

An Analysis of Three Puzzles in the Logic of Intention

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 generalize our formal approach to modelling PRS agents away from PRS-specific assumptions to more general theories of rationality, while not losing the concreteness of the connection between an agent's mental states and formal models, using three "puzzles" in the logic of intention to motivate our extended approach. We show how the theory can be used to represent solutions to the puzzles and draw out insights into how agent architectures may be extended to handle these more complex scenarios.

1 Introduction

In previous work, we developed formal models of a class of BDI agent architectures based on PRS, Georgeff and Lansky [7]. We have developed an operational semantics aiming to succinctly abstract the main properties of this family of architectures [16], models of the "mental states" of agents based on this architecture [17], a formal derivation of some "rationality postulates" of those agents [18], extended the account to incorporate belief update and attempts [19], and examined model generation algorithms for use in model checking for agents of this class [20]. The theory is based on a logic, Agent Dynamic Logic (ADL), that combines elements from Computation Tree Logic [6], Propositional Dynamic Logic [13], and BDI Logic [14]. The motivation of all this work is to develop a logical framework that is at once rigorous in providing formal notions of belief, desire and intention, yet which is also closely enough aligned to the operational behaviour of this architecture to enable formal reasoning about the behaviour of agents in this class.

Some aspects of this prior work are specific to the PRS architecture, and do not extend to more general theories of rational agency. This has been deliberate, in order to capture the behaviour of the PRS family in as concise a manner as possible. Two major assumptions peculiar to PRS are (i) that the agent attempts to execute only one action at a time, and (ii) that the agent selects a plan to fulfil an achievement goal only at the time of executing that plan (at the latest possible moment). These assumptions reflect more the simplicity of PRS and/or the characteristics of the environments in which PRS agents are expected to operate, rather than general principles of rationality.

The main objective of this paper is to generalize the formal theory away from these PRS-specific assumptions, while not losing the concreteness of the connection between an agent's mental states and formal models. To do this, we use as motivation three puzzles in the logic of intention and action, two discussed extensively by Bratman [2]

(a “Video Game” example and a “Strategic Bomber” example), and one presented here derived from the Lottery Paradox. These examples highlight some of the complexities of Bratman’s account which have not been captured in previous formal models of rational agency. Our work is a step towards capturing these more complex properties.

One lesson from developing the formal models used in ADL is that the theory of action modelling the dynamics of the environment must be clearly distinguished from the beliefs of the agent, whereas in approaches such as Cohen and Levesque [4], the theory of action is part of the agent’s beliefs. For PRS, these must be kept distinct because the PRS agent does no reasoning about actions and plans except insofar as it selects a plan from a given plan library to achieve some goal. Thus the theory of action *describes* the agent’s behaviour, but is not part of the agent’s explicit beliefs. Bratman, however, considers highly sophisticated rational agents who know their own theory of action; nevertheless, this theory still needs to be distinguished from the beliefs of an agent about the environment during the execution of some plan. Thus ADL uses two modal operators to capture this distinction, a modal operator **B** for the agent’s beliefs, and a modal operator **A** to capture the theory of action implicit in the formal models.

In ADL, the intentions of the agent are those actions the agent chooses from its plans that eventually it will successfully perform, represented using a formula $\lambda\pi$ for a program term π , and are relative to a *situation*, a state in a branching-time execution structure that includes accessibility relations for beliefs, alternative actual situations, and a subworld relation denoting the successful executions of an action. The language of programs in ADL extends that for PDL in including special actions of the form *achieve* γ where γ is a PC formula, and *attempt* π where π is a program term. The language also includes test statements $\alpha?$ and program terms for conditional and iterative constructs, defined as follows:

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

The semantics of ADL is given in terms of computation trees, extending the approach of Harel [8]. Union is defined in terms of “tree merging”, sequencing in terms of “tree adjoining” and iteration in terms of the transitive closure of the sequencing operation.

