

Formalisations of capabilities for BDI-agents

Lin Padgham¹ and Patrick Lambrix²

¹*Department of Computer Science, RMIT University, Melbourne, Australia*

²*Department of Computer and Information Science, Linköpings universitet, Linköping, Sweden*

Abstract. Intentional agent systems are increasingly being used in a wide range of complex applications. Capabilities has recently been introduced into some of these systems as a software engineering mechanism to support modularity and reusability while still allowing meta-level reasoning. This paper presents possible formalisations of capabilities within the framework of beliefs, goals and intentions and indicates how capabilities can affect agent reasoning about its intentions. We define a style of agent commitment which we refer to as a *self-aware* agent which allows an agent to modify its goals and intentions as its capabilities change. We also indicate which aspects of the specification of a BDI interpreter are affected by the introduction of capabilities and give some indications of additional reasoning which could be incorporated into an agent system on the basis of both the theoretical analysis and the existing implementation.

Keywords: agent representation formalisms, agent theory, BDI-agents, agent capabilities

1. Introduction

Agent systems are becoming increasingly popular for solving a wide range of complex problems. Intentional agent systems have a substantial base in theory as well as a number of implemented systems that are used for challenging applications such as air-traffic control and space systems (Rao and Georgeff, 1995). One of the strengths of the BDI *Belief, Desire, Intention* class of systems (including IRMA (Bratman *et al.*, 1988), PRS (Georgeff and Ingrand, 1989), JACK (Busetta *et al.*, 1999b), JAM (Huber, 1999) and UMPRS (Lee *et al.*, 1994)) is their strong link to theoretical work, in particular that of Rao and Georgeff (Rao and Georgeff, 1991), but also Cohen and Levesque (Cohen and Levesque, 1990), Bratman *et al.* (Bratman *et al.*, 1988), Shoham (Shoham, 1993) and Wooldridge (Wooldridge, 2000). Although the theory is not implemented directly in the systems it does inform and guide the implementations (Rao and Georgeff, 1992).

In this paper we investigate how a notion of *capability* can be integrated into the BDI logic of Rao and Georgeff (Rao and Georgeff, 1991), preserving the features of the logic while adding to it in ways that eliminate current intuitive anomalies and mismatches between the theory and implemented systems.



© 2004 Kluwer Academic Publishers. Printed in the Netherlands.

We understand capability as the ability to act rationally towards achieving a particular goal, in the sense of having an abstract plan type that is believed to achieve the goal. Our notion of capability is based on the philosophical idea that “can” implies both *ability* and *opportunity* (Cross, 1986; McCall, 1970). Our understanding of capability is equivalent to the former. Depending on circumstances a capability may not always result in an achievable plan for realising the goal, but it is a prerequisite for such. Lack of either ability or opportunity imply “cannot”. We argue that it is irrational to commit to a goal if one does not have the ability (capability), but that with respect to opportunity it is sufficient to believe that an opportunity may exist in the future. This is somewhat different to the notion of ‘CAN’ explored by Moore (Moore, 1985) and used more recently by Lin, Levesque, Lesperance and Scherl (Lesperance *et al.*, 2000; Lin and Levesque, 1998), where “can” includes both ability and opportunity and is based on the idea of the goal being necessarily achievable, rather than possibly achievable. We argue that for agents operating in complex and dynamic worlds, it is seldom if ever possible to reason about things being necessarily achievable, and that what rational agents use in their planning and acting is more a notion of possibility. It is rational to adopt a goal if it is possibly achievable; rationality does not require the goal to be necessarily achievable.

We describe a possible formal relationship of capabilities to the other BDI concepts of *beliefs*, *goals* and *intentions*. The addition of capabilities enriches the existing formal model and allows for definition of a self-aware agent which takes on and remains committed to goals only if it has a capability for such goals. The formalisation we introduce deals only with a single agent, but we indicate directions for development that would be suitable for dealing with rational behaviour in a multi-agent system which takes into account the known capabilities of other agents.

This work is partially motivated by the development and use of a *capability* construct in JACK, a java based BDI agent development environment (Busetta *et al.*, 1999b), which follows the basic abstract interpreter described in (Rao and Georgeff, 1992). We indicate how capabilities can be integrated into this abstract interpreter and also indicate some issues for consideration in the implementation of capabilities that are highlighted by this work. This work can be seen as part of the ongoing interplay between theory and practice in the area of BDI agent systems. It provides a foundation for exploring some of the practical reasoning mechanisms involving capabilities and for further developing the theory as well as informing the ongoing implementations.

The remainder of the paper is organised as follows. In section 2 we motivate the introduction of capabilities into the BDI-framework and show what kind of reasoning can be achieved. Section 3 describes the BDI-logic that is the basis of our approach while our approach is described in section 4. An extension of (section 5) and an alternative to (section 6) our approach are also briefly described. In section 7 we describe an abstract interpreter which follows the logic of our system. We also make some recommendations for limited changes to the implementation of capabilities in JACK in order to support the kind of reasoning suggested in section 2, in a way that is consistent with the theory of capability as used by us. Finally, in section 8 we compare the notion of capability that we have explored, to related notions of “can” and “ability” in the philosophical literature. We also compare our work to some other work on capability/can/ability in agent systems. Section 9 concludes the paper.

2. Using Capabilities in Reasoning

Most BDI systems contain a *plan library* made up of plans which are essentially abstract specifications for achieving certain goals or doing sub-tasks on the way to achieving a goal. Each plan is associated with a triggering event (which may be an event of type *achieve goal ϕ*). Each plan may also have a list of pre-conditions or a *context* which describes the situation in which the plan is intended to be used. The context condition may be used to bind variables which are then used in the plan body. The plan body is the code which executes the plan. This may contain invocations of sub-goals which allow new plans to flesh out the detail of the plan, calls to external “actions” (e.g. by other agents), or other code in the plan or host language.

We understand having a *capability (for) ϕ* as meaning that the agent has at least one plan that has as its trigger the goal ϕ . The context condition of the plan can be understood as representing the situation in which the agent has the opportunity to achieve ϕ using this plan. In the terms of Cross (Cross, 1986) those worlds in which the context condition of one of the plans to achieve ϕ is true are precisely those worlds in which it is reasonable for the agent to demonstrate that it has the ability to achieve ϕ .

At any given time the agent may be unable to actually use this plan (if the pre-conditions or context are not true then it does not have the opportunity to achieve its goal using this plan). However, if there is no plan for ϕ then clearly no amount of opportunity will enable the agent

to act intentionally in such a way that it brings about ϕ . Thus we say that the agent does not have capability ϕ .¹

This notion of capability ϕ means then that it is *possible* that the agent, if it acts rationally, can achieve ϕ in some future world. Much of the work on reasoning about what agents “can” do, assumes a stronger notion of what we call capability - namely that “can” ϕ , or capability ϕ means that if the agent acts rationally it will *necessarily* achieve ϕ in some future world (e.g. (Moore, 1985; Lesperance *et al.*, 2000; Lin and Levesque, 1998)). We believe that this stronger notion is too restrictive for the kind of reasoning about rational action that is needed in agent systems. Whilst it is possible to know what opportunities exist in the present, goals and intentions are primarily future directed (Bratman *et al.*, 1988), and in a dynamic and uncertain environment, it is seldom possible to reason about what opportunities will necessarily exist in some future world. For rational action it is sufficient that the agent limit its goals to things it may have an opportunity to achieve in some future world.

Given that agents are not omniscient and that perception, or discovering facts about the world may require effort, adopting a goal may actually direct a rational agent’s perception, or information finding action, to watch for or create the opportunity needed to realise its capability. Moore provides an illustrative example where he explains that in order to say an agent “can” open a safe, the agent needs to know how to do the action (or execute the plan) to dial a combination and open a lock, but it also needs to know the combination (Moore, 1985). We would argue that our agent has a capability to open the safe, but it will only have the opportunity if it knows the combination (and perhaps other factors). Having the goal to open the safe, and knowing the opportunity requirements, may lead the agent to direct its perception to acquiring the relevant knowledge - - for example observing carefully when someone else opens the safe.

In complex real systems there is usually far more information potentially available than can reasonably be processed. By knowing the situation in which a capability to achieve a goal can be used, the agent can focus its perception on watching for the appropriate situation or opportunity to use its capability.

