

## Agent Capabilities: Extending BDI Theory

Lin Padgham<sup>1</sup> and Patrick Lambrix<sup>2</sup>

<sup>1</sup>RMIT University, Melbourne, Australia

<sup>2</sup>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 one of these systems as a software engineering mechanism to support modularity and reusability while still allowing meta-level reasoning. This paper presents a formalisation 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.

### 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) and Shoham (Shoham 1993). 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. We understand capability as the ability to

react rationally towards achieving a particular goal. Depending on circumstances a capability may not always result in an achievable plan for realising the goal, but it is a pre-requisite for such.

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 to achieve 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 recently reported 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 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.

### 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 subtasks on the way to achieving a goal. Each plan is associated with a triggering event (which may be an event of type *achieve goal X*). 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 subgoals which allow new plans to flesh out the detail of the plan, calls to external “actions”, or other code in the plan or host language.

We understand having a *capability to achieve X* as mean-

ing that the agent has at least one plan that has as its trigger the goal event *achieve X*. That is the agent has at least one way it knows how to achieve X in some situation. At any given time the agent may be unable to actually use this plan (depending on whether its pre-conditions match the state of the world), but having some such plan is clearly a pre-requisite to being able to achieve X.<sup>1</sup>

In the description of the implementation of capabilities in JACK (Busetta *et al.* 1999a) a capability 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. The interface is specified partially in terms of what events generated external to the capability, can be handled by this capability. Thus a part of the interface to a capability will be a list of the goal achievement events that the capability is designed to handle. Additional subgoal events and the plans that deal with these can be hidden within the internals of the capability. The interface also specifies what events generated by the capability 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* may contain a set of plans to construct a schedule in a certain domain. The knowledge base fragment defined as part of this capability 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 responds to while the only events it generates that are seen externally are events which notify the schedule or which notify failure to generate a schedule.

It is easy to see how this abstraction of a set of plans into a capability could be used to advantage in finding plans that are a possibility for responding to a specific event. Rather than examining all plans it is only necessary to look within the plans of either the generating capability, or a capability that has the relevant event specified as part of its external interface. Naturally this relies on appropriate software engineering design and will preclude the system “discovering” a plan within the internals of a capability that achieves a goal that is not specified as part of the interface of the capability. This is consistent with the practical reasoning approach inherent in these systems which relies on forward chaining based on specified triggers (combined with the ability to manage failure and retry alternative mechanisms to achieve goals and subgoals). The abstraction of sets of plans into capabilities also provides a mechanism for name scoping which is a practical help in building large and complex systems.

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

---

<sup>1</sup>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 all BDI systems of which we are aware are implemented, and is part of the mechanism which allows for efficient practical reasoning.

the expectation is that their behaviour also changes. One way to achieve this could be to use capabilities. A capability 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 subset of the set of *desires*, 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 lays the initial foundation for addressing some of these issues. 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.

## Semantics of R&G BDI Logic

The logic developed by Rao and Georgeff<sup>2</sup> (e.g. (Rao and Georgeff 1991; 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

---

<sup>2</sup>Due to space limitations we are unable to fully define R&G’s logic here, though we attempt to give the basic idea. The reader is referred to (Rao and Georgeff 1991) for full formal definitions.

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.

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  $<_{strong}$  and modal operators are not closed under implication with respect to a weaker modality, making formulae such as:

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

satisfiable. 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, whereas *path formulae* are evaluated over a path in a time-tree.<sup>3</sup> The modal operator *optional* is said to be true of a path formula  $\theta$  at a particular point in a time-tree if  $\theta$  is true of at least one path emanating from that point. The operator *inevitable* is said to be true of a path formula  $\theta$  at a particular point in a time-tree if  $\theta$  is true of all paths emanating from that point. The logic also includes the standard temporal operators  $\bigcirc$  (next),  $\diamond$  (eventually),  $\square$  (always) and  $\bigcup$  (until) which operate over path formulae.

Figure 1 illustrates evaluation of some formulae in a belief, goal or intention world (i.e. a time-tree).

A belief  $\alpha$ , (written  $\text{BEL}(\alpha)$ ) implies that  $\alpha$  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 axiomatisation for beliefs is the standard weak-S5 (or KD45) modal system. For goals and intentions the D and K axioms are adopted.

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

The key axioms of what Rao and Georgeff refer to as the *basic I-system* (Rao and Georgeff 1991) are as follows<sup>4</sup>

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

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

<sup>3</sup>See (Rao and Georgeff 1991) for definitions of state and path formulae.

<sup>4</sup>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).

$$\text{AI3 } \text{INTEND}(\text{does}(\text{a})) \supset \text{does}(\text{a})$$

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

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

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

$$\text{AI7 } \text{done}(\text{a}) \supset \text{BEL}(\text{done}(\text{a}))$$

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

This 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 intentionally, and as long as nothing happens that is inconsistent with the agent's expectations, then these agents will eventually achieve their goals.