The organization of this paper is as follows. We first describe the three “puzzles” in the logic of intention that are addressed in this work, and then provide a critique of earlier formal approaches to these problems (newly identifying shortcomings in that earlier work). We then present definitions of ADL that extend the theory towards more general models of rational agency, and indicate how the three puzzles are handled in the framework. The exercise yields insights into how agent architectures may be extended to handle these more complex scenarios.

2 Three Puzzles in the Logic of Intention

2.1 The “Package Deal” Problem: Strategic Bomber

The “package deal” problem, Bratman [2, Ch. 10], concerns the relationship between choice and intention and the question of whether an agent intends the consequences of its intended actions. The example discussed by Bratman compares a “Strategic Bomber” with a “Terror Bomber”, both engaged in a war with the enemy.

Terror Bomber plans to bomb a school in enemy territory, thereby killing the children, and so to terrorize the enemy population. Strategic Bomber plans to bomb a munitions plant, thereby undermining the enemy's war effort. However, Strategic Bomber knows that the munitions plant is next to a school, and by bombing the munitions plant, he will also kill the children. Strategic Bomber has taken this into consideration in choosing his course of action. Does Strategic Bomber intend to kill the children?

The problem arises because agents, in their reasoning about what to do, choose between alternative "scenarios", sets of actions and their consequences taken into account in their reasoning. So in particular, Strategic Bomber considers and deliberately chooses to kill the children (as part of a larger scenario involving bombing the munitions plant). So shouldn't this mean that Strategic Bomber intends to kill the children? Bratman argues that it does not, essentially because "killing the children" does not play the three characteristic functional roles of intention, e.g. the agent does not pursue means towards killing the children (so would not plan to kill the children, for example, if they were evacuated from the school). The question is how to capture these complex properties of intentions in a formal model.

2.2 Motivational Potential: The "Video Game" Example

Bratman [1] uses the example of a video game to argue against what he calls the "Simple View", the view that if I do *A* intentionally then I intend to *A*. The argument hinges on Bratman's requirement for intentions to be strongly consistent, relative to the agent's beliefs, and hence the example provides support for this requirement.

Consider a video game, played with two hands, whose aim is to hit one of two targets. A missile shot using the left hand heads towards one target; one shot using the right towards another. The game ends when one of the targets is hit, and it is impossible (and the agent knows this) to hit both targets simultaneously (in this case, the game shuts down just prior to this happening). A natural strategy is for the agent to continually guide two missiles, one with each hand, each towards the appropriate target, each trying to hit that target. Does the agent intend to hit the targets?

The *strong consistency* requirement on beliefs and intentions is never defined precisely by Bratman; the closest statement approaching a definition is that an agent's intentions are strongly consistent relative to its beliefs when it is 'possible for [its] plans taken together to be successfully executed in a world in which [its] beliefs are true', Bratman [3, p. 19], though Bratman realizes that this is in need of further elaboration, Bratman [2, p. 179, note 3].

Now given this requirement, in the video game example, the agent does not intend to hit either target. Let the targets be target 1 and target 2. Then the two intentions, the intention to hit target 1 and the intention to hit target 2, are not strongly consistent with the belief that it is impossible to hit both target 1 and target 2. The question is that if the agent does not intend to hit target 1 nor intend to hit target 2, what does the agent

intend? Bratman [1] suggests a number of possibilities, such as the agent intends to shoot at the targets, to try to hit the targets, to hit each target if it can, or to hit one of the two targets. The questions for formal modelling are which of these is right and how can the differences between them be captured.

2.3 Rational Belief and Intention: The Lottery Problem

The Lottery Paradox, usually attributed to Kyburg [10], is a problem about the rationality of belief under uncertainty.

