

Intention and Rationality for PRS-Like Agents

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 elaborate on our earlier work on modelling the mental states of PRS-like agents by considering the logical properties of intentions and various rationality postulates for PRS-like agents that are a consequence of this modelling, in particular Bratman's asymmetry thesis, the side effect problem for intentions and Rao and Georgeff's non-transference principles. We show that PRS-like agents do enjoy many of the accepted logical properties of intention and rationality postulates as automatic consequences of our approach, though, as we have argued before, PRS-like agents do not have intentions in the sense required by Bratman's theory.

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 [4], Pratt's Propositional Dynamic Logic [8] and Rao and Georgeff's BDI Logic [9]. 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 [6], Georgeff and Ingrand [5], as well as derivative architectures such as UM-PRS, C-PRS, AgentSpeak(L), JACK Intelligent AgentsTM and JAM.

In Wobcke [12], we defined an abstract PRS-like architecture extending that of Rao and Georgeff [11] that aimed to capture some of the essential elements common to this family of architectures. In this paper, we focus on the logical properties of (agents and programs developed for) this architecture. We consider two aspects of agency, the logical properties of intention, and the postulates of rationality developed by Rao and Georgeff [9,10], who followed the approach of Cohen and Levesque [3] in aiming to capture formally some of the principles of rationality proposed by Bratman [1]. In particular, we examine Bratman's asymmetry thesis, the side effect problem for intentions, and Rao and Georgeff's non-transference principles. We show that PRS-like agents do enjoy many of the accepted properties of intention and rationality postulates as consequences of our approach, though, as we have argued before, PRS-like agents have only a

limited notion of intention that does not fulfil the functional roles required of Bratman’s theory of intention. The technical aspects of this work derive from our earlier work on Agent Dynamic Logic, Wobcke [13], for formalizing beliefs, desires and intentions, and for defining the semantics of agent programs.

The paper is organized as follows. After a summary of our previous work on Agent Dynamic Logic, we consider in Section 3 the logic of intention, and in Section 4 the rationality postulates.

2 Agent Dynamic Logic

In this section, we summarize our approach to defining the semantics of PRS-like agents’ mental states; see Wobcke [13] for more explanation. Our 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).¹ ADL models capture the agent’s beliefs, desires and intentions as viewed from a single point in time, but not the changes in the agent’s mental state that may result as time progresses and the agent is forced to change beliefs, reconsider goals and possibly abandon intentions. In this respect, it is similar to other approaches to modelling BDI agents using temporal structures, e.g. Cohen and Levesque [3], Rao and Georgeff [9].

The main technical problem addressed in ADL is the interpretation of agent programs as computations over branching time structures (with emphasis on defining the meaning of special actions *achieve* γ , which correspond to subgoals in the plan language). The interpretation of an agent program is formally a set of BDI interpretations, each such BDI interpretation representing the ways a program can be executed starting from a particular situation. Following Rao and Georgeff [9], each BDI interpretation is essentially a set of worlds, where a *world* is a computation tree (each node in such a tree having an associated state), together with accessibility relations capturing the agent’s beliefs, desires and intentions. It is necessary to use BDI interpretations here (rather than just computation trees) because PRS-like agent programs implicitly refer, in the tests in conditional and iterative statements, to the beliefs of the agent, so the formal modelling of the agent’s actions must also include explicit reference to the beliefs of the agent.

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

¹ Note that Rao and Georgeff’s logic was based on CTL* rather than CTL. We use CTL for simplicity.

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 begin with a summary of the CTL definitions.