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

The *IC-system* requires capability-accessible worlds exactly analogous<sup>5</sup> to belief-accessible worlds, goal-accessible worlds and intention-accessible worlds.  $\text{CAP}(\phi)$  is then defined as being true if it is true in all the capability-accessible worlds. If  $\mathcal{C}$  is the accessibility relation with respect to capabilities, then

$$M, v, w_t \models \text{CAP}(\phi) \text{ iff } \forall w' \in \mathcal{C}_t^w: M, v, w_t \models \phi^6$$

We adopt the D and K axioms for capabilities, i.e. capabilities are closed under implication and consistent.

## 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*.

### Belief-Capability Compatibility:

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

$$\text{AIC1 } \text{CAP}(\alpha) \supset \text{BEL}(\alpha)$$

<sup>5</sup>It is also possible to have a variant where capability-accessible worlds are also required to always be sub-worlds of belief-accessible worlds. This variant and its ramifications are considered in a longer version of this paper which will be available as an RMIT technical report.

<sup>6</sup>All the details of the supporting framework are not given here due to space limitations, but follow straightforwardly from (Rao and Georgeff 1991).

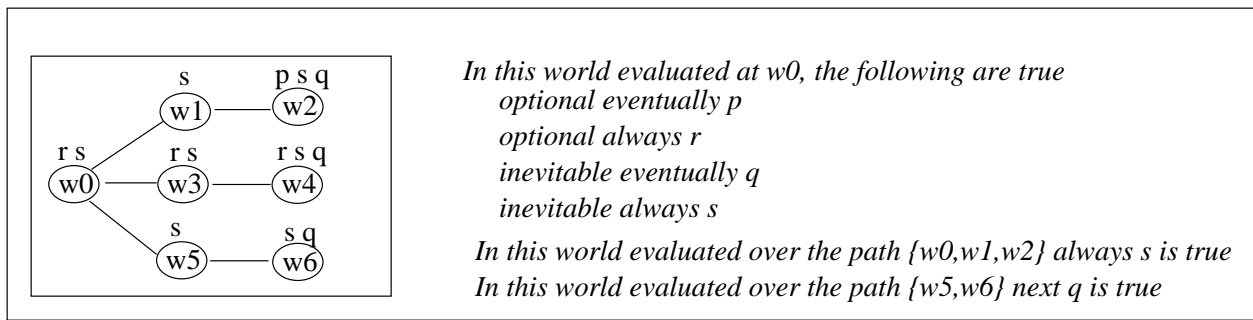


Figure 1: Diagram illustrating evaluation of formulae in a world.

Thus if an agent has the capability that *optional*( $\psi$ ) is true, this also implies a belief that *optional*( $\psi$ ) is true. This should not be read as having a capability to achieve X implies that I believe X is true. The natural language semantics is closer to the statement that if I have a capability to achieve X (at time t), then I believe that it is possible for X to be true (at time t). Statements where  $\alpha$  is a simple predicate rather than a formula involving *optional* must be evaluated at a particular time point. So  $\text{CAP}(\text{rich}) \supset \text{BEL}(\text{rich})$  means that if I am capable of being rich now then I believe I am rich now. Importantly it does not mean that if I have a capability of being rich in the future, I believe that I am rich in the future - I believe only that there is some possible future where I am rich. We 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  $\text{CAP}(\text{rich}) \supset \text{BEL}(\text{rich})$  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:<sup>7</sup>

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

#### Capability-Goal Compatibility

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

**AIC2**  $\text{GOAL}(\alpha) \supset \text{CAP}(\alpha)$

**CIC2**  $\forall w' \in \mathcal{C}_t^w, \exists w'' \in \mathcal{G}_t^w$  such that  $w'' \sqsubseteq w'$ .

#### 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

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

actually follows from AI1 and AI6.

#### Beliefs about Capabilities

If the agent has a capability  $\alpha$  then it believes that it has a capability  $\alpha$ .

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

**CIC3**<sup>8</sup>  $\forall w' \in \mathcal{B}_t^w, \forall w'' \in \mathcal{C}_t^{w'}$  we have  $w'' \in \mathcal{C}_t^w$

#### Capabilities regarding Goals

If an agent has a goal  $\alpha$  then it has the capability to have the goal  $\alpha$ .

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

**CIC4**  $\forall w' \in \mathcal{C}_t^w, \forall w'' \in \mathcal{G}_t^{w'}$  we have  $w'' \in \mathcal{G}_t^w$

#### Capabilities regarding Intentions

If an agent has an intention  $\alpha$  it also has the capability to have the intention  $\alpha$ .

#### Follows from AIC2 and AI6

**INTEND**( $\alpha$ )  $\supset$  **CAP**(**INTEND**( $\alpha$ ))

semantic condition:

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

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  $\text{CAP}(\alpha) \equiv \text{BEL}(\text{CAP}(\alpha))$ . As mentioned in (Rao and Georgeff 1991) this has the effect of collapsing mixed nested modalities to their simpler non-nested forms.