The capabilities implemented in JACK are not exactly the notion of capability we have developed so far in terms of reasoning about goals and intentions. However, they are useful in realising efficient

¹ This assumes that all plans explicitly state what goals they achieve, and does not take account of goals being achieved as a result of side-effects. This is consistent with how many BDI systems of which we are aware are implemented, and is part of the mechanism which allows for efficient practical reasoning.

reasoning about capability. To avoid confusion we will refer to JACK's implementation of capabilities as *capability modules*.

A capability module in JACK is essentially a set of plans, a fragment of the knowledge base that is manipulated by those plans and a specification of the interface to the capability module (Busetta *et al.*, 1999a). The interface is specified partially in terms of what events generated external to the capability module, can be handled by it. Thus a part of the interface to a capability module is a list of the goal achievement events that it is designed to handle. Additional sub-goals and the plans that deal with these can be hidden within the internals of the capability module. The interface also specifies what events generated internally are to be visible externally and gives information as to what portion of the knowledge base fragment is used by the capability.

As an example a *scheduling capability module* may contain a set of plans to construct a schedule in a certain domain. The knowledge base fragment defined as part of this module may have knowledge about the objects to be scheduled, their priorities, and various other information that is generated and used as a schedule is being built. There may be a single external goal event called *achieve-schedule* which this capability module responds to, while the only events it generates that are seen externally are events which output a schedule or which notify failure to generate a schedule.

Reasoning about whether or not an agent should adopt a particular top-level goal, can be done by examining its capability module declarations, rather than by examining all plans. For this reasoning to be sound the sets of plans must be abstracted into capability modules in such a way that the module is self contained (i.e. does not necessarily rely on other capability modules for its successful execution), or that its dependencies are explicit and are reasoned about. Abstracting the representation to allow reasoning over capability modules, rather than plans, supports the efficient real-time reasoning that is a critical part of these kind of agent systems. If the agent system is a closed multi-agent system, it is reasonable to assume that any plans needed for successful execution of a goal handled by a capability module, will be found somewhere within the system, if not within that capability module. However, if capabilities are able to be added and deleted (or activated and de-activated) dynamically, or if agents are relying on the capabilities of other agents which may come and go, in order to achieve their goals, then it becomes necessary to represent any sub-goals within a capability module which require another capability, or perhaps the assistance of another agent.

Busetta *et al.* (Busetta *et al.*, 1999a) describe how agents can be built by incorporating specific capabilities. A growing amount of work

in multi-agent systems discusses agents with varying “roles”. If an agent changes roles dynamically the expectation is that their behaviour also changes. One way to achieve this could be to use capabilities. A capability module could specify and implement the things that an agent could do within a particular role. As an agent changed role, appropriate capabilities could then be activated or de-activated.

While a capability (in general language usage) cannot be regarded as a mental attitude similar to beliefs, desires, goals and intentions, beliefs about capabilities (both one’s own and others) are clearly important mental attitudes for reasoning about action.

When we talk about *goals* and *intentions* we expect that they are related to aspects of the world that the agent has (at least potentially) some control over. While it is reasonable to talk about an agent having a desire for it to be sunny tomorrow, having a goal for it to be sunny tomorrow makes little intuitive sense - unless of course our agent believes it can control the weather. Just as *goals* are constrained to be a consistent sub-set of the set of *desires*, and of *beliefs* we would argue that they should also be constrained to be consistent with its *capabilities* (at least within a single agent system - this needs to be modified for multi-agent systems but the notion of capability remains relevant; for multi-agent systems one must also consider capabilities of agents other than oneself). As intentions are commitments to achieve goals these also are intuitively limited to aspects of the world the agent has some control over. Consequently, we would wish our agent’s goals and intentions to be limited by its capabilities (or what it believes to be its capabilities).

Capabilities may also provide a suitable level at which agents in a multi-agent heterogeneous system have information about other agents. An agent observing an (external) event that it may not itself have the capability to respond to, may pass on the event to another agent if it believes that agent has the capability to respond to the event. (Beliefs about) capabilities of other agents may also provide a mechanism for supporting co-operation. An agent in a multi-agent system may contact or try to influence some other agent with the required capability, or alternatively may make decisions about its own actions based on the believed capabilities of other agents. Goals of an agent in a multi-agent system are likely to be constrained (in some way) by the capabilities of other agents as well as one’s own capabilities.

We explore a possible formalisation of capabilities within BDI logic that incorporates them naturally as constraining goals and intentions, while being themselves constrained by beliefs - a rational agent cannot believe itself to have a capability to achieve something which it does not believe is a possible state of affairs in some future world. We first

summarise the BDI logic of Rao and Georgeff and then explore how this can be extended to incorporate capabilities - currently in the context of a single agent reasoning about its own capabilities, although we are also working on extending this to multi-agent systems.

3. The BDI logic of Rao and Georgeff

The logic developed by Rao and Georgeff (e.g. (Rao and Georgeff, 1991; Rao and Georgeff, 1992)) is a logic involving multiple worlds, where each world is a *time-tree* of world states with branching time future and single time past. The various nodes in the future of the time-tree represent the results of different events or agent actions. The different worlds (i.e. different time-tree structures) result from incomplete knowledge about the current state of the world and represent different scenarios of future choices and effects based on differing current state. Formally, we have the following definition (Rao and Georgeff, 1991).²

DEFINITION 1. *An interpretation M is defined as a tuple $\langle W, E, T, \prec, U, \mathcal{B}, \mathcal{G}, \mathcal{I}, \Phi \rangle$. W is a set of worlds, E is a set of primitive event types, T is a set of time points, \prec is a total, transitive and backward-linear binary relation on time points, U is the universe of discourse, and Φ is a mapping of first-order entities to elements in U for any given world and time point. A situation w_t is a world w at a given time point t . \mathcal{B} , \mathcal{G} and \mathcal{I} are accessibility relations for beliefs, goals and intentions, respectively. $\mathcal{B} \subseteq W \times T \times W$, and similarly for \mathcal{G} and \mathcal{I} . We will use \mathcal{B}_t^w for the worlds accessible from world w at time t .*

A world w of W is a tuple $\langle T_w, A_w, S_w, F_w \rangle$ where $T_w \subseteq T$ is a set of time points in w and A_w is \prec restricted to w . S_w and F_w are arc functions that map adjacent time points to events in E . S_w represents successfully occurring events while F_w represents failed events.

The syntax of the language is given in figure 1.

The main value of Rao and Georgeff's formalism is that it avoids anomalies present in some other formalisms whereby an agent is forced to accept as goals (or intentions) all side effects of a given goal (or intention). Modalities are ordered according to a strength relation \langle_{strong} where $BEL \langle_{strong} GOAL \langle_{strong} INTEND$, and modal operators are

² As seen in definition 1, the original Rao and Georgeff formalism defines event types and a mechanism for defining the success and failure of events. Our extension of the BDI formalism does not concern events and we ignore this part of the original formalism in the remainder of this paper.

state formulae:

- - any first-order formula is a state formula.
- - if ϕ_1 and ϕ_2 are state formulae, and x is a variable, then $\neg\phi_1$, $\phi_1 \vee \phi_2$, and $\exists x: \phi_1(x)$ are state formulae.
- - if ϕ is a state formula then $\text{BEL}(\phi)$, $\text{GOAL}(\phi)$ and $\text{INTEND}(\phi)$ are state formulae.
- - if ψ is a path formula, then $\text{optional}(\psi)$ is a state formula.

path formulae:

- - any state formula is a path formula.
- - if ψ_1 and ψ_2 are path formulae, then $\neg\psi_1$, $\psi_1 \vee \psi_2$, $\psi_1 \cup \psi_2$, $\diamond \psi_1$, $\bigcirc \psi_1$ are path formulae.

abbreviations:

- - $\phi_1 \wedge \phi_2$ is defined as $\neg(\neg\phi_1 \vee \neg\phi_2)$
- - $\forall x: \phi(x)$ is defined as $\neg \exists x: \neg \phi(x)$
- - $\text{inevitable}(\psi)$ is defined as $\neg \text{optional}(\neg\psi)$
- - $\square \psi$ is defined as $\neg \diamond \neg \psi$

Figure 1. Syntax of the Rao and Georgeff logic.

not closed under implication with respect to a weaker modality, making formulae such as:

$$\text{GOAL}(\psi) \wedge \text{BEL}(\text{inevitable}(\square(\psi \supset \gamma))) \wedge \neg\text{GOAL}(\gamma)$$

satisfiable. That is it is possible to not have a goal for something which one believes to be a logical consequence of a goal one does have. Thus it is possible to have a goal to go to the dentist, to believe that going to the dentist necessarily involves pain, but *not* have a goal to have pain.