Consider a fair lottery with a million tickets and only one winning ticket. An agent with one ticket has a very small chance of winning, hence it is rational for the agent to believe that it will have a losing ticket. However, by similar reasoning, it is rational for the agent to believe that any given ticket is a losing ticket. But the agent also believes that there will be a winning ticket. So it seems that the agent's beliefs, each one of which is rational, are together inconsistent, a clear violation of rationality.

The Lottery Paradox has been extensively discussed in the literature, and it is not possible in this paper to survey all the proposed approaches. One reasonable way out of this paradox, proposed by Pollock [12], is to deny that the agent is rational in believing it has a losing ticket (Pollock treats this as a case of "collective defeat", i.e. the collection of equally supported beliefs that each ticket will not win, together with the belief that one ticket will win, defeats the conclusion that any particular ticket will not win).

The *Lottery Problem* that I want to raise here is: is it possible for the agent to intend to win the lottery, and further, if so, is the agent rational to so intend? After all, if the agent is irrational in believing it will lose, the strong-consistency requirement on intentions and beliefs does not block the (rational) agent's intention to win the lottery. But now the agent will start planning to win the lottery, planning what to do with the money, etc., and this seems to be irrational behaviour. The problem is that Bratman's theory seems to leave no room for the agent to intend and plan on the basis of the likelihoods of achieving the intended outcomes, including extremely improbable outcomes such as winning the lottery, and this, at root, is because beliefs are "flat out", not graded with any degree of certainty. The issue for formal modelling is how to model the agent's beliefs and intentions while respecting the answers given to the above questions.

3 Critique of Previous Approaches

Despite the strong influence of Bratman's work in the BDI agents literature, there have been surprisingly few attempts to develop logics of belief, desire and intention. There are, broadly speaking, two approaches to modelling intention, corresponding roughly to Bratman's "two faces" of intention. According to Bratman [1], intentions are Janus-faced, looking to the present (in controlling action) and to the future (in planning). In their future-directed aspect, actions of the agent are modelled as attempts with no guarantee of success; viewed from the aspect of the (ongoing) present, the world is

modelled as a sequence of states that actually occur, and intentions are incorporated into such models. The approaches of Cohen and Levesque [4] and Wooldridge [21] are of the latter type, taking as a starting point models of actions inspired by computational logic; those of Rao and Georgeff [14,15] and Wobcke [17,19] are more oriented towards the future-directed aspects (so notions of the agent's possible actions and the failure of an agent's attempts are more pertinent in these approaches). However, the main difficulty is to capture both present-directed and future-directed aspects of intention in one formalism.

Perhaps the best known work is that of Cohen and Levesque [4], in which, rather than formalizing intentions directly, intentions are reduced to persistent goals. Persistence is related to Bratman's notion of stability [2], but is much simpler, in referring only to a temporal property. More precisely, an intention towards an action a to achieve a goal p , relative to some condition q , is characterized, with respect to a state in a sequence of states, as a persistent goal that the agent has that a be done and then p hold immediately after believing that a will happen and then p hold. For a goal to be a *persistent goal* (P-GOAL) relative to some condition q , the agent must believe the goal does not currently hold, and have the goal and continue to have the goal until either believing it to be achieved, believing it to be impossible or believing q to be false.

However, because the intentions of the agent are relative to a sequence of states (world), the agent's intentions can only be determined once it is known which world the agent inhabits, as this will determine which goals persist. While in one sense, whether a goal persists should depend on the world (e.g. a goal should be dropped if some adverse circumstance in the world arises), in another sense the intentions of the agent should be able to be related to the agent's mental state at any point in time without reference to a particular sequence of states, which is impossible with Cohen and Levesque's theory as intentions vary from world to world.