We will refer to the axioms AI2, AI3, AI6, AI7, 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.<sup>9</sup>

#### Commitment Axioms

Rao and Georgeff define three variants of a commitment axiom, which taken together with the basic axioms define

<sup>8</sup>This (and CIC4) is subtly different from the analogue of what is in (Rao and Georgeff 1991), which appears to be slightly wrong. The explanation with proof and counter-example for the original formulation will be given in the full paper.

<sup>9</sup>AI1 follows from AIC1 and AIC2. AI4 follows from AIC1, AIC2 and AI6. AI5 follows from AIC1 and AIC4.

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 to achieve that intention.

The *self-aware agent* is defined by the *basic IC-system* plus the following axiom which we call AIC9d.<sup>10</sup>

**AIC9d**  $\text{INTEND}(\text{inevitable} \diamond \phi) \supset$   
 $\text{inevitable}(\text{INTEND}(\text{inevitable} \diamond \phi))$   
 $\bigcup (\text{BEL}(\phi) \vee \neg \text{CAP}(\text{optional} \diamond \phi))$

It is then possible to extend theorem 1 in (Rao and Georgeff 1991) 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.<sup>11</sup> Self-awareness can be combined with either open-mindedness or single-mindedness to obtain self-aware-open-minded and self-aware-single-minded agents.

### Properties of the Logic

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

**Theorem 1** For modalities  $R_1$  and  $R_2$  such that  $R_1 <_{\text{strong}} R_2$ , the following formulae are satisfiable:

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

**Proof:** We prove the result for BEL and CAP. The proof for the other pairs of modalities is similar. Assume  $\text{BEL}(\phi)$ . Then,  $\phi$  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  $\phi$  is not true in one of these worlds, then  $\phi$  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 to do  $\phi$ , believe that  $\phi$  implies  $\gamma$ , but not have the capability to do  $\gamma$ .<sup>12</sup>

<sup>10</sup>This numbering is chosen because of the relationship of AIC9d to AI9a, AI9b, and AI9c in the original *I-system*.

<sup>11</sup>These theorems and proofs are not shown here due to space restrictions. They are available in the longer version of the paper.

<sup>12</sup>The alternative formulation referred to in footnote 5 does not have this property with respect to capabilities.

**Theorem 2** For modalities  $R_1$  and  $R_2$  such that  $R_1 <_{\text{strong}} R_2$ , 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,  $\phi$  is true in every capability-accessible world. To be able to infer that  $\gamma$  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  $\gamma$  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 R&G BDI logic and allow definition of further interesting types of agents. We look now 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.

### Implementation aspects

An abstraction of a BDI-interpreter which follows the logic of the basic *I-system* is given in (Rao and Georgeff 1992).<sup>13</sup> The first stages in the cycle of this abstract interpreter are to generate and select plan options. These are filtered by beliefs, goals and current intentions. Capabilities now provide an additional filter on the options we generate and select. Similarly capabilities must be considered when dropping beliefs, goals and intentions. In a system with dynamic roles capabilities themselves may also be dropped. Thus we obtain this slightly modified version of the interpreter in (Rao and Georgeff 1992).

#### BDI with capabilities interpreter:

```
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.
```

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 def-

<sup>13</sup>Due to lack of space we cannot give more than the most basic summary here of this interpreter and its relation to the logic.

initions of the various kinds of agents (blindly committed, single-minded, open-minded and self-aware) has to 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 to do  $\phi$ , believe that  $\phi$  implies  $\gamma$ , but not have the capability to do  $\gamma$  (see before), 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<sup>14</sup> but would have an impact on how the option generation and selection phases of the abstract interpreter work.

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:

BEL(*inevitable*  $\Box$ (rain)  $\supset$  (garden-wet))

BEL(*optional*  $\Diamond$ (rain))

In the R&G 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(optional  $\Diamond$ (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*.

## Conclusion and Future Work

The formalisation of capabilities and their relationships to beliefs, goals and intentions is a clean extension of an existing theoretical framework. Advantages of the extension include eliminating mismatch between theory and what happens in actual systems, better mapping of theory to intuition, indication of areas for development of implemented reasoning in line with the theory and highlighting of issues for consideration in actual implementations.

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.

<sup>14</sup>The necessary modification is essentially to require that all capability-accessible worlds are sub-worlds of belief-accessible worlds. However this breaks the symmetry of the current formalisation where capability accessible worlds are exactly analogous to belief/goal/intention accessible worlds.

Goals would then be constrained by a combination of one’s own capabilities plus beliefs about other agent’s capabilities.

## References

- 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.
- 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.
- 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.
- Y. Shoham. Agent oriented programming. *Artificial Intelligence*, 60:51–92, 1993.