Unlike the logic of predicate calculus BDI logic formulae are always evaluated with respect to particular time points. The logic has two kinds of formulae; *state formulae* are evaluated at a specific point in a time-tree (a situation), whereas *path formulae* are evaluated over a path in a time-tree. The modal operator *optional* is said to be true of a path formula θ at a particular point in a time-tree if θ is true of at least one path emanating from that point. The operator *inevitable* is said to be true of a path formula θ at a particular point in a time-tree if θ is true of all paths emanating from that point. The logic also includes the standard temporal operators \bigcirc (next), \diamond (eventually), \square (always) and \cup (until) which operate over path formulae.

A belief α , (written $\text{BEL}(\alpha)$) implies that α is true in all belief-accessible worlds. Similarly, a goal ($\text{GOAL}(\alpha)$) is something which is true in all goal-accessible worlds and an intention ($\text{INTEND}(\alpha)$) is true in all intention-accessible worlds. The accessibility relations are called \mathcal{B} , \mathcal{G} and \mathcal{I} for BEL, GOAL and INTEND, respectively. The

$M, v, w_t \models q(y_1, \dots, y_n)$ iff $\langle v(y_1), \dots, v(y_n) \rangle \in \Phi[q, w, t]$ where $q(y_1, \dots, y_n)$ is a predicate formula.

$M, v, w_t \models \neg\phi$ iff $M, v, w_t \not\models \phi$

$M, v, w_t \models \phi_1 \vee \phi_2$ iff $M, v, w_t \models \phi_1$ or $M, v, w_t \models \phi_2$

$M, v, w_t \models \exists x: \phi(x)$ iff $M, v_d^x, w_t \models \phi$ for some d in U

$M, v, (w_{t_0}, w_{t_1}, \dots) \models \phi$ iff $M, v, w_{t_0} \models \phi$

$M, v, (w_{t_0}, w_{t_1}, \dots) \models \bigcirc \psi$ iff $M, v, (w_{t_1}, \dots) \models \psi$

$M, v, (w_{t_0}, w_{t_1}, \dots) \models \diamond \psi$ iff $\exists k \geq 0: M, v, (w_{t_k}, \dots) \models \psi$

$M, v, (w_{t_0}, w_{t_1}, \dots) \models \psi_1 \cup \psi_2$ iff

(a) $\exists k \geq 0: M, v, (w_{t_k}, \dots) \models \psi_2$ and

$\forall j, 0 \leq j < k: M, v, (w_{t_j}, \dots) \models \psi_1$

or (b) $\forall j \geq 0: M, v, (w_{t_j}, \dots) \models \psi_1$

$M, v, w_{t_0} \models \text{optional}(\psi)$ iff there exists a fullpath $(w_{t_0}, w_{t_1}, \dots)$ such that

$M, v, (w_{t_0}, w_{t_1}, \dots) \models \psi$

$M, v, w_t \models \text{BEL}(\phi)$ iff $\forall w' \in \mathcal{B}_t^w: M, v, w'_t \models \phi$

$M, v, w_t \models \text{GOAL}(\phi)$ iff $\forall w' \in \mathcal{G}_t^w: M, v, w'_t \models \phi$

$M, v, w_t \models \text{INTEND}(\phi)$ iff $\forall w' \in \mathcal{I}_t^w: M, v, w'_t \models \phi$

Figure 2. Semantics of the Rao and Georgeff logic.

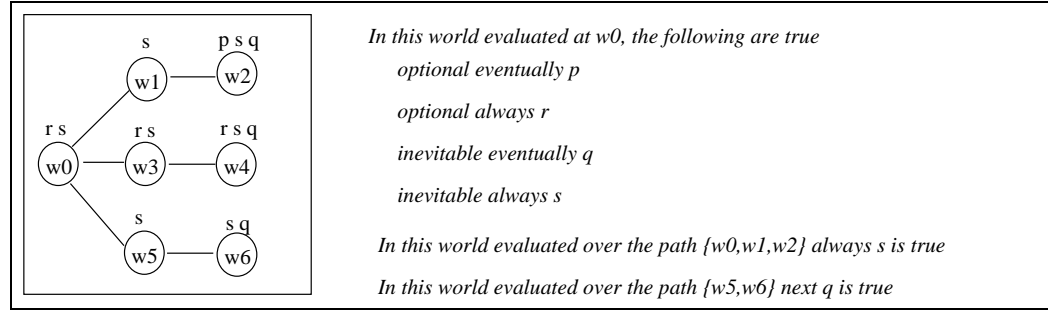


Figure 3. Diagram illustrating evaluation of formulae in a world.

axiomatisation for beliefs is the standard weak-S5 (or KD45) modal system. For goals and intentions the D and K axioms are adopted.

Figure 2 gives the semantics of the language. In the figure, M is an interpretation, w_t is a situation and v is a variable assignment. Further, v_d^x is the function that yields d for the variable x and is the same as v everywhere else.³ Figure 3 illustrates evaluation of some formulae in a belief, goal or intention world (i.e. a time-tree).

The logic requires that goals be compatible with beliefs (and intentions compatible with goals). This is enforced by requiring that for

³ We note that for $M, v, w_t \models \text{BEL}(\phi)$ to hold, w'_t needs to exist for each world w' that is belief-accessible from w_t . Similarly for GOAL and INTEND.

each belief-accessible world w at time t , there must be a goal-accessible sub-world of w at time t . This ensures that no formula can be true in all goal-accessible worlds unless it is true in a belief-accessible world. There is a similar relationship between goal-accessible and intention-accessible worlds. Intuitively, a world is a sub-world (\sqsubseteq) of another world if it is a copy of the first world except that some time branches may be missing. For a formal definition we refer to (Rao and Georgeff, 1991).

The key axioms of what Rao and Georgeff refer to as the *basic I-system* (Rao and Georgeff, 1991) are as follows⁴

AI1 $\text{GOAL}(\alpha) \supset \text{BEL}(\alpha)$

An agent that adopts a formula as a goal (e.g. optional \diamond p) also believes that formula.

AI2 $\text{INTEND}(\alpha) \supset \text{GOAL}(\alpha)$

An agent adopts intentions only towards things that are goals.

AI4 $\text{INTEND}(\phi) \supset \text{BEL}(\text{INTEND}(\phi))$

if an agent intends something it believes that it intends it.

AI5 $\text{GOAL}(\phi) \supset \text{BEL}(\text{GOAL}(\phi))$

if an agent has something as a goal then it believes that it has it as a goal.

AI6 $\text{INTEND}(\phi) \supset \text{GOAL}(\text{INTEND}(\phi))$

if an agent intends something it has the goal to intend it.

AI8 $\text{INTEND}(\phi) \supset \text{inevitable } \diamond (\neg \text{INTEND}(\phi))$

intentions are always dropped eventually

Associated with the axioms AI1, AI2 and AI4-AI6 are a number of semantic conditions:

CI1 $\forall w' \in \mathcal{B}_t^w, \exists w'' \in \mathcal{G}_t^w: w'' \sqsubseteq w'$

For each belief-accessible world, w at time t , there exists a goal-accessible sub-world of w , at time t .

CI2 $\forall w' \in \mathcal{G}_t^w, \exists w'' \in \mathcal{I}_t^w: w'' \sqsubseteq w'$

For each goal-accessible world, w at time t , there exists an intention-accessible sub-world of w , at time t .

CI4 $\forall w' \in \mathcal{B}_t^w, \forall w'' \in \mathcal{I}_t^{w'}: w'' \in \mathcal{I}_t^w$

This restricts intention-accessible worlds from belief-accessible worlds to be intention-accessible worlds.

