Intelligent Agents: Issues and Logics[*]

John-Jules Meyer
Utrecht University
Institute of Information and Computing Sciences
Intelligent Systems Group
P.O. Box 80.089
3508 TB Utrecht, The Netherlands

May 9, 2002

Abstract

In this paper we review some issues of research in intelligent agents, and particularly some logical theories that have been proposed in the literature to describe aspects of intelligent agents.

1 Introduction

Intelligent agents have become the subject of much research lately. The field of agent technology and its foundations lies in between the disciplines of artificial intelligence and mainstream computer science, in particular that of software engineering. Agents are pieces of software (or hardware) that display a certain degree of autonomy. They are not completely reliant on the user's commands, but are able to react to the environment in an autonomous manner, and, even more remarkable, they will take initiative on their own to perform actions to influence their environment in a certain desirable way (goal-directed or proactive behaviour). Agents are thus very much 'situated' in their environment: they are able to both sense ('perceive') and affect the environment they inhabit. Therefore, agents can be viewed as successors to the traditional knowledge-based or expert systems which could reason with symbolic representations of some universe of discourse (or knowledge domain), and could render expert-level advise on problems related to that domain. These expert systems thus acted as a kind of artificial consultants, but could take no actions on their own, and did not affect an inhabited environment.

On the other hand, programming (in terms of) agents, called agent-oriented programming may also be viewed as a successor to the now

[*] This tutorial is partially supported by the SURF-project "Landelijk Onderwijsweb Kennistechnologie (LOK)"

popular object-oriented (OO) programming paradigm in software engineering. Here one must think typically of designing systems of multiple agents (so-called *multi-agent systems (MAS)*) that communicate in order to cooperate in some sensible way (for instance to solve a task together). Although objects in the OO setting already possess some kind of autonomy (they have their own data types and methods that can be called by other objects), there is no mention of any autonomy in the sense of displaying initiative or being proactive. In agent-oriented programming these notions are central, and, moreover, communication among agents is much less a matter of just invoking a method of another agent, but rather asking questions to other agents which these other agents may (or may not) handle in their own way. Also the content of communication (the messages itself) seems to be different in the sense of higher-level than in OO programming: in MAS's communication may take place involving the typical agent-related notions of beliefs, desires and goals.

Agent notions such as autonomy, reactiveness and proactiveness are often coined by means of the notion of a 'mental state'. Such a mental state comprises the attitudes of an agent: informational ones, dealing with knowledge and belief (updating and revising these as new information comes available, thus including reasoning and learning capabilities), and motivational ones, dealing with wishes / desires, goals, intentions, commitments.

2 Issues

In order to realise agents one should address several aspects at different levels. In research on agents the following areas are distinguished: theories, architectures and (agent programming) languages ([41]).

Theories concern descriptions of agents, in particular their behaviour, often in terms of their informational and motivational attitudes. Most of these propose (modal) logics for reasoning about these attitudes.

Architectures pertain to more or less generic organisation schemes of how agents are (to be) built, varying from abstract pictures with indications of some essential components to concrete system descriptions.

Languages concern more or less dedicated programming languages designed to program agents in terms of agent concepts (or, using the agent metaphor).

Of course, there are many interrelations between theories, architectues and languages. For example, an agent programming language may be designed using the concepts of some particular agent theory, and may employ or realise some particular agent architecture.

Moreover one has to distinguish the 'micro-level' from the 'macro-level': the former pertains to the internal structure of an agent, whereas with the latter the 'societal' or 'interagent' level is referred to. Both on micro- and macrolevel one can discuss theories, architectural issues and issues with respect to languages. For example, with respect to the internal ('micro'-) level one has the well-known BDI theory of Rao & Georgeff [31], a logic to reason about the agent's beliefs, desires and intentions. Other logical theories for single agents include the approach of Cohen & Levesque ([2])[1] and the KARO logic ([18]. BDI theory has given rise to an architecture based on the BDI notions, called the BDI architecture. As to languages for single agents there have been several proposals like AgentSpeak(L) ([30]), GOLOG ([23]) and 3APL[2] ([16]).

