b^* models the execution paths of π in the actual world and may have *non-final* situations, situations at leaf nodes in a computation tree where execution can continue (other worlds in b represent the agent’s belief states, and final situations in b^* represent paths where program execution cannot continue). Each primitive action π (except for the special action *achieve* γ – see below) is modelled as in PDL as a binary relation on states (independent of the agent’s beliefs), denoted R_π . In ADL, composite actions are modelled as sets of reduced BDI interpretations whose worlds derive from these state relations, one for each state in which the action is executable and each possible observation.

Definition 4. A PRS interpretation is a pair $\langle \mathcal{S}, \mathcal{R} \rangle$, where \mathcal{S} is a set of states and \mathcal{R} is a family of sets of reduced BDI interpretations \mathcal{R}_π based on \mathcal{S} , one such set for each program π , for which (i) for each atomic program π except those of the form *achieve* γ , $R_\pi \subseteq R_{\text{attempt } \pi}$, and (ii) in any reduced BDI interpretation $b \in \mathcal{R}_\pi$, transitions in b^* are state transitions from R_π while state transitions in other worlds in b are from $R_{\text{observe } \alpha}$ where α is some conjunction of literals β for which the postcondition of π logically implies either β or $\neg\beta$.

Definition 5. The binary state relation $R_{\text{observe } \alpha}$ is defined (for α a conjunction of literals) as follows.

$$(s, t) \in R_{\text{observe } \alpha} \text{ iff } s \models \alpha \text{ and } t = \{s - \bar{\alpha} \cup \alpha^+\}$$

where if α is $\alpha_1 \wedge \dots \wedge \alpha_n$, α^+ is the set of α_i and $\bar{\alpha}$ is the set of literals complementary to the α_i .

It is obvious that if s is a state (corresponding to a consistent complete theory over the base propositional language \mathcal{L}) and $(s, t) \in R_{\text{observe } \alpha}$, then t is also a state (i.e. is also a consistent and complete theory over \mathcal{L}), so the binary state relation is well defined. It is also clear that if K is a belief set of the agent (a consistent, but not necessarily complete, theory over \mathcal{L}) and K_α^* is the revised belief set as defined in PRS-like systems (extending the definition above

to any consistent theory), then if $K = \cap\{s : s \in \mathcal{S}\}$, $K_\alpha^* = \cap\{t : s \in \mathcal{S}, K \subseteq s \text{ and } (s, t) \in R_{\text{observe } \alpha}\}$. This ensures that the state transitions based on an observation action taken from a set of epistemic alternatives correctly capture the update of the agent's belief set.

Definition 6. *The relations $\mathcal{R}_{\text{achieve } \gamma}$ corresponding to an achievement subgoal γ are defined as follows.*

$$\mathcal{R}_{\text{achieve } \gamma} = \uplus\{\mathcal{R}_\pi \cap \Gamma : \pi \in L, \text{post}(\pi) \vdash \gamma\}$$

Here Γ is the set of reduced BDI interpretations b where the final situations of b^* satisfy γ , and $\text{post}(\pi)$ is the postcondition of a plan π in the plan library L .

Definition 7. *Let $\langle \mathcal{S}, \mathcal{R} \rangle$ be a PRS interpretation and $\langle \mathcal{T}, \prec, \mathcal{S}, \mathcal{W}, \mathcal{B} \rangle$ be a reduced BDI interpretation. For a world $w \in \mathcal{W}$ over $\langle \mathcal{T}, \prec \rangle$ containing a point t , let w^t be the subworld of w over $\langle \mathcal{T}_t, \prec_t \rangle$, the subtree of $\langle \mathcal{T}, \prec \rangle$ generated from t . Then w satisfies $\text{do}(\pi)$ and $[\pi]\alpha$ at a time point t in w as follows.*

$$\begin{aligned} w \models_t \text{do}(\pi) & \quad \text{if there is a reduced BDI interpretation } b \text{ in } \mathcal{R}_\pi \text{ such} \\ & \quad \text{that } b^* \text{ is isomorphic to a prefix of } w^t \\ w \models_t [\pi]\alpha & \quad \text{if for every reduced BDI interpretation } b \text{ in } \mathcal{R}_\pi \text{ such} \\ & \quad \text{that } b^* \text{ is isomorphic to a prefix of } w^t, w \models_u \alpha \text{ for} \\ & \quad \text{every point } u \text{ in } w \text{ corresponding to a leaf node of } b^* \end{aligned}$$