CI5 $\forall w' \in \mathcal{B}_t^w, \forall w'' \in \mathcal{G}_t^{w'}: w'' \in \mathcal{G}_t^w$

This restricts goal-accessible worlds from belief-accessible worlds to

⁴ AI1 and AI2 only hold for so-called O-formulae which are formulae with no positive occurrences of inevitable outside the scope of the modal operators. See (Rao and Georgeff, 1991) for details. Also \supset is implication (not superset). Further, AI3 and AI7 in the original framework deal with events and are not shown here.

be goal-accessible worlds.

CI6 $\forall w' \in \mathcal{G}_t^w, \forall w'' \in \mathcal{I}_t^{w'}: w'' \in \mathcal{I}_t^w$

This restricts intention-accessible worlds from goal-accessible worlds to be intention-accessible worlds.

The conditions CI4, CI5 and CI6 are subtly different from the conditions in (Rao and Georgeff, 1991), which appear to be slightly wrong. The (Rao and Georgeff, 1991) versions of CI4, CI5 and CI6 respectively are:

$\forall w' \in \mathcal{B}_t^w, \forall w'' \in \mathcal{I}_t^w: w'' \in \mathcal{B}_t^{w'}$

$\forall w' \in \mathcal{B}_t^w, \forall w'' \in \mathcal{G}_t^w: w'' \in \mathcal{B}_t^{w'}$

$\forall w' \in \mathcal{G}_t^w, \forall w'' \in \mathcal{I}_t^w: w'' \in \mathcal{G}_t^{w'}$

We first show that our version of CI4 leads to the desired results. Assume CI4 and $M, v, w_t \models \text{INTEND}(\phi)$. Then we want to prove that $M, v, w_t \models \text{BEL}(\text{INTEND}(\phi))$ or $\forall w^1 \in \mathcal{B}_t^w: M, v, w_t^1 \models \text{INTEND}(\phi)$ or $\forall w^1 \in \mathcal{B}_t^w: \forall w^2 \in \mathcal{I}_t^{w^1}: M, v, w_t^2 \models \phi$. Given CI4 we know that $\forall w^1 \in \mathcal{B}_t^w: \forall w^2 \in \mathcal{I}_t^{w^1}: w^2 \in \mathcal{I}_t^w$. Further, as $M, v, w_t \models \text{INTEND}(\phi)$, we know that $\forall w^2 \in \mathcal{I}_t^w: M, v, w_t^2 \models \phi$. This gives the result.

Our version of CI4 restricts intention-accessible worlds from belief-accessible worlds to be intention-accessible worlds. This is the key to the result as we have assumed that all intention-accessible worlds model ϕ . The (Rao and Georgeff, 1991) version of CI4, however, does not say anything about intention-accessible worlds from belief-accessible worlds. It allows for intention-accessible worlds from belief-accessible worlds not to be intention-accessible. Therefore, these worlds can model ϕ or $\neg \phi$ and thus $\text{BEL}(\text{INTEND}(\phi))$ does not necessarily hold. Similar comments hold for CI5 and CI6.

The framework can then be used as a basis for describing and exploring various commitment axioms that correspond to agents that behave in various ways with respect to commitment to their intentions. Rao and Georgeff describe axioms for what they call a blindly committed agent, a single-minded agent and an open-minded agent, showing that as long as an agent's beliefs about the current state of the world are always true, as long as the agent only acts intentionally⁵, and as long as nothing happens that is inconsistent with the agent's expectations, then these agents will eventually achieve their goals.

4. Semantics of Capabilities

As discussed previously it makes little intuitive sense to have a goal and an intention for the sun to shine, unless an agent also has some

⁵ This includes not dropping goals as this would lead to dropping intentions.

mechanism for acting to achieve this world state. We extend the BDI logic of Rao and Georgeff's *I-system* (Rao and Georgeff, 1991; Rao and Georgeff, 1992) to incorporate capabilities which constrain agent goals and intentions to be compatible with what it believes are its capabilities. We call our extended logic the *IC-system* (Padgham and Lambrix, 2000).

The *IC-system* requires capability-accessible worlds exactly analogous⁶ to belief-accessible worlds, goal-accessible worlds and intention-accessible worlds. $CAP(\phi)$ is then defined as being true if ϕ is true in all the capability-accessible worlds. If \mathcal{C} is the accessibility relation with respect to capabilities, then

$$M, v, w_t \models CAP(\phi) \text{ iff } \forall w' \in \mathcal{C}_t^w: M, v, w'_t \models \phi$$

We adopt the K and D axioms for capabilities, i.e. capabilities are closed under implication and consistent. Similarly to the belief, goal and intention accessible worlds, we also constrain the capability accessible worlds based on their compatibility with the worlds accessible via the other modalities. Therefore, in the next section we give axioms and semantic conditions to capture the desired interrelationships among an agent's beliefs, capabilities, goals and intentions.

4.1. COMPATIBILITY AXIOMS

The first two axioms of the basic *I-system* described in the previous section have to do with the compatibility between beliefs and goals, and goals and intentions. We add two further compatibility axioms relating to capabilities. Note that the compatibility axioms refer only to so-called O-formula, i.e. formula that do not contain any positive occurrences of *inevitable* outside the scope of the modal operators.

Belief-Capability Compatibility

This axiom states that if the agent has an O-formula α as a capability, the agent believes that formula.

AIC1 $CAP(\alpha) \supset BEL(\alpha)$

Thus if an agent has the capability that *optional*(ψ) is true, this also implies a belief that *optional*(ψ) is true. This should not be read as having a capability for α implies that α is believed to be true. The natural language semantics is closer to the statement that if an agent has a capability for α (at time t), then the agent believes that it is possible for α to be true (at time t). Statements where α is a simple predicate rather than a formula involving *optional* must be evaluated at a particular time point. So $CAP(\text{rich}) \supset BEL(\text{rich})$ means that if an agent is capable of being rich now then the agent believes he is rich

⁶ See section 5 for an alternative definition.

now. Importantly it does not mean that if the agent has a capability of being rich in the future, he believes that he is rich in the future - he believes only that there is some possible future where he is rich. To say that the agent is capable of being rich in the future we would write $CAP(\text{optional} \diamond \text{rich})$ or $CAP(\text{inevitable} \diamond \text{rich})$. In the first case we state that in the capability-accessible worlds there always is a future in which the agent is rich. However, there is no guarantee that this future will actually be the future for the agent. If the agent has the capability of being rich in every possible future, then the second case may be used. We observe that this is a much stronger statement than the first statement. We also note that intuitively it only really makes sense to talk about capabilities (and goals and intentions) with respect to future time, so the semantics of formulae such as $CAP(\text{rich}) \supset BEL(\text{rich})$ are intuitively awkward though not problematic. This is inherent in the original logic and applies to goals and intentions at least as much as to capabilities. It could be addressed by limiting the form of valid formulae using CAP, GOAL and INTEND but we have chosen to remain consistent with the original BDI logic.

The semantic condition associated with this axiom is:⁷

CIC1 $\forall w' \in \mathcal{B}_t^w, \exists w'' \in \mathcal{C}_t^w: w'' \sqsubseteq w'$.

For each belief-accessible world, w at time t, there exists a capability-accessible sub-world of w, at time t.

As mentioned before, intuitively, a world is a sub-world (\sqsubseteq) of another world if it is a copy of the first world except that some time branches may be missing. Figure 4 illustrates a structure that satisfies the semantic condition CIC1.

Capability-Goal Compatibility

This axiom and associated semantic condition states that if the agent has an O-formula α as a goal, then the agent also has α as a capability. This constrains the agent to adopt as goals only formulae where there is a corresponding capability.

AIC2 $GOAL(\alpha) \supset CAP(\alpha)$

Having a goal for something, implies having a capability for that something.

⁷ $\mathcal{B}, \mathcal{C}, \mathcal{G}$ and \mathcal{I} are the accessibility relations with respect to beliefs, capabilities, goals and intentions respectively.

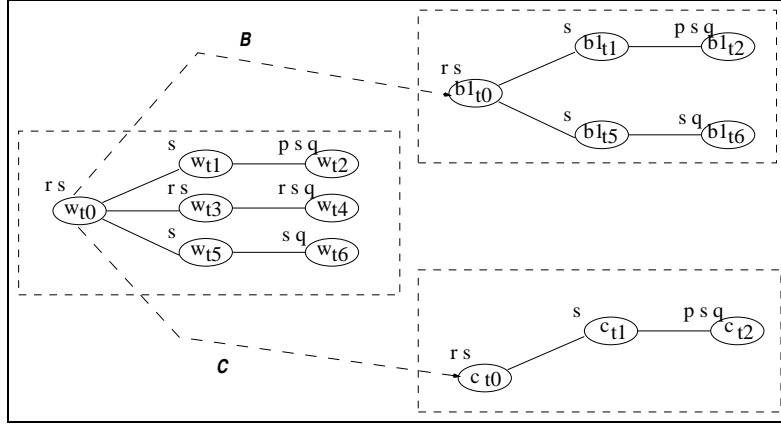


Figure 4. A structure satisfying CIC1.