Definition 1. *The CTL path formulae are defined as follows. If α and β are state formulae, then $\Box\alpha$ (henceforth α holds), $\Diamond\alpha$ (eventually α holds), $\bigcirc\alpha$ (α holds in the next state) and $\alpha\mathcal{U}\beta$ (eventually β holds and α holds until β holds) are path formulae.*

Definition 2. *The CTL state formulae are defined as follows. First, any formula of the base propositional language \mathcal{L} is a state formula. Second, if α is a path formula, then $A\alpha$ (α holds on all paths) and $E\alpha$ (α holds on some path) are state formulae.*

Definition 3. *A CTL time tree $\langle\mathcal{T}, \prec\rangle$ is a nonempty set of time points \mathcal{T} and a binary relation \prec on \mathcal{T} that is irreflexive, transitive, discrete, serial, backwards linear and rooted.*

Definition 4. *Let \mathcal{S} be a nonempty set of states. A world w over a CTL time tree $\langle\mathcal{T}, \prec\rangle$ based on \mathcal{S} is a function on \mathcal{T} giving a state $w_t \in \mathcal{S}$ for each time point $t \in \mathcal{T}$ (in the context of a particular world w , w_t is called a situation). For convenience, say that time points and paths in \mathcal{T} are also time points and paths in w .*

The semantics of CTL state and path formulae can be given with respect to (time points and paths in) worlds using the following definitions.

Definition 5. *For any point t in a CTL time tree $\langle\mathcal{T}, \prec\rangle$, the subtree of $\langle\mathcal{T}, \prec\rangle$ generated from t , denoted $\langle\mathcal{T}_t, \prec_t\rangle$, is defined to be that subtree of $\langle\mathcal{T}, \prec\rangle$ consisting of the set of points $\mathcal{T}_t = \{u \in \mathcal{T} : t \preceq u\}$, where \prec_t is defined as \prec restricted to \mathcal{T}_t , and $t \preceq u$ is an abbreviation for $t \prec u$ or $t = u$.*

Definition 6. *Let $\langle\mathcal{T}, \prec\rangle$ be a CTL time tree, let p be a path in \mathcal{T} and let t be a time point in p . The successor $s_p(t)$ of t in p is that state $u \in p$ for which $t \prec u$ but there is no $v \in p$ with $t \prec v \prec u$.*

Definition 7. *Let \mathcal{S} be a set of states and let \mathcal{L} be a language for expressing time-independent properties of states (we assume there is a satisfaction relation \models between states and formulae of \mathcal{L}). Let w be a world over a CTL time tree $\langle\mathcal{T}, \prec\rangle$ based on \mathcal{S} . Then w satisfies a CTL formula at a time point t in \mathcal{T} as follows.*

$w \models_t \alpha$	if $w_t \models \alpha$, for α a formula of \mathcal{L}
$w \models_t A\alpha$	if $w \models_p \alpha$ for every path p in $\langle\mathcal{T}_t, \prec_t\rangle$
$w \models_t E\alpha$	if $w \models_p \alpha$ for some path p in $\langle\mathcal{T}_t, \prec_t\rangle$
$w \models_p \Box\alpha$	if $w \models_u \alpha$ for every $u \in p$
$w \models_p \Diamond\alpha$	if $w \models_u \alpha$ for some $u \in p$
$w \models_p \bigcirc\alpha$	if $w \models_{s_p(t)} \alpha$
$w \models_p \alpha\mathcal{U}\beta$	if there is $u \in p$ with $w \models_u \beta$, and $w \models_v \alpha$ for every $v \in p$ with $v \prec u$

To enable CTL to be used for modelling rational agents, Rao and Georgeff [9] developed a logic we call BDI-CTL, which extended the language of CTL state formulae to include modal operators for modelling beliefs, desires (goals) and intentions. In our reformulation, the modal operators **B** and **G** apply to BDI-CTL state formulae, while the modal operator **I** applies to programs. Thus intentions are directed towards actions represented as programs. The basic semantic notion is a BDI interpretation, which in Rao and Georgeff's framework, is a set of worlds over the subtrees of a single time tree, where this time tree is a branching time structure as described above, and each "world" consists of an assignment of a state to each time point in the tree over which it is based. The formal definitions of BDI interpretations and of the accessibility relations rely on the notion of a subworld of a world. A subworld of a world w contains, for each time point t , a subset of the possible futures of t that w admits, according to whether or not the corresponding path is contained in the subtree from which the subworld is derived.