Definition 8. *A world w is a prefix of a world w' if for each end node n of w , there is a world w_n such that replacing each n in w by w_n results in w' .*

Definition 9. *A world w^1 in a reduced BDI interpretation $\langle \mathcal{T}_1, \prec_1, \mathcal{S}, \mathcal{W}_1, \mathcal{B}_1 \rangle$ is isomorphic to a world w^2 in a reduced BDI interpretation $\langle \mathcal{T}_2, \prec_2, \mathcal{S}, \mathcal{W}_2, \mathcal{B}_2 \rangle$ if there is a one-one correspondence f between \mathcal{T}_1 and \mathcal{T}_2 such that for all $t, u \in \mathcal{T}_1$, $t \prec_1 u$ iff $f(t) \prec_2 f(u)$, and for all $t \in \mathcal{T}_1$, w_t^1 is equivalent to $w_{f(t)}^2$.*

Definition 10. *A situation $w_{t_1}^1$ in a reduced BDI interpretation $\langle \mathcal{T}_1, \prec_1, \mathcal{S}, \mathcal{W}_1, \mathcal{B}_1 \rangle$ is equivalent to a situation $w_{t_2}^2$ in a reduced BDI interpretation $\langle \mathcal{T}_2, \prec_2, \mathcal{S}, \mathcal{W}_2, \mathcal{B}_2 \rangle$ if $w_{t_1}^1 = w_{t_2}^2$ (i.e. they are equal as states) and $\{u_{t_1}^1 : \mathcal{B}_1(w^1, t_1, u, t_1)\} = \{v_{t_2}^2 : \mathcal{B}_2(w^2, t', v, t_2)\}$ (i.e. they satisfy the same basic beliefs).*

Definition 11. *Two situations $w_{t_1}^1$ and $w_{t_2}^2$ in reduced BDI interpretations b^1 and b^2 share a common history if all corresponding pairs of situations $w_{t_i}^1$ and $w_{t_i}^2$ in the sequences of situations $[w_{t_0}^1, \dots, w_{t_n}^1]$ prior to $w_{t_1}^1$ in b^1 and $[w_{t_0}^2, \dots, w_{t_n}^2]$ prior to $w_{t_2}^2$ in b^2 are equivalent (allowing here the possibility that both sequences are empty).*

We can now state the constraints on the sets of BDI interpretations \mathcal{R}_π that ensure that the program construction operators respect their operational definitions.

$$\begin{aligned} \mathcal{R}_{\pi;\chi} &= \mathcal{R}_\pi \oplus \mathcal{R}_\chi \\ \mathcal{R}_{\pi \cup \chi} &= \mathcal{R}_\pi \uplus \mathcal{R}_\chi \\ \mathcal{R}_{\pi^*} &= \mathcal{R}_\pi^* \text{ (the reflexive transitive closure of } \mathcal{R}_\pi \text{ under } \oplus) \\ \mathcal{R}_{\alpha?} &= \{b : b \in B_1, b^* \models \mathbf{B}\alpha \text{ and } b^* \text{ is isomorphic to a world in } \mathcal{R}_{\text{env}}\} \\ \mathcal{R}_{\neg\alpha?} &= \{b : b \in B_1, b^* \not\models \mathbf{B}\alpha \text{ and } b^* \text{ is isomorphic to a world in } \mathcal{R}_{\text{env}}\} \end{aligned}$$

Here B_1 is the set of reduced BDI interpretations all of whose worlds are of depth 1. Note that test actions in the PRS-like architecture consume one cycle during which the environment may change. Hence the test $\alpha?$ ($-\alpha?$) succeeds only when the agent believes (does not believe) α , but the execution context, as captured in the non-final situations in the model, are those resulting from the changes in the environment, as reflected in \mathcal{R}_{env} .

The relations \mathcal{R}_π for attempts are also subject to the following constraints.

$$\begin{aligned} \mathcal{R}_{attempt \ \Lambda} &= \mathcal{R}_\Lambda, \mathcal{R}_{attempt \ \alpha?} = \mathcal{R}_{\alpha?}, \mathcal{R}_{attempt \ -\alpha?} = \mathcal{R}_{-\alpha?} \\ \mathcal{R}_{attempt \ \pi; \chi} &= \mathcal{R}_{attempt \ \pi; attempt \ \chi} \\ \mathcal{R}_{attempt \ \pi \cup \chi} &= \mathcal{R}_{attempt \ \pi \cup attempt \ \chi} \\ \mathcal{R}_{attempt \ \pi^*} &= \mathcal{R}_{(attempt \ \pi)^*} \\ \mathcal{R}_{attempt \ achieve \ \gamma} &= \biguplus \{ \mathcal{R}_\pi : \pi \in L, post(\pi) \vdash \gamma \} \end{aligned}$$

As above, $post(\pi)$ is the postcondition of a plan π in the plan library L . The last constraint formally captures the intuition that execution of a plan in the plan library counts as an attempt to achieve its postcondition.