CIC2 $\forall w' \in \mathcal{C}_t^w, \exists w'' \in \mathcal{G}_t^w: w'' \sqsubseteq w'$. For each capability-accessible world, w at time t , there exists a goal-accessible sub-world of w , at time t .

4.2. MIXED MODALITY AXIOMS

Axioms AI4, AI5 and AI6 define the relationships when the BEL, GOAL and INTEND modalities are nested. We add two new axioms and a corollary along with semantic conditions to capture the relationship between CAP and each of the other modalities. We note that the original axiom AI4 actually follows from AI1 and AI6.

Beliefs about Capabilities

If the agent has a capability α then it believes that it has a capability α .

AIC3 $\text{CAP}(\alpha) \supset \text{BEL}(\text{CAP}(\alpha))$

CIC3 $\forall w' \in \mathcal{B}_t^w, \forall w'' \in \mathcal{C}_t^{w'}: w'' \in \mathcal{C}_t^w$

This restricts capability-accessible worlds from belief-accessible worlds to be capability-accessible worlds.

Capabilities regarding Goals

If an agent has a goal α then it has the capability to have the goal α .

AIC4 $\text{GOAL}(\alpha) \supset \text{CAP}(\text{GOAL}(\alpha))$

CIC4 $\forall w' \in \mathcal{C}_t^w, \forall w'' \in \mathcal{G}_t^{w'}: w'' \in \mathcal{G}_t^w$

This restricts goal-accessible worlds from capability-accessible worlds to be goal-accessible worlds.

Capabilities regarding Intentions

If an agent has an intention α it also has the capability to have the intention α .

Follows from AIC2 and AI6

$\text{INTEND}(\alpha) \supset \text{CAP}(\text{INTEND}(\alpha))$

semantic condition:

$\forall w' \in \mathcal{C}_t^w, \forall w'' \in \mathcal{I}_t^{w'}: w'' \in \mathcal{I}_t^w$

This restricts intention-accessible worlds from capability-accessible worlds to be intention-accessible worlds.

Strengthening of this group of axioms by replacing implication with equivalence would result in the expanded version of the equivalences mentioned in (Rao and Georgeff, 1991) namely $\text{INTEND}(\alpha) \equiv \text{BEL}(\text{INTEND}(\alpha)) \equiv \text{CAP}(\text{INTEND}(\alpha)) \equiv \text{GOAL}(\text{INTEND}(\alpha))$ and $\text{GOAL}(\alpha) \equiv \text{BEL}(\text{GOAL}(\alpha)) \equiv \text{CAP}(\text{GOAL}(\alpha))$. Equivalence strengthening would also give $\text{CAP}(\alpha) \equiv \text{BEL}(\text{CAP}(\alpha))$, which would imply that the agent has full knowledge about its capabilities. As mentioned in (Rao and Georgeff, 1991) this has the effect of collapsing mixed nested modalities to their simpler non-nested forms.

We refer to the axioms AI2, AI6, AI8, AIC1, AIC2, AIC3 and AIC4 as the *basic IC-system*. We note that all axioms of the *I-system* remain true, although some are consequences rather than axioms.⁸

4.3. COMMITMENT AXIOMS

Rao and Georgeff define three variants of a commitment axiom, which taken together with the basic axioms define what they call a *blindly committed agent*, a *single-minded agent* and an *open-minded agent*. The blindly committed agent maintains intentions until they are believed true, the single-minded agent maintains intentions until they are believed true or are believed impossible to achieve, while the open-minded agent maintains intentions until they are believed true or are no longer goals.

We define an additional kind of agent which we term a *self-aware agent* which is able to drop an intention if it believes it no longer has the capability for that intention.

⁸ AI1 follows from AIC1 and AIC2. AI4 follows from AIC1, AIC2 and AI6. AI5 follows from AIC1 and AIC4.

The *self-aware agent* is defined by the *basic IC-system* plus the following axiom which we call AIC9d.⁹

$$\begin{aligned} \mathbf{AIC9d} \quad & \text{INTEND}(\text{inevitable} \diamond \phi) \supset \\ & \text{inevitable}(\text{INTEND}(\text{inevitable} \diamond \phi) \\ & \cup (\text{BEL}(\phi) \vee \neg \text{CAP}(\text{optional} \diamond \phi))) \end{aligned}$$

Intending ϕ implies either continuing to intend ϕ , or the agent believes ϕ is true, or it does not have the capability for ϕ in any possible future.

As an example, a self-aware agent that has the intention of being rich in every possible future, will keep this intention until he believes he is rich or until he does not have the capability of being rich in the future any more. This last fact would mean that there are capability-accessible worlds where the agent does not become rich in any possible future in that world.

It is then possible to show that a self-aware agent will inevitably eventually believe its intentions, and to prove a new theorem that under certain circumstances the self-aware agent will achieve its intentions. Self-awareness can be combined with either open-mindedness or single-mindedness to obtain self-aware-open-minded and self-aware-single-minded agents.

THEOREM 1. *A basic self-aware agent, with the basic IC-system and the axiom AIC9d, satisfies the following property, that if an agent intends something, and retains the capability for that something until it believes it is true, then it will eventually believe that that something is true:*

$$\begin{aligned} & \text{INTEND}(\text{inevitable} \diamond \phi) \wedge \\ & \text{inevitable}(\text{CAP}(\text{optional} \diamond \phi) \cup \text{BEL}(\phi)) \\ & \supset \text{inevitable}(\diamond \text{BEL}(\phi)) \end{aligned}$$

Proof:

Assume the premise. Then from AIC9d we can conclude $\text{inevitable}(\text{INTEND}(\text{inevitable} \diamond \phi) \cup (\text{BEL}(\phi) \vee \neg \text{CAP}(\text{optional} \diamond \phi)))$. By AI8 and the definition of \cup we can conclude $\text{inevitable}(\diamond (\text{BEL}(\phi) \vee \neg \text{CAP}(\text{optional} \diamond \phi)))$. Given the fact that $\text{inevitable}(\text{CAP}(\text{optional} \diamond \phi) \cup \text{BEL}(\phi))$, we can conclude that $\text{inevitable}(\diamond \text{BEL}(\phi))$. ♣

As an example, the theorem states that if an agent is a self-aware agent, who intends to be rich in the future and retains his capability of being rich in the future until he believes he is rich, then the agent will believe that he is rich at some point of time in the future.

⁹ This numbering is chosen because of the relationship of AIC9d to AI9a, AI9b, and AI9c in the original *I-system*.

A *competent agent* (Cohen and Levesque, 1990) is an agent that satisfies the axiom of true beliefs, i.e. $BEL(\phi) \supset \phi$ (**AI10**). It can then be shown that a competent self-aware agent *will* achieve its intentions, rather than just believing so. However, as discussed in (Rao and Georgeff, 1991), AI10 is often difficult to live up to for agents as it requires true beliefs about future realisation of its intentions. AI10 may therefore need to be restricted to current beliefs or to beliefs about primitive actions.

THEOREM 2. *A competent basic self-aware agent, with the basic IC-system and the axioms AIC9d and AI10, satisfies the following property, that if an agent intends something, and retains the capability for that something until it believes it is true, then that something will eventually be true:*

$$\begin{aligned} & INTEND(inevitable\Diamond\phi) \wedge \\ & inevitable(CAP(optional\Diamond\phi) \cup BEL(\phi)) \\ & \supset inevitable(\Diamond\phi) \end{aligned}$$

Proof: Follows directly from the proof of the theorem 1 and AI10.