On the other hand, with respect to the societal ('macro'-) level one has theories about societal behaviour, including aspects of communication, coordination but also norms and obligations etc. Also logical theories concerning mutual and common beliefs, intentions etc. (the multi-agent versions of BDI plus typical social notions). There is also the very interesting but somewhat elusive issue of so-called *emergent* behaviour, constituting behaviour of a multi-agent system (typically with very many agents) that emerges from the interactions between the individual agents, and cannot be predicted from the behaviour from the individual agents. This topic is being investigated extensively. A logical approach to MAS is proposed by e.g. Singh [34]. As to architectures for MAS we mention the InteRRap architecture [29] which includes a layer for handling the communication between agents. There are also several agent programming languages for MAS: even the first agent language AGENT0 [37] had already communication elements, and the language CONGOLOG [13] is an extension of the language GOLOG to handle concurrency. Concurrent METATEM [11] is a language for programming MAS's based on (executable) temporal logic. Finally we must mention dedicated agent communication languages (ACLs) to program communication between agents such as KQML [10] and FIPA-ACL [12].

One now obtains the following scheme:

	theories	architectures	languages
micro	BDI theory / logic, C & L , KARO	BDI arch.	(single) AOP (AgentSpeak(L), GOLOG, 3APL)
macro	societal behaviour (communication, coordination, norms, ...), emergent behaviour	InteRRap	multi-agent AOP (AGENT0, CONGOLOG, Concurrent METATEM, ACLs (KQML, FIPA))

A description of research on architectures and programming languages is beyond the scope of this paper. We now concentrate on theories, and particularly the logics that are involved in these. (As a consequence we will not say much about many more advanced aspects of societal behaviour,

[1] Although, as we shall see, in this theory modalities are indexed by agents, we nevertheless treat it as a *single* agent theory, since the relation between the modalities for different agents is not studied.

[2] Although some work has been done on multi-agent aspects of 3APL such as communication, cf. [15], we treat it here as a *single* agent language.

such as emergent behaviour, since – although it is being studied extensively – as far as I know – until now, but perhaps necessarily so by its very nature! – it eludes the use of logic for a proper treatment.)

In order to describe (the attitudes of) agents one may resort to logics which are tailored to express the notions above, such as knowledge, belief, desires, goals, etc. Here modal logic comes in as it has been the traditional tool in philosophy to analyse these notions in a systematic and formal, rigorous manner.

We begin with the modal logic(s) of knowledge and belief.

3 Epistemic and Doxastic Logic

Epistemic (doxastic) logic is the logic of knowledge (belief). It is a modal logic with modal operator **K** (or **B**) indicating that the formula that it is given as an argument is known (believed). Stemming from philosophy epistemic and doxastic logic have been adopted by computer scientists and AI researchers in the 1980's in order to describe aspects of knowledge appearing in distributed and knowledge-based computer systems ([9, 27]).

Formally, epistemic logic is treated as follows. The language is obtained by taking classical (propositional) logic augmented by a clause for the knowledge or belief operator. We assume a set $\mathcal{P}$ of atomic propositions.

Definition 3.1 *Language of epistemic / doxastic formulas.*

- *every atomic formula in $\mathcal{P}$ is an epistemic (doxastic) formula*
- *if φ_1 and φ_2 are epistemic (doxastic) formulas, then $\neg\varphi_1, \varphi_1 \vee \varphi_2$ are epistemic (doxastic) formulas*
- *if φ is an epistemic (doxastic) formula, then $\mathbf{K}\varphi$ $(\mathbf{B}\varphi)$ is an epistemic (doxastic) formula*

Other propositional connectives (such as $\wedge, \rightarrow, \leftrightarrow$) are introduced as (the usual) abbreviations.

Definition 3.2 *Models for epistemic and doxastic logic are usually taken to be Kripke structures of the form $\mathcal{M} = \langle W, \vartheta, R\rangle$, where:*

- *W is a non-empty set of states (or worlds)*
- *ϑ is a truth assignment function per state*
- *R is an accessibility relation on W for interpreting the modal operator **K** or **B**. In the former case it is assumed to be an equivalence relation, while for the latter it is assumed to be euclidean, transitive and serial.*

Remark 3.3 *The set of states (worlds) that are accessible from a certain state (world) must be viewed as epistemic alternatives for this world: if the agent is in this state he is not able to distinguish these accessible worlds due to his (lack of) knowledge/belief on the true nature of his state: as far he is concerned he could be in any of the alternatives.*

The reason that for modelling knowledge the accessibility relation is taken to be an equivalence relation, can be understood as follows: the agent, being in a state, considers a set of alternatives which contains the state he is in (so the agent considers his true state as an alternative) and which are all alternatives of each other.

For belief this would be too strong: in particular, for belief it is not reasonable to assume that the agent always considers his true state as an alternative, since he may be mistaken. So, for belief, weaker assumptions are assumed, which nevertheless result in a number of interesting validities below.