Definition 8. A BDI interpretation is a tuple $\langle \mathcal{T}, \prec, \mathcal{S}, \mathcal{W}, \mathcal{B}, \mathcal{I} \rangle$ 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{B} is a subset of $\mathcal{W} \times \mathcal{T} \times \mathcal{W}$ defined only for tuples (w, t, w') for which t is a time point in w and w' , 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 9. 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 .

Any particular world $w \in \mathcal{W}$ defines a set of futures the agent considers possible starting from the initial situation in that world. Since each world is based on a branching time structure, the actions are modelled as being non-deterministic. However, since the agent has only partial information about its environment, it may be ignorant of exactly which situation it is embedded in; the agent considers only that it is embedded in a situation in which its current beliefs are true. This ignorance is captured using the relation \mathcal{B} . 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 $\mathcal{B}(w, t, w')$. The relation on situations is assumed to be serial, transitive and Euclidean, and moreover, the relations \mathcal{B} and \mathcal{I} are assumed to satisfy the condition that $\mathcal{B}(w, t, w')$ iff $\mathcal{B}(\mathcal{I}(w, t), t, \mathcal{I}(w', t))$, i.e. the epistemic alternatives of $\mathcal{I}(w, t)$ at t are the intended subworlds of the epistemic alternatives of w at t . This simple relationship between \mathcal{B} and \mathcal{I} is used below to derive some basic properties of the relation between beliefs and intentions.

The intentions of the agent are those actions the agent considers that eventually it will successfully perform, according its current view of the environment, i.e. not taking into account the potential for the environment to change so as to force the agent to revise or abandon its intentions. Intentions are modelled

using the function \mathcal{I} , which defines, for any particular world, which futures (sub-worlds) in that world are intended by the agent: in the “intended” futures, all the agent’s actions are performed successfully. More formally, the intentions of an agent with respect to a world w at t are those action formulae π for which on all possible futures in $\mathcal{I}(w, t)$, the agent eventually does π , i.e. the intentions of an agent are represented by formulae of the form $\mathbf{BA}\Diamond do(\pi)$ (for the purposes of exposition, we defer the truth definition for $do(\pi)$). Finally, the goals of the agent are simply those formulae γ such that the agent intends to perform the action *achieve* γ . This is in keeping with the behaviour of the PRS-like interpreter, which has no mechanism for adopting a goal independently of a plan to achieve that goal.

Definition 10. *Let $\langle \mathcal{T}, \prec, \mathcal{S}, \mathcal{W}, \mathcal{B}, \mathcal{I} \rangle$ 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.*

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

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 [7]. If α is a BDI-CTL state formula, $[\pi]\alpha$ is understood to mean that α holds in all terminating situations arising from an execution of π , while $\langle \pi \rangle\alpha$, which is defined as $\neg[\pi]\neg\alpha$, is understood to mean that α holds in one terminating execution of π . In computation tree semantics, terminating situations are simply the leaf nodes of the trees (note that infinite branches correspond to programs that do not terminate). 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 semantics of ADL is based on what we call *PRS interpretations*, which provide an interpretation for each program in terms of a set of BDI interpretations. As mentioned above, the need for BDI interpretations (rather than computation trees or worlds) arises because PRS agent programs refer to the beliefs of the agent in the tests of conditional and iterative statements, so the models of the PRS programs must also include reference to the agent’s beliefs. The definition of a PRS interpretation also enables us to define the satisfaction conditions for $do(\pi)$ as needed for modelling intentions.

Each program is modelled as a set of BDI interpretations that arises by varying the initial state and belief accessibility relation \mathcal{B} , and in which \mathcal{I} is defined as the identify function (i.e. $\mathcal{I}(w, t) = w$ for all w, t). The function \mathcal{I} plays no role in the interpretation of agent programs (so could be omitted), while the \mathcal{B} relation is used to define the meaning of the test statements. In any BDI interpretation b modelling a program π , one distinguished world b^*

models the execution paths of π 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 by a set of worlds over a computation tree of depth 1, one world for each state $s \in \mathcal{S}$ in which the action is executable. Each such world with root s has one child situation for each possible outcome of executing the action π in s , and each of these situations is non-final. In addition, there is a special “action” *env* that models the changes in the environment that occur over cycles in which the agent tests its beliefs.