The problem can be illustrated with a simple counterexample. Suppose on Monday at noon I decide on the goal of having lunch at the cafe, and I persist in having this goal until I later have lunch at the cafe. Now on Tuesday at noon I decide exactly the same thing, except on my way to the cafe I hear music and instead go to a nearby concert, and have my lunch later at home. On Cohen and Levesque's account, on Monday at noon I had the P-GOAL (hence the intention) to have lunch at the cafe, while on Tuesday at noon, I did not have this P-GOAL and intention (at no stage did I ever think it achieved or infeasible to have lunch at the cafe, so the P-GOAL was only relative to the fact that there was no concert). On the temporal understanding of "persistent goal" this is right: as my behaviour showed, I was not really "committed" on Tuesday to having lunch at the cafe, preferring instead the concert, but the problem is that, intuitively, there is no difference between my mental states at noon on Monday and at noon on Tuesday (recall I do not know about the concert on Tuesday until later), so why should I be considered to have an intention (to have lunch at a cafe) on Monday that I don't have on Tuesday? The problem stems from the fact that the sequences of states in Cohen and Levesque's models represent courses of events of the actual world and the agent's beliefs and goals in the actual world, and not the possible actions the agent could or would have performed in a range of "possible worlds".

Finally, in terms of addressing the logical properties of intentions, as Cohen and Levesque themselves note, the approach of reducing intention to persistent goals is problematic in that the logic of P-GOAL possesses the right properties for capturing the side-effect free principle, that a rational agent need not intend the believed necessary consequences (side-effects) of an intention, but ‘because of what [they] believe to be the wrong reasons,’ Cohen and Levesque [4, p. 238].

Properly capturing the future-directed aspect of intention requires taking into account the nondeterminism of actions and the possibility of the agent’s attempts to fail. An alternative approach more consistent with this aspect of intention was presented by Rao and Georgeff [14,15], whose formalism is based on branching-time structures, in which a situation (a state in a branching-time structure) can have multiple successor situations determined by the different choices available to the agent. One other difference between this approach and Cohen and Levesque’s is that intention is treated as a primitive concept.

Rao and Georgeff [15] explicitly address the side-effect problems using the Strategic Bomber example, however in this formalism, “bombing the plant” and “killing the children” are modelled as propositions, not actions. Thus it is open to Rao and Georgeff to use standard devices in modal logic to ensure the desired conclusions, such as allowing a possible world with a situation that satisfies “bombing the plant” but not “killing the children”, and making such a situation goal and intention-accessible to a situation modelling the state of the world, but not belief-accessible. However, the problem now is how to relate those situations to the agent; such situations relate to the agent only insofar as they satisfy (or not) the agents beliefs, goals and intentions, so this does not (without further elaboration) provide any explanation of any underlying properties of those beliefs, desires and intentions. The difficulty is that such a situation is not one that can actually occur (that is, is not even a possible state the agent will consider or encounter), so it is difficult to independently motivate the inclusion of such a situation in a model of the agent’s reasoning or behaviour.

The formalism of Rao and Georgeff [14] is more complicated, despite appearing earlier. In this model, “transitions” between situations in branching-time structures are labelled with events, for each transition at most one event that either succeeds or “fails”. The aim is to capture the attempts of an agent, which may succeed or fail; note that a failed event labelling a transition refers to the action the agent was trying to do, not to what actually happened as the result of that failed attempt. In any case, it is not clear under this approach how the side-effect free principles *for actions*, in particular for the Strategic Bomber example, should be modelled. Presumably “bombing the plant” and “killing the children” are events, but the restriction in the models to at most one successful or failed event per transition means that the relation between actions of killing the children *by* bombing the plant cannot be modelled, as properly this requires modelling two actions using the one transition (treating “killing the children” as a proposition holding at situations would not work in general). Moreover, as Strategic Bomber succeeds in killing the children (though that was not his intention), even allowing multiple successful events per transition does not solve the problem, because there would still need to be a way to differentiate those successful events the agent intended from those it did not (c.f. the motivational potential of the intention, Bratman [2, Ch. 8]).