Definition 3.4 *(Interpretation of epistemic / doxastic formulas.) In order to determine whether an epistemic (doxastic) formula is true in a model/state pair $\mathcal{M}, w$ (if so, we write $\mathcal{M}, w \models \varphi$), we stipulate:*

- $\mathcal{M}, w \models p$ *iff* $\vartheta(w)(p) = \text{true}$, *for* $p \in \mathcal{P}$
- *The logical connectives are interpreted as usual.*
- $\mathcal{M}, w \models \mathbf{K}\varphi(\mathbf{B}\varphi)$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $R(w, w')$

Remark 3.5 *The last clause can be understood as follows: an agent knows (believes) a formula to be true if the formula is true in all the epistemic alternatives that the agent considers at the state he is in (represented by the accessibility relation).*

Definition 3.6 *(validity.)*

- *Validity of a formula with respect to a model* $\mathcal{M} = \langle W, \vartheta, R \rangle$ *is defined as:* $\mathcal{M} \models \varphi \Leftrightarrow \mathcal{M}, w \models \varphi$ *for all* $w \in \mathcal{M}$.
- *Validity of a formula is defined as validity with respect to all models:* $\models \varphi \Leftrightarrow \mathcal{M} \models \varphi$ *for all models* $\mathcal{M}$ *of the form considered.*

Validities in epistemic logic with respect to the given models (which we will refer to the 'axioms' of knowledge) are:

Proposition 3.7

- $\models \mathbf{K}(\varphi \to \psi) \to (\mathbf{K}\varphi \to \mathbf{K}\psi)$
- $\models \mathbf{K}\varphi \to \varphi$
- $\models \mathbf{K}\varphi \to \mathbf{K}\mathbf{K}\varphi$
- $\models \neg\mathbf{K}\varphi \to \mathbf{K}\neg\mathbf{K}\varphi$

Remark 3.8 *The first axiom says that knowledge is closed under implication: if both the implication $\varphi \to \psi$ and the antecedent φ is known then also the conclusion ψ is known. This is of course a very 'idealised' property of knowledge, but its validity is at the very heart of using so-called normal modal logic as we do here. (If one wants to deny this property, one has to resort to 'nonstandard' approaches, cf. [27].) The second axiom expresses that knowledge is true. (One cannot honestly, truthfully and justifiably state to* know *something that is false.) The third and fourth axioms express a form of introspection: the agent knows what it knows, in* *the sense that it knows that it knows something (the second axiom), and, moreover, it knows what it does* not *know (the third axiom). Of course, this may be very unrealistic to assume for some intelligent agents, such as humans, but often it makes sense to assume it in the case of artificial agents, either by virtue of their finitary nature or by way of some idealisation. In any case it makes life easier, since the resulting logic, called* **S5**, *is very elegant (has relatively simple models) and enjoys several pleasant properties ([27]).*

With respect to doxastic logic we obtain the following validities:

Proposition 3.9

- $\models \mathbf{B}(\varphi \to \psi) \to (\mathbf{B}\varphi \to \mathbf{B}\psi)$
- $\models \neg\mathbf{B}\mathtt{ff}$
- $\models \mathbf{B}\varphi \to \mathbf{B}\mathbf{B}\varphi$
- $\models \neg\mathbf{B}\varphi \to \mathbf{B}\neg\mathbf{B}\varphi$

Remark 3.10 *Again we observe the introspection properties, but the second axiom now states that an agent's belief is not inconsistent, which is weaker than the property that belief should be true. Also note the first axiom which states that also belief is closed under implication, which may be regarded as even 'more idealised' a property than for knowledge! (Again see [27] for alternatives.)*

One may wonder whether the knowledge and belief modalities are interrelated in some meaningful way. Although in the literature (for example [21, 17, 38, 39]), and indeed also in several versions (e.g. in [24], Chapter 5) of the richer KARO logic, which we will encounter in the sequel, several interesting possibilities for such an interaction have been investigated, we will assume in this paper only the natural (but see [38, 39]) property that knowledge implies belief: $\mathbf{K}\varphi \to \mathbf{B}\varphi$.