Definition 11. 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 BDI interpretations \mathcal{R}_π based on \mathcal{S} , one such set of BDI interpretations for each program π .

Definition 12. Let $\langle \mathcal{S}, \mathcal{R} \rangle$ be a PRS interpretation and $\langle \mathcal{T}, \prec, \mathcal{S}, \mathcal{W}, \mathcal{B}, \mathcal{I} \rangle$ be a 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 the state formula $do(\pi)$ and $[\pi]\alpha$ at a time point t in w as follows.

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

Definition 13. 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 14. A world w^1 in a BDI interpretation $\langle \mathcal{T}_1, \prec_1, \mathcal{S}, \mathcal{W}_1, \mathcal{B}_1, \mathcal{I}_1 \rangle$ is isomorphic to a world w^2 in a BDI interpretation $\langle \mathcal{T}_2, \prec_2, \mathcal{S}, \mathcal{W}_2, \mathcal{B}_2, \mathcal{I}_2 \rangle$ 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 15. A situation w_t in a BDI interpretation $\langle \mathcal{T}_1, \prec_1, \mathcal{S}, \mathcal{W}_1, \mathcal{B}_1, \mathcal{I}_1 \rangle$ is equivalent to a situation w'_t in a BDI interpretation $\langle \mathcal{T}_2, \prec_2, \mathcal{S}, \mathcal{W}_2, \mathcal{B}_2, \mathcal{I}_2 \rangle$ if $w_t = w'_t$ (i.e. they are equal as states) and $\{u_t : \mathcal{B}_1(w, t, u)\} = \{v'_t : \mathcal{B}_2(w', t', v)\}$ (i.e. they satisfy the same basic beliefs).

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 16. Let w^1 and w^2 be worlds (in 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 17. Let w^1 and w^2 be worlds (in 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 .

If b^1 and b^2 are BDI interpretations, $b^1 \oplus b^2$ is the set of all worlds formed by simultaneously concatenating, for each non-final situation w_s^1 of each world w^1 in b^1 , w^1 and a world $w^2(s)$ in b^2 whose initial situation is equivalent to w_s^1 (if one exists), and $b^1 \uplus b^2$ is that set of worlds obtained by merging all pairs of worlds w^1 and w^2 in $b^1 \cup b^2$ whose initial situations are equivalent.

We can finally give the constraints on the sets of BDI interpretations \mathcal{R}_π that ensure that each respects the operational semantics of the program construction operators.

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

Here B_1 is the set of 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 $\alpha?$ ($\neg\alpha?$) 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} .

Definition 18. The relations \mathcal{R}_π for the empty program Λ and for programs achieve γ corresponding to a subgoal γ are defined as follows.

$$\begin{aligned} \mathcal{R}_\Lambda &= B_0 \\ \mathcal{R}_{achieve \gamma} &= \uplus \{ \mathcal{R}_\pi \cap \{b : b^* \models pre(\pi) \wedge context(\pi)\} : \pi \in L, post(\pi) \vdash \gamma \} \end{aligned}$$

Here B_0 is the set of BDI interpretations all of whose worlds are of depth 0, and $pre(\pi)$, $post(\pi)$ and $context(\pi)$ are the precondition, postcondition and context

(respectively) of a plan π in the plan library L . The precondition and postcondition of Λ are just *true*.

Thus Λ is executable in every situation in every world, and leaves the state unchanged, while *achieve* γ is modelled as a set of BDI interpretations in which γ is achieved by the successful execution of a (hierarchical) plan built from plans in the plan library. Note that this condition is actually a set of recursive definitions of $\mathcal{R}_{achieve\ \gamma}$ for all formulae γ at once; it is recursive because the plans π may include subgoals of the form *achieve* δ (and here, of course, δ may be γ). Thus the meaning of *achieve* γ involves a least fixpoint construction.