In earlier work, Wobcke [17,19], motivated by these problems, we developed a logic called Agent Dynamic Logic (ADL), in which possible worlds corresponded to branching-time execution structures of an agent program, and in which intentions referred to actions successfully performed by the agent. In Wobcke [18], we showed how to automatically derive various logical properties of intention. However, the definitions there were too strong, for, while intention satisfied the side-effect free principle, in that an agent intending to bomb the plant did not intend to kill the children, it was not possible to model an agent that did intend actions that are consequences of its intended actions, so the agent which intended both to bomb the plant *and* kill the children could not be modelled. More formally, the formula $\lvert \pi \wedge \lvert (\pi \cup \psi) \rvert$ was not satisfied unless $\pi \cup \psi$ was the same action as π (when π is an action, $\pi \cup \psi$ subsumes π and can be viewed as a consequence of π). The basic reason for this is that the definitions assumed, as in Rao and Georgeff [14], that only one primitive action could be attempted by the agent in any one situation (as is the case for typical BDI agents, though see Pollack [11]). That this assumption is too strong can be seen by looking at an example of action specialization. Suppose there are two equally desirable plans with the postcondition of being in San Francisco, the “Buridan” case discussed by Bratman [2, p. 11]. On the earlier definitions, an agent intending to execute one of the plans could not also intend to achieve being in San Francisco. This works for PRS agents, which select a plan to achieve a goal only at the time of execution. However, this will not work for planning agents in general, and so won’t adequately cope with the Strategic Bomber example where reasoning and plan selection precede execution.

4 Modelling the Examples

The original semantics of action in ADL was based on execution structures, which provided, for each program term π in the language, a computation tree \mathcal{R}_π with each node corresponding to a state of the environment together with some associated set of belief-alternative states, defined by an accessibility relation \mathcal{B} . Branching in execution structures corresponded to nondeterminism in the execution of actions. With each mental state of the agent was associated a PRS-interpretation, also an execution structure (modelling the possible executions of the current set of plans of the PRS agent), with in addition, a subworld relation \mathcal{I} defining the successful action executions (those achieving the postcondition) and an accessibility relation \mathcal{A} giving the “actual alternative” situations, those situations that the agent could be in given its past history, starting from some initial state of the environment and partially executing plans and making observations. An agent was defined to have an intention to do an action π in a situation σ in an execution structure if for any \mathcal{A} -related situation σ' of σ , some initial part of the execution structure emanating from σ' was isomorphic to the execution structure in \mathcal{R}_π whose root had the same environment state and set of belief-alternatives as σ' .

However, as discussed above, this definition is too strong, as it embodies an assumption, acceptable for PRS, that only one atomic action is attempted at any point in time. To properly model the examples, in particular Strategic Bomber where two actions (“bombing the plant” and “killing the children”) are performed with the same behaviour, as also studied by Israel, Perry and Tutiya [9], we need to modify the definitions as follows. An

agent is now defined to have an intention to do an action π in a situation σ in an execution structure if for any \mathcal{A} -related situation σ' of σ , some initial part of the execution structure emanating from σ' is isomorphic to a *subworld of* the execution structure in \mathcal{R}_π whose root has the same environment state and set of belief-alternatives as σ' . This definition makes valid the formula $I\pi \Rightarrow I(\pi \cup \psi)$, so care must be taken to interpret formulae such as $I(\pi \cup \psi)$, which no longer mean an agent has a choice between π and ψ (it being satisfied, for example, if the agent already intends π).