4 Desires and Intentions

Besides knowledge and belief there are several other modalities which may be of interest in the context of (multi-) agent systems. The first that comes to mind perhaps is that agents may also be endowed with *desires* that motivate them to perform actions. The philosopher Bratman has argued that to capture the essence of intelligent agents one must go a step beyond this: also *intentions* have to be included in the description of the mental state of an agent. Intentions must be viewed as wishes that are committed to by the agent ([1]). This is important, Bratman argues, for coherent behaviour. For example, consider an agent that wishes to prepare eggs for a meal. Suppose it desires both a hard-boiled egg and a scrambled, fried one, but it has only the disposal of one egg. Then it has to make a choice: either boil it or scramble and fry it, and moreover, if it has made this choice it determines plans for the future to realise its wishes, and it is important to stay committed to the choice(s) made earlier. For example, it makes little sense to first boil the egg and then try to scramble it, or first scramble it and then boil it. Once one of the two wishes has been selected (committed to) the agent must stick to the realisation of it without switching to the other wish at a moment that this cannot be realised any more! Thus, intentions provide what Bratman calls a "*screen of admissibility*" for adopting other intentions.

In the following sections we discuss three logics in which the notion of intention is formalised, but before we go into the details of these rather complicated logics where also notions of actions and/or time play a role, we first consider motivational modalities such as desires and intentions in isolation.

In principle we could extend the epistemic (or doxastic) logic from the previous section to a multi-modal logic of, say, belief, desires and intention by adding modal operators **D** and **I** for desires and intentions, respectively. We would then obtain a rather simple extension of the logical framework of the previous framework, which we shall call a 'bdi' logic for the moment (for belief, desires and intentions):

Definition 4.1 *Language of bdi formulas.*

- *every atomic formula in $\mathcal{P}$ is a bdi formula*
- *if φ_1 and φ_2 are bdi formulas, then $\neg\varphi_1, \varphi_1 \vee \varphi_2$ are bdi formulas*
- *if φ is a bdi formula, then $\mathbf{B}\varphi, \mathbf{D}\varphi, \mathbf{I}\varphi$ is a bdi formula*

Models for these formulas are simple extensions of the models for doxastic logic:

Definition 4.2 *Models for bdi logic are Kripke structures of the form $\mathcal{M} = \langle W, \vartheta, B, D, I\rangle$, where:*

- *W is a non-empty set of states (or worlds)*
- *ϑ is a truth assignment function per state*
- *B, D, I are accessibility relations on W for interpreting the modal operators $\mathbf{B}, \mathbf{D}$, and $\mathbf{I}$, respectively. B is assumed to be euclidean, transitive and serial, and I is assumed to be serial.*

Remark 4.3 *The idea is the same as with epistemic / doxastic logic. The only thing is that now we have three accessibility relations on the set of states. For the belief-related one the interpretation is as before; for the desire-related one the accessibility relation points at states that are 'desired' (from the perspective of the current state of evaluation, while the intention-related accessibility relation yields states that are 'intended' alternatives of the current state / world. The constraint on B is the same as before (for the same reasons as before), whereas the constraints on D, I are much weaker: for I it is assumed that the relation is serial, that is, there is always at least one 'intended alternative' state (resulting in the property below that intentions are not inconsistent), while for D we do not require any constraint (which implies that desires may even be inconsistent).*

Given these models, the interpretation of bdi formulas is as expected:

Definition 4.4 *(Interpretation of bdi formulas.) In order to determine whether a bdi formula is true in a model/state pair $\mathcal{M}, w$ (if so, we write $\mathcal{M}, w \models \varphi$), we stipulate:*

- $\mathcal{M}, w \models p$ *iff* $\vartheta(w)(p) = \text{true}$, *for* $p \in \mathcal{P}$
- *The logical connectives are interpreted as usual.*
- $\mathcal{M}, w \models \mathbf{B}\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $B(w, w')$
- $\mathcal{M}, w \models \mathbf{D}\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $D(w, w')$
- $\mathcal{M}, w \models \mathbf{I}\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $I(w, w')$

Remark 4.5 *For belief this exactly as before; for desire and intention it is completely analogous: e.g. something (expressed by some formula) is desired (at some state) if this formula holds in all the desired alternative states of that state.*

Using the same definition of validity as before, we thus obtain the same properties for belief as before, and the following for desire and belief:

Proposition 4.6

- $\models \mathbf{B}(\varphi \to \psi) \to (\mathbf{B}\varphi \to \mathbf{B}\psi)$
- $\models \neg\mathbf{B}\mathtt{ff}$
- $\models \mathbf{B}\varphi \to \mathbf{B}\mathbf{B}\varphi$
- $\models \neg\mathbf{B}\varphi \to \mathbf{B}\neg\mathbf{B}\varphi$
- $\models \mathbf{D}(\varphi \to \psi) \to (\mathbf{D}\varphi \to \mathbf{D}\psi)$
- $\models \mathbf{I}(\varphi \to \psi) \to (\mathbf{I}\varphi \to \mathbf{I}\psi)$
- $\models \neg\mathbf{I}\mathtt{ff}$