4.4. PROPERTIES OF THE LOGIC

The logic allows for believing things without having the capability for this, i.e. $BEL(\phi) \wedge \neg CAP(\phi)$ is satisfiable. This means that, for instance, you can believe the sun will inevitably rise, without having a capability for this. Also $inevitable(\Box BEL(\phi)) \wedge \neg GOAL(\phi)$ is satisfiable. Similarly, one can have the capability for something without having the goal for this. In general, a modal formula does not imply a stronger modal formula, where $BEL <_{strong} CAP <_{strong} GOAL <_{strong} INTEND$.

THEOREM 3. *For modalities R_1 and R_2 such that $R_1 <_{strong} R_2$, the following formulae are satisfiable:*

- (a) $R_1(\phi) \wedge \neg R_2(\phi)$
- (b) $inevitable(\Box R_1(\phi)) \wedge \neg R_2(\phi)$

Proof: We prove the result for BEL and CAP. The proof for the other pairs of modalities is similar. Assume $BEL(\phi)$. Then, ϕ is true in every belief-accessible world. For every belief-accessible world there is a capability-accessible world. However, \mathcal{C} may map to worlds that do not correspond to any belief-accessible world. If ϕ is not true in one of these worlds, then ϕ is not a capability. This shows the satisfiability of (a). Similar reasoning yields (b). ♣

As we have seen before, the modalities are closed under implication. However, another property of the logic is that a modal operator is not closed under implication with respect to weaker modalities. For instance, an agent may have the capability for ϕ , believe that ϕ implies γ , but not have the capability for γ .¹⁰ As an example, an agent may have the capability for a rain dance, believe that doing a rain dance will make it rain, but he may not have the capability for rain.¹¹ This solves the side-effect problem as mentioned in section 3.

THEOREM 4. *For modalities R_1 and R_2 such that $R_1 <_{strong} R_2$, the following formulae are satisfiable:*

- (a) $R_2(\phi) \wedge R_1(\text{inevitable}(\Box(\phi \supset \gamma))) \wedge \neg R_2(\gamma)$
- (b) $R_2(\phi) \wedge \text{inevitable}(\Box R_1(\text{inevitable}(\Box(\phi \supset \gamma)))) \wedge \neg R_2(\gamma)$

Proof: We prove the result for BEL and CAP. The proof for the other pairs of modalities is similar. Assume $\text{CAP}(\phi)$ and $\text{BEL}(\text{inevitable}(\Box(\phi \supset \gamma)))$. Then, ϕ is true in every capability-accessible world. To be able to infer that γ is true in each capability-accessible world, we would need that $\phi \supset \gamma$ is true in each capability-accessible world. We know that for every belief-accessible world $\text{inevitable}(\Box(\phi \supset \gamma))$ is true and that for each belief-accessible world there is a capability-accessible world. However, \mathcal{C} may map to other worlds, where this is not true and thus γ is not a capability. This shows the satisfiability of (a). Similar reasoning yields (b). ♣

The formal semantics of capabilities as defined fit well into the existing BDI logic of Rao and Georgeff and allow definition of further interesting types of agents. In section 7 we look at how this addition of capabilities affects the specification of an abstract interpreter for BDI systems and also what issues and questions arise for implementations as the result of the theoretical exploration. First, however, we look at an extension of and an alternative for the basic *IC-system*.

5. Extension of the IC-system

As indicated previously our semantics for capability does not support reasoning that capability ϕ , and the belief that capability ϕ always

¹⁰ The alternative formulation referred to in section 5 does not have this property with respect to capabilities.

¹¹ The agent believes that it will rain, however. One may argue that an agent that believes to have the capability for ϕ and believes that ϕ leads to γ , also believes that he has the capability for γ . However, this would require the introduction of a new axiom such as $\text{BEL}(\text{CAP}(\phi)) \wedge \text{BEL}(\text{inevitable}(\Box(\phi \supset \gamma))) \supset \text{BEL}(\text{CAP}(\gamma))$.

implies capability γ to lead to capability γ . While this semantics is justified by the work of others (Cross, 1986), and indeed seems intuitively correct in some cases, it may also be the case that for some applications we would prefer a semantics that does allow the above reasoning.

In this section we define an extension of the basic framework where the semantics is such that the inference above is valid, supporting the above reasoning.

5.1. COMPATIBILITY AXIOM

The *IC2-system* requires in addition to the *IC-system* the following axiom:

AIC1b $BEL(\textit{inevitable } \phi) \supset CAP(\textit{inevitable } \phi)$

The axiom states that if an agent believes that something is inevitably true then it also has the capability for this. Thus, if an agent believes it is inevitable that the sun will rise then it is assumed that it also has this capability.

The semantic condition associated with this axiom is:

CIC1b $\forall w' \in \mathcal{C}_t^w, \exists w'' \in \mathcal{B}_t^w: w' \sqsubseteq w''$.

Thus every capability-accessible world is a sub-world of some belief-accessible world.

5.2. COMMITMENT AXIOMS

The definition of self-aware agent is not affected by the addition of AIC1b. In practice, however, the interaction between capabilities and beliefs is different.

Theorems 1 and 2 are still valid in this system.

5.3. PROPERTIES OF THE LOGIC

Theorem 3 holds for formulae of the form *optional* ϕ .

THEOREM 5. *The following formulae are satisfiable:*

- (a) $BEL(\textit{optional } \phi) \wedge \neg CAP(\textit{optional } \phi)$
- (b) $\textit{inevitable}(\Box BEL(\textit{optional } \phi)) \wedge \neg CAP(\textit{optional } \phi)$

Proof: Assume $BEL(\textit{optional } \phi)$. Then, *optional* ϕ is true in every belief-accessible world, i.e. there is a path in every belief-accessible world such that ϕ is true. Every capability-accessible world is a sub-world of some belief-accessible world and for every belief-accessible world there is a capability-accessible sub-world. However, the capability-accessible worlds do not need to contain the branch where ϕ is true

and therefore *optional* ϕ does not need to be true in every capability-accessible world. This shows the satisfiability of (a). Similar reasoning yields (b). ♣

Theorem 4 does not hold anymore for CAP and BEL. As this is the theorem that states that it is possible to have $CAP(\phi) \wedge BEL(\text{inevitable}(\Box(\phi \supset \gamma))) \wedge \neg CAP(\gamma)$, this is precisely what we want. We obtain the additional theorem that gives the desired implication that having capability ϕ and believing that ϕ inevitably implies γ is equivalent to having the capability γ .

THEOREM 6. (a) $CAP(\phi) \wedge BEL(\text{inevitable}(\Box(\phi \supset \gamma))) \supset CAP(\gamma)$
 (b) $CAP(\phi) \wedge \text{inevitable}(\Box BEL(\text{inevitable}(\Box(\phi \supset \gamma)))) \supset CAP(\gamma)$

Proof: Follows from AIC1b and the K axiom for CAP. ♣

6. Alternative to the IC-system

In the basic *IC-system* an agent adopts goals for which there is a corresponding capability. An alternative way to constrain goal adoption is to allow the agent to adopt as goals only formulae where there is a corresponding *belief* to have the capability. This means that we have an alternative capability-goal compatibility which states that if an agent has an O-formula α as a goal, then the agent must believe to have α as a capability.

Capability-Goal Compatibility

AIC2' $GOAL(\alpha) \supset BEL(CAP(\alpha))$

CIC2' $\forall w' \in \mathcal{B}_t^w, \forall w'' \in \mathcal{C}_t^{w'}, \exists w''' \in \mathcal{G}_t^w: w''' \sqsubseteq w'$.

We observe that in this case AI1 cannot be obtained from AIC1 and AIC2', and therefore must be stated as a separate axiom.

Let the *IC'-system* be the system satisfying AI1, AI2, AI4-AI6, AI8, AIC1, AIC2', AIC3 and AIC4. The *IC-system* and the *IC'-system* both satisfy $GOAL(\alpha) \supset BEL(CAP(\alpha))$ for O-formulae. Further, in the *IC-system* goals are always constrained by capabilities. Thus, if the agent does not have the capability, then it cannot have the goal. In the *IC'-system*, however, the situation can occur that the agent does not have the capability but still has the goal. This can happen when the agent believes it has the capability.¹²

¹² The axiom in the *IC'-system* seems more intuitive for extending to multi-agent systems. There one would require that $GOAL(a, \alpha) \supset BEL(a, CAP(a, \alpha) \vee \exists a': CAP(a', \alpha))$.