The program construction operators, sequencing, alternation and iteration, are modelled as operations on sets of BDI interpretations. The operation for sequencing is a kind of “concatenation” of worlds, denoted \oplus , analogous to concatenation of computation sequences. For alternation, we utilize an operation, denoted \uplus , that merges two worlds if they have equivalent initial situations.

Definition 12. *Let w^1 and w^2 be worlds (in reduced BDI interpretations) over time trees $\langle \mathcal{T}_1, \prec_1 \rangle$ and $\langle \mathcal{T}_2, \prec_2 \rangle$. Let S be the set of end points of \mathcal{T}_1 , and let S' be the subset of elements s of S for which w_s^1 is a non-final situation and equivalent to w_r^2 , where r is the root of \mathcal{T}_2 . For each element s of S' , let $w^2(s)$ be a world isomorphic to W_2 over a time tree $\langle \mathcal{T}_2^s, \prec_2^s \rangle$, whose accessibility relations are the same as w^2 on corresponding elements. Then the concatenation of w^1 and w^2 , denoted $w^1 \oplus w^2$ is defined over a tree consisting of $\mathcal{T}_1 - S'$ and all the sets \mathcal{T}_2^s with a precedence ordering \prec extending \prec_1 and all the \prec_2^s by also defining $t_1 \prec t_2$ if $t_1 \in \mathcal{T}_1 - S'$, $t_1 \prec_1 s$ and $t_2 \in \mathcal{T}_2^s$. The non-final situations of $w^1 \oplus w^2$ are defined to be those of all the $w^2(s)$ (there are no non-final situations if S' is empty).*

Definition 13. *Let w^1 and w^2 be worlds (in reduced BDI interpretations) over time trees $\langle \mathcal{T}_1, \prec_1 \rangle$ and $\langle \mathcal{T}_2, \prec_2 \rangle$ with roots r_1 and r_2 , such that $w_{r_1}^1$ and $w_{r_2}^2$ are equivalent. Let the tree \mathcal{T} be defined as the set of time points $\mathcal{T}_1 \cup \mathcal{T}_2$ in which r_1 and r_2 are identified, and with \prec defined as $\prec_1 \cup \prec_2$ (so that the identified r_1 and r_2 is the root of \mathcal{T} , and the children of this node are the children of r_1 from \mathcal{T}_1 and of r_2 from \mathcal{T}_2). Then the merger of w^1 and w^2 , denoted $w^1 \uplus w^2$, is the world defined over the tree $\langle \mathcal{T}, \prec \rangle$ that is inherited from w^1 and w^2 , i.e. $(w^1 \uplus w^2)(t)$ is $w^1(t)$ if $t \in \mathcal{T}_1$ and is $w^2(t)$ if $t \in \mathcal{T}_2$. The non-final situations of $w^1 \uplus w^2$ are defined to be those of w^1 and w^2 .*

4 Logical Properties of Attempts and Belief Update

We now describe some logical properties of action attempts and the logic of belief update implied by the above semantics for ADL. The logic by no means provides a complete axiomatization of these properties.

The first set of properties derive directly from the semantics for attempts.

- (A1) $[attempt \pi]\alpha \Rightarrow [\pi]\alpha$
- (A2) $[attempt \beta?]\alpha \Leftrightarrow [\beta?]\alpha$
- (A3) $[attempt \pi; \chi]\alpha \Leftrightarrow [attempt \pi; attempt \chi]\alpha$
- (A4) $[attempt \pi \cup \chi]\alpha \Leftrightarrow [attempt \pi \cup attempt \chi]\alpha$
- (A5) $[attempt \pi^*]\alpha \Leftrightarrow [(attempt \pi)^*]\alpha$

Most of these are straightforward once it is noted that actions within square brackets are understood to be performed successfully. (A1) follows from the constraint that the models of π are all submodels of those of $attempt \pi$. (A2) says that all attempts to test the agents own beliefs are successful (the agent has accurate introspection abilities). (A3)–(A5) decompose complex attempts into constituent components.

A second pair of properties relate to attempts and achievement goals.