4.1 Video Game

In the Video Game example, the agent cannot both intend to hit target 1 and intend to hit target 2. However, what the agent does intend depends, we believe, on the exact strategy adopted. The program language in ADL allows the expression of a number of strategies. Perhaps the most natural strategy is that the agent intends to repeatedly guide the relevant missiles (m_1 and m_2) towards targets 1 and 2 (t_1 and t_2) until either is hit, expressed as the program ‘**repeat** *attempt* (*guide*(m_1, t_1) \cup *guide*(m_2, t_2)) **until** (*hit*(t_1) \vee *hit*(t_2))’. Here “guiding a missile towards a target” is understood as an action that succeeds iff the missile is directed towards the target, but which may also fail (otherwise); in both cases, the agent continues executing the plan. The expression *guide*(m_1, t_1) \cup *guide*(m_2, t_2) here means that the agent has a choice between guiding m_1 to t_1 and m_2 to t_2 , which is needed because ADL does not model the execution of two actions simultaneously. Thus what differentiates this example from the normal cases of intention is that the intended action is a complex program, consisting of a variety of constituent actions (e.g. to guide a missile to target 1 or target 2), and the intention is directed towards an attempt to perform the complex program, not towards the component actions. Thus, returning to Bratman’s suggestions, under this strategy, the agent (i) does not intend to shoot at the targets (if this is equivalent to *guide*(m_1, t_1) \cup *guide*(m_2, t_2)), since the agent only intends to attempt this (though repeatedly), (ii) does not intend (merely) to try to hit the targets (*attempt achieve*(*hit*(t_1) \vee *hit*(t_2))), since (iii) does intend to hit one of the targets (*achieve* (*hit*(t_1) \vee *hit*(t_2))), assuming the complex program is a plan for its achievement. The language of ADL provides no way to express conditional intentions of the form to “hit each target if it can”.

4.2 Strategic Bomber

To correctly model the Strategic Bomber, and to handle the side-effect free principle for actions, the action of “bombing the plant” cannot be subsumed by “killing the children”. But there is no transition on states of the environment for “bombing the plant” in which the children are not also killed. To model this scenario, we therefore make use of the idea that agents “keep track” of the intended effects of their actions, as discussed in Bratman [2, p. 159]. Tracking is handled in ADL through modelling the agent’s changes of belief as the result of observations associated with the agent’s chosen actions. In the case of Strategic Bomber, we can define the observation associated with “bombing the plant” to include the issue of whether the plant is destroyed, but not to include the issue of whether the children are killed. In this way, there will be a transition *on situations* for “bombing the plant”, corresponding to a successful execution (hence is intended), that

is not also a transition for “killing the children” (a transition resulting in the same state of the environment but differing in the agent’s beliefs), so that “bombing the plant” is not subsumed by “killing the children”, and the agent intending the former is not forced to intend the latter. However, the malicious Strategic Bomber, who does intend to kill the children, can also be modelled, as the observation associated with “bombing the plant” for him will include the issue of whether the children are killed, so for him, the action of “bombing the plant” is subsumed by “killing the children”, and the intention to bomb the plant implies the intention to kill the children.

4.3 Lottery Problem

As far as the logic of intention in ADL is concerned, there is no reason to prefer a modelling where the agent can intend, or cannot intend, to win the lottery, since ADL does not deal in probabilities and utilities. However, for the formal semantics to intuitively represent the agent’s reasoning and behaviour, there is strong reason to prefer a model in which the agent *cannot* intend to win the lottery. Moreover, this is preferred if we also understand Bratman’s role of intention in “controlling the conduct” of the agent to mean, not just that an intention to *A* leads to the agent trying to *A*, but that normally the agent is expected to succeed in that execution, as is required for the intention to be used reliably in coordinating the agent’s activities. Hence the simple solution to the problem is to model the agent as intending to *attempt* to win the lottery, by choosing a plan, such as “buy ticket”, whose postcondition is having a ticket, with the possible side-effect of winning the lottery.

This solution suggests a simple extension to standard BDI architectures, which typically include plans whose postconditions are *both* the success conditions on their execution and the motivation for their consideration. The example suggests differentiating these two functions, reserving the postcondition for the success condition (here having a ticket), and providing the plan with an additional *goal* (here winning the lottery). As in the Video Game example, the unusual aspect of this situation is that the agent’s intention is merely an attempt whose success falls outside the agent’s control; thus the situation is similar to that of speech acts, as discussed by Cohen and Levesque [5], where a speech act such as a ‘request’ is modelled as an attempt (to get the hearer to perform some action). Again, the agent needs to adopt a plan whose postcondition is the successful execution of the attempt and whose goal is the further perlocutionary effect. So this mechanism would enable a plan-based approach to communication via speech acts to be incorporated into a BDI architecture.