We can then define an alternative self-aware agent by the *IC'-system* plus the following axiom AIC9d'.

$$\begin{aligned} \mathbf{AIC9d'} \quad & \text{INTEND}(\text{inevitable} \diamond \phi) \supset \\ & \text{inevitable}(\text{INTEND}(\text{inevitable} \diamond \phi)) \\ & \cup (\text{BEL}(\phi) \vee \neg \text{BEL}(\text{CAP}(\text{optional} \diamond \phi))) \end{aligned}$$

It is then possible to prove alternatives for theorems 1 and 2.

We observe also that the *IC-system* and the *IC'-system* become equivalent in the case of competent agents (satisfying AI10, $\text{BEL}(\phi) \supset \phi$). This is because AIC2' can be obtained from AIC2 and AIC3, while AIC2 can be obtained from AIC2' and AI10.

7. Implementation aspects

An abstraction of a BDI-interpreter which follows the logic of the basic *I-system* is given in (Rao and Georgeff, 1992). The system maintains three global basic data structures representing beliefs, goals and intentions. We add now a data structure for capabilities. Each of these data structures allow for update and query operations. Further, there also exists a global event queue. The first stages in the cycle of this abstract interpreter are to generate and select plan options. These are filtered by current beliefs, goals and intentions. Capabilities (in the *IC-system*) or beliefs about capabilities (in the *IC'-system*) now provide an additional filter on the options we generate and select. Once the options are selected, the intention structure is updated and intentions on the intention structure are executed. Then, external events are collected. We note that update operations to the internal data structures can be put on the global event queue continuously and therefore do not need an explicit procedure. The drop procedures in the original interpreter drop beliefs, goals and intentions based on the achievement of goals or on the impossibility to achieve them. In the new procedures capabilities (in the *IC-system*) and beliefs about capabilities (in the *IC'-system*) must be considered when dropping beliefs, goals and intentions. In a system with dynamic roles capabilities themselves may also be dropped. Also, in the *IC2-system* beliefs must be considered for dropping capabilities. Thus we obtain this slightly modified version of the interpreter in (Rao and Georgeff, 1992) as shown in figure 5.

This abstract interpreter is at a very high level and there are many details which must be considered in the actual implementation that are hidden in this abstraction. One important implementation detail that is highlighted by the definitions of the various kinds of agents (blindly committed, single-minded, open-minded and self-aware) has to

```

initialise-state();
do
  options :=
    option-generator(event-queue,B,C,G,I);
  selected-options :=
    deliberate(options,B,C,G,I);
  update-intentions(selected-options,I);
  execute(I);
  get-new-external-events();
  drop-successful-attitudes(B,C,G,I);
  drop-impossible-attitudes(B,C,G,I);
until quit.

```

Figure 5. BDI with capabilities interpreter

do with when intentions should be dropped. With respect to capabilities the axiom AIC9d highlights the fact that if capabilities are allowed to change during execution it may be necessary to drop some intentions when a capability is lost/removed.

The observation that it is possible for an agent to have the capability for ϕ , believe that ϕ implies γ , but not have the capability for γ (*IC-system*), highlights an area where one may wish to make the agent more “powerful” in its reasoning by disallowing this situation. This is possible by a modification of the logical formalisation (*IC2-system*) but has an impact on how the option generation and selection phases of the abstract interpreter work. However, we observe that in this case the effect of a plan is not necessarily part of the plan, but may have been derived through a belief.

In (Rao and Georgeff, 1992) an example is given to illustrate the workings of the specified abstract interpreter. In this example John wants to quench his thirst and has plans (which are presented as a special kind of belief) for doing this by drinking water or drinking soda, both of which then become options and can be chosen as intentions (instantiated plans that will be acted on).

It is also possible to construct the example where the agent believes that rain always makes the garden wet, and that rain is eventually possible, represented as:

$$\text{BEL}(\textit{inevitable} \Box(\text{rain}) \supset (\text{garden-wet}))$$

$$\text{BEL}(\textit{optional} \Diamond(\text{rain}))$$

In the Rao and Georgeff formalism which does not differentiate between plans and other kinds of beliefs this would allow our agent to adopt (rain) as a GOAL. However, in the absence of any plan in the plan

library for ever achieving rain this does not make intuitive sense - and in fact could not happen in implemented systems. With the *IC-system* presented here we would also require $CAP(\text{optional } \diamond(\text{rain}))$ thus restricting goal adoption to situations where the agent has appropriate capabilities (i.e. plans).

This example demonstrates that in some respects the *IC-system* is actually a more correct formalisation of implemented BDI systems than the original *I - system*.

8. Related work

A number of systems have started to implement capability-like entities. In JACK (Busetta *et al.*, 1999a) capabilities are essentially a set of plans, a fragment of the KB and an interface to the capability. In the KAOs system (Bradshaw *et al.*, 1997) capabilities are the services or functions that an agent can provide. LARKS (Sycara *et al.*, 1999) defines a capability specification as a frame containing slots for context, types, input, output, constraints on input and output, conceptual and textual descriptions. None of these programming constructs are explicitly related to theoretical formalisms involving capability, ability, or a notion of “can”.

There is a large body of theoretical work exploring the notion of ability, achievability or “can”. More recently there is also work on defining these concepts within computational systems that are able to act in the world. For instance, in (Lin and Levesque, 1998; Lesperance *et al.*, 2000) achievability is defined as what goals can be achieved by a robot given a basic action theory describing an initial state of the world and some primitive actions available to the robot. Ability to achieve a goal involves knowing what to do when in order to arrive at a goal state. More precisely, an agent can achieve a goal in a situation if there exists an action selection function such that the agent knows in the original situation that it can get to a situation where the goal holds. In (Dung, 1998) capability is defined as a function C such that for a plan p and a state s , $C(p,s)$ represents the set all possible execution processes which could occur when the agent is executing p from s .

This work follows the general style of Moore’s definition of “can” which requires that in order to say that an agent “can” ϕ , it must be the case that if the agent wants ϕ , and the agent acts rationally, then it will achieve ϕ . As discussed previously, this is a much more restrictive view of capability than that which we use.

Munindar Singh defines a concept of “know-how” (Singh, 1998), which is closely related to capability. He states that “an agent x knows

how to achieve p , if he is able to bring about p through his actions” (Singh, 1998, page 14). Singh also notes that ability is a concept separate from opportunity, but that logical consideration of ability without opportunity is technically complex. In his own formalism, know-how is anchored to a point in time, and requires that in order to know how to achieve p , the agent is able to select a series of actions, from that point in time, that will result in p being true. Although the point in time is most easily the current point - in which case opportunity must also be present, it is possible for it to be some future time point. Our sense of capability could then be described as the agent believing that it is possible that there is a future time point where the agent “knows how”.

We consider our definition of capability is precisely what is needed to guide an agent in whether or not it is rational to adopt a particular goal. However, the stronger notion of capability is possibly what is required in order to commit to a plan - a rational agent chooses the means by which it will achieve its goal, based on opportunity as well as what we have called capability. In this sense the agent “can” achieve ϕ when it has both a plan or way of achieving ϕ and an opportunity to use that plan. It is at this point that the rational agent should commit to the particular plan. This stronger notion of capability is that which is used in work such as that by Shapiro et al. (Shapiro et al., 1995) and allows reasoning about guarantees of achieving a goal. The two notions of capability are perhaps appropriate to different stages of commitment - the weaker notion is appropriate in order to make a rational commitment to “intend that”, whereas the stronger notion is appropriate in order to make the commitment to “intend how”.

This ambiguity over the meaning of “can” is discussed in the philosophical literature, with a number of ideas as to what is the basis of the ambiguity and what must be examined in order to obtain the correct intuitive meaning. Some authors (e.g. (McCall, 1970)) make a distinction between *individual actions* and *action types*. It is possible to explain the difference between our view of capability, and that of Levesque et al (Lin and Levesque, 1998; Lesperance *et al.*, 2000) as being primarily this difference between generic plan types and specific instantiated plan instances.