3 Logical Properties of Intention for PRS-Like Agents

We now discuss the logic of intention implied by the above semantics for ADL. First note that, with our semantics, there are no special logical requirements for goals and intentions. The set of goals (and of intentions) can even be inconsistent, partly because of the implicit temporal aspects of our modal operators, i.e. the formulae $\mathbf{G}\gamma$ and $\mathbf{G}\neg\gamma$ are not inconsistent – they simply mean that the agent wants to achieve γ and $\neg\gamma$ at some points in the future (as determined by the current plans), though not necessarily at the same point. This differs from the presentation in Rao and Georgeff [9] in which the modal operators \mathbf{B} , \mathbf{G} and \mathbf{I} all modify propositional formulae with the temporal modality explicit, so an agent's goals γ and $\neg\gamma$ are represented using formulae $\mathbf{G}\diamond\gamma$ and $\mathbf{G}\diamond\neg\gamma$.

However, the language of ADL programs is more complicated than modal logic, in including program terms for conditional and iterative constructs, as in PDL. The meaning of these constructs follows the definitions from PDL, in which the *if-then-else* and *while* statements are defined in terms of the sequencing, union, iteration and test primitives as follows.

$$\begin{aligned} \mathbf{if}\ \alpha\ \mathbf{then}\ \pi\ \mathbf{else}\ \chi &\equiv (\alpha?; \pi) \cup (\neg\alpha?; \chi) \\ \mathbf{while}\ \alpha\ \mathbf{do}\ \pi &\equiv (\alpha?; \pi)^*; \neg\alpha? \end{aligned}$$

Note, however, that the conditional and iterative constructs in PRS agent programs behave very differently: the tests in the PRS statements are tests, not on the state of the world, but on the agent's belief state (a distinction not needed for standard PDL).

We can now present some logical properties of intention that are consequences of the ADL semantics.

- (I1) $\models \mathbf{I}(\pi; \chi) \Rightarrow \mathbf{I}\pi \wedge \mathbf{I}\chi$
- (I2) $\not\models \mathbf{I}(\mathbf{if}\ \alpha\ \mathbf{then}\ \pi\ \mathbf{else}\ \chi) \Rightarrow (\mathbf{I}\pi \vee \mathbf{I}\chi)$
- (I3) $\not\models \mathbf{I}\pi \Rightarrow \mathbf{I}(\pi \cup \chi)$
- (I4) $\models \mathbf{I}(\mathbf{achieve}\ \gamma) \Rightarrow \mathbf{I}(\pi_1 \cup \dots \cup \pi_n)$ where the π_i are the plans whose postcondition logically implies γ
- (I5) $\not\models \mathbf{I}(\mathbf{achieve}\ \gamma) \wedge \mathbf{I}(\mathbf{achieve}\ \delta) \Rightarrow \mathbf{I}(\mathbf{achieve}\ (\gamma \wedge \delta))$

With (I1), performing the action $\pi; \chi$ requires the performance of π followed by χ , hence performing both actions. For (I2), performing an *if-then-else* statement requires that the agent perform one or other of the branches of the statement, though *which* branch is determined only by the result of a test on beliefs

that may take place in the future. Hence the agent need not have a commitment (at the present time) to any particular branch. In contrast, (I3) relates to the choice(s) involved in commitment to an intention, and represents the fact that once a choice has been made, the choice no longer exists. Here $\mathbf{I}(\pi \cup \chi)$ represents the performance of either π or χ , but $\mathbf{I}\pi$ represents a commitment to π – the latter does not imply the former. Note, though, that the formula represents a “choice” of the agent only to the extent that which of the actions is executed is partly determined by the situation at execution time. (I4) is a consequence of the semantics for *achieve* actions, and mirrors the property defined by Cavedon and Rao [2], who obtain this property by the use of special “plan” worlds in which the only events are those actions contained in the agent’s plans. Finally, (I5) is a consequence of the future directedness of intention: an agent intending to achieve γ and δ intends that γ and δ eventually hold, but not necessarily at the same time, so need not intend to achieve $\gamma \wedge \delta$.