5 Conclusion

Motivated by three examples from the literature on rational agency and intention, we generalized our earlier approach to modelling BDI agents away from PRS-specific assumptions, whilst maintaining the close correspondence between an agent’s internal states and the formal models. We showed how a modification to the formal theory enabled the examples to be modelled, in particular by (i) allowing more than one action to be performed with the same behaviour, and (ii) including a notion of goals which may be indirect effects of plans in BDI agent architectures.

Acknowledgement

This work is funded by an Australian Research Council Discovery Project Grant.

References

1. Bratman, M.E. (1984) 'Two Faces of Intention.' *The Philosophical Review*, **93**, 375–405.
2. Bratman, M.E. (1987) *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA.
3. Bratman, M.E. (1990) 'What is Intention?' in Cohen, P.R., Morgan, J. & Pollack, M.E. (Eds) *Intentions in Communication*. MIT Press, Cambridge, MA.
4. Cohen, P.R. & Levesque, H.J. (1990) 'Intention is Choice with Commitment.' *Artificial Intelligence*, **42**, 213–261.
5. Cohen, P.R. & Levesque, H.J. (1990) 'Rational Interaction as the Basis for Communication.' in Cohen, P.R., Morgan, J. & Pollack, M.E. (Eds) *Intentions in Communication*. MIT Press, Cambridge, MA.
6. Emerson, E.A. & Clarke, E.M. (1982) 'Using Branching Time Temporal Logic to Synthesize Synchronization Skeletons.' *Science of Computer Programming*, **2**, 241–266.
7. Georgeff, M.P. & Lansky, A.L. (1987) 'Reactive Reasoning and Planning.' *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, 677–682.
8. Harel, D. (1979) *First-Order Dynamic Logic*. Springer-Verlag, Berlin.
9. Israel, D.J., Perry, J.R. & Tutiya, S. (1991) 'Actions and Movements.' *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, 1060–1065.
10. Kyburg, H.E., Jr. (1961) *Probability and the Logic of Rational Belief*. Wesleyan University Press, Middletown, CT.
11. Pollack, M.E. (1991) 'Overloading Intentions for Efficient Practical Reasoning.' *Noûs*, **25**, 513–536.
12. Pollock, J.L. (1995) *Cognitive Carpentry*. MIT Press, Cambridge, MA.
13. Pratt, V.R. (1976) 'Semantical Considerations on Floyd-Hoare Logic.' *Proceedings of the Seventeenth IEEE Symposium on Foundations of Computer Science*, 109–121.
14. 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.
15. 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.
16. 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.
17. 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.
18. Wobcke, W.R. (2002) 'Intention and Rationality for PRS-Like Agents.' in McKay, B. & Slaney, J. (Eds) *AI 2002: Advances in Artificial Intelligence*. Springer-Verlag, Berlin.
19. Wobcke, W.R. (2004) 'Model Theory for PRS-Like Agents: Modelling Belief Update and Action Attempts.' in Zhang, C., Guesgen, H.W. & Yeap, W.K. (Eds) *PRICAI 2004: Trends in Artificial Intelligence*. Springer-Verlag, Berlin.
20. Wobcke, W.R., Chee, M. & Ji, K. (2005) 'Model Checking for PRS-Like Agents.' in Zhang, S. & Jarvis, R. (Eds) *AI 2005: Advances in Artificial Intelligence*. Springer-Verlag, Berlin.
21. Wooldridge, M.J. (2000) *Reasoning About Rational Agents*. MIT Press, Cambridge, MA.