However, we think the distinction between ability and opportunity as discussed by Cross (Cross, 1986), with particular reference to the importance of the ability sense of “can”, is the most important distinction between our work and that of others using the notion to reason about what computational agents “can” do. We note Cross’ argument that if “can” is a modal operator (as we have defined it to be), then it is necessary that the accessibility relation be relative to an individual,

otherwise anomalies occur when formalising natural language expressions. For example (taken from (Cross, 1986)) the sentences:

Bill can balance a banana on his nose while I stand on his shoulders,
and

I can stand on Bill's shoulders while he balances a banana on his nose
would both be formalised as:

CAN(Bill balances a banana on his nose while I stand on his shoulders)

However in order to capture the intuitive meaning (and differences) in the above sentences, we need something more like:

(Bill CAN(Bill balances a banana on his nose while I stand on his shoulders)), and

(I CAN(Bill balances a banana on his nose while I stand on his shoulders))

We agree completely with this, but as we are reasoning only about the capabilities of a single agent, this relativisation is implicit.

We note that the situation we have in our basic IC-system, where $CAP(\phi) \wedge BEL(\phi \rightarrow \gamma)$ does not imply $CAP(\gamma)$ also exists in the model theoretic explanation of “can” given by Cross, even though this seems at first glance to be odd. He argues that when can means ability (as it does for us), then the above implication should indeed not follow.

van Linder, van der Hoek and Meyer have done extensive work on formalising ability as part of their work on formal theories of agent behaviour (van Linder *et al.*, 1998; van Linder *et al.*, 1999). They follow the tradition of separating ability from opportunity and requiring both for successful execution of actions. In their formal logic executing an action when an opportunity does not exist, results in a counterfactual state of affairs from which no further action is possible. Their work differs from ours in that they are concerned with a finer level of granularity and are concerned to be able to reason about logical composition of ability. We assume composite capabilities to be represented directly - a plan to achieve ϕ indicates a capability for ϕ and we assume that the plan does not include sub-tasks for which the agent does not have the capability (though a sub-task may include requesting help from another agent to successfully realise the goal).

There is a fundamental difference in our approach as compared to that of van Linder, van der Hoek and Meyer (and indeed others) in that we do not attempt to build a formalism to allow reasoning about when actions will succeed. Rather our reasoning is about when it is *rational* to commit to a goal, or to commit to a plan as a particular way of achieving a goal. This rationality requires only that there is a reasonable chance of success, not that success is guaranteed. Well engineered plan sets ensure that sub-tasks within a plan (or capability) are matched to an agent's capabilities (or that there is an expectation

that another agent with the appropriate capability will co-operate). If the opportunity exists for execution of the most abstract plan, then the well engineered plan-set will ensure that opportunities are created for the execution of sub-tasks (though exogenous factors may always interfere).

9. Conclusion and Future Work

The formalisation of capabilities and their relationships to beliefs, goals and intentions is a clean extension of the existing BDI theoretical framework. It provides a theoretical basis for adoption of goals which eliminates a current source of mismatch between theory and implemented systems in that the theory allows adoption of goals which there is no way to achieve. Current implemented BDI systems tend to require both ability and opportunity before a commitment can be made. However, we have explored in other work (Thangarajah *et al.*, 2002) the problems that can arise from this, where an agent should commit to achievement of a goal (or at the very least retain the desire) although there is no current opportunity that allows commitment to a particular course of action to achieve the goal.

The implementation of capability modules in JACK is seen as an abstraction of our notion of capability and we note that if reasoning is to be done at the level of capability then it is important that capabilities be self-contained, or that dependencies be explicitly represented so that they can be reasoned about. This becomes important particularly if capabilities are dynamic, or if agents exist in an open system where the existence of other agents with needed capabilities cannot be guaranteed.

Exploration of how an agent's knowledge of other agents' capabilities affects its own goals and intentions requires further work and some modifications to the axioms relating goals to capabilities. This seems to require a framework which allows for beliefs about other agent's capabilities. Goals would then be constrained by a combination of one's own capabilities plus beliefs about other agent's capabilities.

References

- J.M. Bradshaw, S. Dutfeld, P. Benoit, and J.D. Woolley. KAOS: Toward an industrial-strength open agent architecture. In *Software Agents*, Bradshaw, ed, AAAI Press, pages 375–418, 1997.
- M.E. Bratman, D.J. Israel, and M.E. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4(4):349–355, 1988.

- P. Busetta, N. Howden, R. Rönquist, and A. Hodgson. Structuring BDI agents in functional clusters. In *Proceedings of the Sixth International Workshop on Agent Theories, Architectures, and Languages - ATAL 99*, 1999.
- P. Busetta, R. Rönquist, A. Hodgson, and A. Lucas. Jack intelligent agents - components for intelligent agents in Java. In *AgentLink News Letter*, pages 2–5, January 1999.
- P. Cohen and H. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
- C.P. Cross. ‘Can’ and the Logic of Ability. In *Philosophical Studies*, 50, pages 53–64, 1986.
- P.M. Dung. A formal methodology for verifying situated agents. In *Proceedings of the National Conference on Artificial Intelligence - AAAI 98*, pages 637–642, 1998.
- M. Georgeff and F. Ingrand. Decision-making in an embedded reasoning system. In *Proceedings of the International Joint Conference on Artificial Intelligence - IJCAI 89*, pages 972–978, August 1989.
- M. Huber. Jam: A BDI-theoretic mobile agent architecture. In *Proceedings of the Third International Conference on Autonomous Agents - Agents 99*, pages 236–243, Seattle, WA, 1999.
- J. Lee, M. Huber, P.G. Kenny, and E.H. Durfee. UM-PRS: An implementation of the procedural reasoning system for multi-robot applications. In *Proceedings of the Conference on Intelligent Robotics in Field, Factory, Service and Space - CIRFFSS 94*, pages 842–849, Houston, TX, 1994.
- Y. Lesperance, H. Levesque, F. Lin, and R. Scherl. Ability and knowing how in the situation calculus. *Studia Logica*, 66(1):165–186, 2000.
- F. Lin and H. Levesque. What robots can do: robot programs and effective achievability. *Artificial Intelligence*, 101:201–226, 1998.
- M. Lind. Possibilities for Action. *Center for Human-Machine Interaction*, report CHMI-7-2000, 2000.
- S. McCall. Ability as a Species of Possibility, In *The Nature of Human Action*, M. Brand, ed, Glenview, Scott, Foresman and Company , 1970.
- J.J. Meyer., W. van der Hoek, and B. van Linder. A Logical Approach to the Dynamics of Commitments. *Artificial Intelligence*, 113:1–40, 1999.
- R. C. Moore. A Formal Theory of Knowledge and Action In *Formal Theories of the Commonsense World*, J.C. Hobbs and R.C. Moore, ed, Ablex, Norwood, NJ, pages 319–358 1985.
- L. Padgham and P. Lambrix. Agent capabilities: Extending BDI theory. In *Proceedings of Seventeenth National Conference on Artificial Intelligence - AAAI 2000*, pages 68–73, Austin, TX, 2000.
- A. Rao and M. Georgeff. Modeling rational agents within a BDI-architecture. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference - KR 91*, pages 473–484, 1991.
- A. Rao and M. Georgeff. An abstract architecture for rational agents. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference - KR 92*, pages 439–449, 1992.
- A. Rao and M. Georgeff. BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multi-Agent Systems - ICMAS 95*, San Francisco, USA, 1995.
- S. Shapiro, Y. Lesperance, and H.J. Levesque. Goals and Rational Action in the Situation Calculus - A Preliminary Report. In *Notes of the AAAI Fall Symposium*

- on Rational Agency: Concepts, Theories, Models, and Applications*, pages 117–122, 1995.
- Y. Shoham. Agent oriented programming. *Artificial Intelligence*, 60:51–92, 1993.
- M. P. Singh. Know-How. In Anand S. Rao and Michael J. Wooldridge (editors), *Foundations of Rational Agency*, Applied Logic Series, Kluwer, 1999, pages 105–132.
- K. Sycara, M. Klusch, S. Widoff, and J. Lu. Dynamic service matchmaking among agents in open information environments. *SIGMOD Record*, 28(1):47–53, 1999.
- J. Thangarajah, L. Padgham, and J. Harland. Representation and Reasoning for Goals in BDI Agents. *Australian Computer Science Conference*, 2002.
- B. van Linder, W. van der Hoek, and J.J. Meyer. Formalizing abilities and opportunities of agents. *Fundamenta Informaticae*, 34:53–101, 1998.
- M. Wooldridge. *Reasoning About Rational Agents*. The MIT Press, 2000.