4 Rationality Postulates for PRS-Like Agents

Rao and Georgeff [10] present a discussion of rationality for BDI agents, building on the approach of Cohen and Levesque [3]. One of their main aims is to capture formally some of the principles of rationality proposed by Bratman [1], in particular, the asymmetry thesis, side effect free principle, and what they call non-transference principles. The *asymmetry thesis* is the thesis that (i) it is inconsistent for a rational agent to both intend to perform some action and believe s/he will not do it, and (ii) it *is* consistent for a rational agent to intend to perform some action whilst not believing s/he will do it (“doing it” here means executing the action successfully, so the thesis means that the agent need not be certain that the attempt to perform the intended action will succeed, but cannot believe it will inevitably fail). The *side effect free principle* is that a rational agent need not intend the believed necessary consequences (side effects) of an intention. The *non-transference principle* for intentions states that a rational agent need not adopt an intention to achieve some condition that it believes will inevitably become true in the future.

More formally, Rao and Georgeff represent these principles using schemas of the following form, adapted to the language of ADL. The adaptation is necessary because in Rao and Georgeff’s formalism, intentions are directed towards propositions (intentions that α be true), rather than towards actions (intentions to perform π).

- (A1) $\models \mathbf{I}\pi \Rightarrow \neg \mathbf{B}\mathbf{A}\Box \neg do(\pi)$
- (A2) $\not\models \mathbf{I}\pi \Rightarrow \mathbf{B}\mathbf{A}\Diamond do(\pi)$
- (SE) $\not\models \mathbf{I}(achieve \ \gamma) \wedge \mathbf{B}\mathbf{A}\Box(\gamma \Rightarrow \delta) \Rightarrow \mathbf{I}(achieve \ \delta)$
- (NT) $\not\models \mathbf{B}\mathbf{A}\Diamond \gamma \Rightarrow \mathbf{I}(achieve \ \gamma)$

Note that there are different varieties of (SE) depending on the “strength” of belief in the connection between γ and δ (the postulate above shows the strongest connection, and hence the hardest to refute).²

² In ADL, the even stronger side effect free principle with $\gamma \vdash \delta$ is also invalid.

In the ADL modelling of PRS-like agents, the only intentions are those relating to actions executed by the agent, though external events are also assumed to occur. Hence (NT) is satisfied, since a condition γ that always eventually arises need not be related to a plan of the agent. Now consider the asymmetry thesis. The crucial difference between Rao and Georgeff's semantics and ADL semantics is the condition that the intended futures in some world are a subworld of that world. Hence the agent can believe there are possible futures in which its intentions are not fulfilled, but cannot believe its actions will always fail. This means that both (A1) and (A2) are satisfied automatically. For (SE), Rao and Georgeff require there to be belief-accessible worlds that are not intention worlds; thus, for example, there is an intention world in which the agent goes to the dentist to relieve a toothache and does not feel pain, even though in all belief-accessible worlds the agent feels pain in this circumstance. In contrast, our explanation for (SE) rests on the behaviour of the PRS-like interpreter in selecting actions: the agent can select a plan (and hence form an intention) to achieve δ only when the postcondition of the plan logically implies δ . Hence it is possible for an agent to intend to execute a plan whose postcondition implies γ but not δ , e.g. the goal of relieving the toothache but not the goal of feeling pain. Thus (SE) is also satisfied in ADL.

The above postulates express relationships between beliefs and intentions, following Bratman [1]; Rao and Georgeff also introduce (without justifying argument) similar postulates relating beliefs and goals, and goals and intentions. Due to the way intentions and goals are defined in ADL, postulates relating beliefs and goals follow from the above postulates, while postulates relating goals and intentions are not needed. For example, the following postulates relating beliefs and goals follow from the definition of $\mathbf{G}\gamma$ as $\mathbf{I}(\text{achieve } \gamma)$. Again, all are satisfied in ADL.