- (B1) $[achieve \gamma]\gamma$
- (B2) $[attempt achieve \gamma]\alpha \Leftrightarrow [\pi_1 \cup \dots \cup \pi_n]\alpha$ where the π_i are the plans whose postcondition logically implies γ

(B1) is a definitive property of achievement goals which should hold in any theory: it basically states that successful executions of $achieve \gamma$ actually achieve γ . However, the point of (B2) is that not all attempts to achieve γ are guaranteed to achieve γ , even those arising from successful executions of a plan whose postcondition implies γ , i.e. (B2) with $achieve \gamma$ instead of $attempt achieve \gamma$ is incorrect. So without a notion of attempts the correct axiom cannot be expressed.

A third set of properties capture simple properties of belief update.

- (C1) $\mathbf{BA} \circ \alpha \Leftrightarrow \mathbf{A} \circ \mathbf{B}\alpha$
- (C2) $[observe \alpha]\mathbf{B}\alpha$
- (C3) $[observe (\beta_1 \wedge \beta_2)]\alpha \Leftrightarrow ([observe \beta_1]\alpha \wedge [observe \beta_2]\alpha)$
- (C4) $[attempt achieve \gamma](\mathbf{B}\gamma \vee \mathbf{B}\neg\gamma)$

(C1) is a definitive property required for modelling belief update using the semantics of action, stating that the set of situations forming the epistemic alternatives of the agent after performing some action are exactly those resulting situations derived from executing the appropriate actions (the action itself or the observe actions) starting from each element of the set of initial epistemic alternatives. (C2) is valid only because we have assumed the agent's observations are veridical and because of a technicality of ADL that an observation of α can succeed only if α is true. (C3) is a simple belief revision property that holds because β_1 and β_2 are restricted to conjunctions of literals, making belief update on complete theories deterministic. (C4) captures the idea that the agent observes the postcondition of an action after attempted execution.

5 Conclusion

In this paper, we extended earlier work on modelling the mental states of PRS-like agents by considering the dynamics of belief. This has required addressing a series of related questions, including the correct modelling of action attempts (in turn both handling different types of success and failure and accommodating the present-directed and future-directed aspects of intention), and capturing the distinctions between two types of “possible” world, those that represent the ignorance of the agent about which world is the actual world, and those that reflect the agent’s ignorance of the state of the world given that it is in some actual world. The resulting theory enables the dynamics of belief to be modelled using the semantics of action, and allows properties of belief update, attempts and observations to be represented in Agent Dynamic Logic.

Acknowledgements

This work is funded by an Australian Research Council Discovery Project Grant. Discussions with Krystian Ji have helped greatly in clarifying the main issues addressed in this paper. Thanks also to the Decision Systems Laboratory at the University of Wollongong for hosting a seminar on an earlier version of this work.

References

1. Bratman, M.E. (1987) *Intention, Plans and Practical Reason*. Harvard University Press, Cambridge, MA.
2. Cohen, P.R. & Levesque, H.J. (1990) ‘Intention is Choice with Commitment.’ *Artificial Intelligence*, **42**, 213–261.
3. Emerson, E.A. & Clarke, E.M. (1982) ‘Using Branching Time Temporal Logic to Synthesize Synchronization Skeletons.’ *Science of Computer Programming*, **2**, 241–266.
4. Georgeff, M.P. & Lansky, A.L. (1987) ‘Reactive Reasoning and Planning.’ *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, 677–682.
5. Harel, D. (1979) *First-Order Dynamic Logic*. Springer-Verlag, Berlin.
6. Pratt, V.R. (1976) ‘Semantical Considerations on Floyd-Hoare Logic.’ *Proceedings of the Seventeenth IEEE Symposium on Foundations of Computer Science*, 109–121.
7. Rao, A.S. & Georgeff, M.P. (1991) ‘Modeling Rational Agents within a BDI-Architecture.’ *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR’91)*, 473–484.
8. van der Hoek, W., van Linder, B. & Meyer, J.-J.Ch. (1999) ‘An Integrated Modal Approach to Rational Agents.’ in Wooldridge, M. & Rao, A. (Eds) *Foundations of Rational Agency*. Kluwer, Dordrecht.
9. Wobcke, W.R. (2002) ‘Modelling PRS-like Agents’ Mental States.’ in Ishizuka, M. & Sattar, A. (Eds) *PRICAI 2002: Trends in Artificial Intelligence*. Springer-Verlag, Berlin.
10. Wooldridge, M.J. (2000) *Reasoning About Rational Agents*. MIT Press, Cambridge, MA.