$$\begin{aligned} \text{(A1)} & \models \mathbf{G}\gamma \Rightarrow \neg\mathbf{BA}\Box\neg\text{do}(\text{achieve } \gamma) \\ \text{(A2)} & \not\models \mathbf{G}\gamma \Rightarrow \mathbf{BA}\Diamond\text{do}(\text{achieve } \gamma) \\ \text{(SE)} & \not\models \mathbf{G}\gamma \wedge \mathbf{BA}\Box(\gamma \Rightarrow \delta) \Rightarrow \mathbf{G}\delta \\ \text{(NT)} & \not\models \mathbf{BA}\Diamond\gamma \Rightarrow \mathbf{G}\gamma \end{aligned}$$

Rao and Georgeff [9] present the following axiom schemes as the “basic system” for modelling BDI agents, reconstructed in ADL below: where our reconstruction differs significantly, the original version is also given.

$$\begin{aligned} \text{(AI1)} & \mathbf{G}\gamma \Rightarrow \mathbf{BE}\Diamond\text{do}(\text{achieve } \gamma) & \{\mathbf{G}\alpha \Rightarrow \mathbf{B}\alpha \text{ for } \alpha \text{ an O-formula}\} \\ \text{(AI2)} & \mathbf{I}(\text{achieve } \gamma) \Rightarrow \mathbf{G}\gamma & \{\mathbf{I}\alpha \Rightarrow \mathbf{G}\alpha\} \\ \text{(AI3)} & \mathbf{I}\pi \Rightarrow \text{do}(\pi) & \{\mathbf{I}\text{do}(e) \Rightarrow \text{do}(e)\} \\ \text{(AI4)} & \mathbf{I}\pi \Rightarrow \mathbf{BI}\pi \\ \text{(AI5)} & \mathbf{G}\gamma \Rightarrow \mathbf{BG}\gamma \\ \text{(AI6)} & \mathbf{I}\pi \Rightarrow \mathbf{GI}\pi \\ \text{(AI7)} & [\pi]\alpha \Rightarrow [\pi]\mathbf{B}\alpha & \{\text{done}(e) \Rightarrow \mathbf{B}(\text{done}(e))\} \\ \text{(AI8)} & \mathbf{I}\pi \Rightarrow \mathbf{BA}\Diamond\neg\mathbf{I}\pi & \{\mathbf{I}\alpha \Rightarrow \mathbf{A}\Diamond\neg\mathbf{I}\alpha\} \end{aligned}$$

For (AI1), an *O-formula* is defined to be a formula with no positive occurrences of A or negative occurrence of E outside the scope of a \mathbf{B} , \mathbf{G} or \mathbf{I} operator. In (AI3) and (AI7), e denotes a primitive action.

(AI1) is meant to capture the “realism” condition, c.f. Cohen and Levesque [3], that an agent does not adopt goals it believes are unachievable. Rao and Georgeff consider the special case of (AI1) with α the O-formula $E\Diamond p$, giving the reading that if the agent has the goal of possibly eventually achieving p , then it believes that it *is* possible to eventually achieve p . In our reconstruction, the “possible future” aspect of the goal is built into the meaning of the modal operator \mathbf{G} , so that this special case of (AI1) follows directly from the definitions. Similarly, for (AI2), we have defined goals so that $\mathbf{I}(\textit{achieve } \gamma) \Leftrightarrow \mathbf{G}\gamma$ is valid (there is no mechanism for a PRS-like agent to have a goal γ without a corresponding intention to achieve γ).

On Rao and Georgeff’s description, (AI3) means that if the agent has an intention to do a primitive action e then the agent does e , though this does not mean that e is done successfully, only that e is attempted by the agent. Thus in the very special case where the agent has only the intention to perform one primitive action at the current time point, that intention is adopted and the action attempted. However, $\mathbf{I}\pi \Rightarrow do(\pi)$ is not valid in ADL because of the future directedness of intention built in to the definition; an agent’s having an intention to perform π implies only that the agent believes it will attempt π at some future time.

The validity of (AI4) and (AI5) in ADL follow from the restriction imposed on the relations \mathcal{B} and \mathcal{I} , the condition that $\mathcal{B}(w, t, w')$ iff $\mathcal{B}(\mathcal{I}(w, t), t, \mathcal{I}(w', t))$. Indeed the converses of (AI4) and (AI5) also follow from this condition, i.e. the agent has correct *and complete* beliefs about its goals and intentions. However, (AI6) is not well formed in ADL: on our interpretation, an intention to perform π gives rise to a goal of achieving the postcondition of π , not a goal of intending π itself. As goals are future directed, the goal to intend π (at some time in the future) is quite different from the (current) intention to perform π .

(AI7) is similar in flavour to (AI3) in that it relies on a distinction between attempted and performed actions. Rao and Georgeff’s interpretation of (AI7) is that after attempting a primitive action e , the agent believes that it has attempted e . Our reconstruction preserves the idea of the agent’s awareness of its own actions. However, our version of (AI7) is not intuitively valid, first because there is no way in ADL to distinguish an agent’s attempted actions from the actions it performs, and second because the agent may not observe all the consequences of its actions, so not believe an action has been performed successfully even when it was performed successfully.

Finally, (AI8), which Rao and Georgeff take to mean that an agent eventually acts on an intention or else abandons that intention, does not hold for PRS-like agents. One reason for this is that this interpretation needs modification when plans can be infinite, because an intention, even though being acted on, may never be fulfilled, even in principle. A more serious reason is that it is part of the operational definition of the PRS-like interpreter that the agent always acts on a plan that has maximal value, so a low-value plan may never be activated.

In summary, (AI1), (AI2), (AI4) and (AI5) follow from the ADL definitions, while the other properties are intuitively invalid for PRS-like agents.

5 Conclusion

In this paper, we extended earlier work on modelling the mental states of PRS-like agents by considering the logical properties of intentions and various rationality postulates for PRS-like agents that are a consequence of our modelling, in particular Bratman's asymmetry thesis, the side effect problem for intentions and Rao and Georgeff's non-transference principles. We showed that PRS-like agents do enjoy many of the accepted logical properties of intention and rationality postulates as automatic consequences of our approach, even though PRS-like agents embody only a weak notion of intention.

References

1. Bratman, M.E. (1987) *Intention, Plans and Practical Reason*. Harvard University Press, Cambridge, MA.
2. Cavedon, L. & Rao, A.S. (1996) 'Bringing About Rationality: Incorporating Plans Into a BDI Agent Architecture.' in Foo, N.Y. & Goebel, R.G. (Eds) *PRICAI'96: Topics in Artificial Intelligence*. Springer-Verlag, Berlin.
3. Cohen, P.R. & Levesque, H.J. (1990) 'Intention is Choice with Commitment.' *Artificial Intelligence*, **42**, 213–261.
4. Emerson, E.A. & Clarke, E.M. (1982) 'Using Branching Time Temporal Logic to Synthesize Synchronization Skeletons.' *Science of Computer Programming*, **2**, 241–266.
5. Georgeff, M.P. & Ingrand, F.F. (1989) 'Decision-Making in an Embedded Reasoning System.' *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 972–978.
6. Georgeff, M.P. & Lansky, A.L. (1987) 'Reactive Reasoning and Planning.' *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, 677–682.
7. Harel, D. (1979) *First-Order Dynamic Logic*. Springer-Verlag, Berlin.
8. Pratt, V.R. (1976) 'Semantical Considerations on Floyd-Hoare Logic.' *Proceedings of the Seventeenth IEEE Symposium on Foundations of Computer Science*, 109–121.
9. 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.
10. Rao, A.S. & Georgeff, M.P. (1991) 'Asymmetry Thesis and Side-Effect Problems in Linear-Time and Branching-Time Intention Logics.' *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, 498–504.
11. Rao, A.S. & Georgeff, M.P. (1992) 'An Abstract Architecture for Rational Agents.' *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning (KR'92)*, 439–449.
12. Wobcke, W.R. (2001) 'An Operational Semantics for a PRS-like Agent Architecture.' in Stumptner, M., Corbett, D. & Brooks, M. (Eds) *AI 2001: Advances in Artificial Intelligence*. Springer-Verlag, Berlin.
13. 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.