Remark 4.7 *The properties for desire and intention are rather weak. This is due to the weak constraints we have put on the accessibility relations for these modal operators. We observe that intentions are not inconsistent, while both for intention and desire we have closure under implications again (which may again be viewed as idealisations for rational agents).*

In order to really describe the informational and motivational attitudes of agents the present bdi logic is too little expressive. In fact, the very meaning of agent is 'acting entity' (from the Latin '*agere*'). So we need to incorporate the notion of action is some way. In the following approaches proposed in the literature this is done in different ways. Another thing that is not treated in the above bdi logic, is whether the notions of belief, desire and intention are related in some way, and if so, how. Also on this issue we will see possible (and really distinct!) answers in the approaches we will treat next. We start with the approach by Cohen and Levesque.

5 Cohen and Levesque's Logic of Intention

Cohen & Levesque, in an influential paper [2] on this subject, give a formal analysis of the notion of intention. Their setting is a modal logic with operators for belief and goals, with a possibility to express the performance of actions, which gives the logic the flavour of a linear-time temporal logic with extra modalities. In this framework they define intentions as certain (persistent) goals. In fact, the formalism is 'tiered' in the sense that it contains an 'atomic layer' describing beliefs, goals and actions of an agent, and a 'molecular layer' in which concepts like intention are defined in terms of the primitives of the atomic layer.

Thus, Cohen & Levesque's definition of intention amounts to:

$$intention = choice + commitment$$

Formally, Cohen and Levesque define a logical language by means of the following primitive operators:

We assume a set $\mathcal{A}$ of atomic actions (or rather atomic action expressions) with typical elements a, b, a set Ag of agent names with typical elements i, j, and a set $Pred$ of 'atomic' predicate formulas $p = P(t_1, ..., t_n)$ where P is a predicate symbol and the t_i are terms, over a certain fixed signature. The set of well-formed formulas with typical elements φ, ψ will be defined below. We also assume a set Var of variables, with typical elements x, y, which includes the sets $\mathcal{A}$ and Ag. (So both the atomic actions and agent names are considered to be variables so that quantification over these in the language below becomes possible.) In the sequel $\mathcal{Z}$ denotes the set of integers.

Definition 5.1 *(Operators.)*

- *$HAPPENS\ \alpha$: action α happens* next
- *$DONE\ \alpha$: action α has* just *happened*
- *$AGT\ i\ a$: agent i is the* only *agent of atomic action a*
- *$BEL\ i\ \varphi$: formula φ follows from i's beliefs*
- *$GOAL\ i\ \varphi$: formula φ follows from i's goals*
- *$a \leq b$: action a is an initial subsequence of b*

Remark 5.2 *As Cohen & Levesque's logic is primarily (based on) a temporal logic, and not a logic of action such as dynamic logic, additional operators are needed to reason about actions in some way. Here we see that this is done by especially the $HAPPENS$ and $DONE$ operators, expressing (immediate) future and past action executions, respectively. The AGT operator expresses the 'actor' of an atomic action. Next the familar 'belief' and 'goal' operators are introduced, and finally there is an operator $\leq$ to compare actions as to intial parts (which will be based on the semantics of these actions, where atomic actions are to be interpreted as event sequences, see below).*

Definition 5.3 *(Language.) The set L_{CL} of well-formed formulas is given by the BNF grammar:*

$$\begin{array}{rcl}
\varphi & ::= & p\ (\in Pred) \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \ldots \mid \exists x\varphi \mid \\
& & HAPPENS\ \alpha \mid DONE\ \alpha \mid AGT\ i\ a \mid \\
& & BEL\ i\ \varphi \mid GOAL\ i\ \varphi \mid a \leq b \\
\alpha & ::= & a(\in \mathcal{A}) \mid \varphi? \mid \alpha_1;\alpha_2 \mid \alpha_1 + \alpha_2 \mid \alpha^*
\end{array}$$

Remark 5.4 *On the dots other familiar operators can be added like $\vee$ and $\forall$. As these may also be introduced by means of the usual abbreviations, we omit these here.*

The following abbreviations are used:

- $\diamond\varphi = \exists x(HAPPENS\ x;\varphi?)$
- $\Box\varphi = \neg\diamond\neg\varphi$
- $LATER\ \varphi = \neg\varphi \wedge \diamond\varphi$
- $HAPPENS\ i\ a = HAPPENS\ a \wedge AGT\ i\ a$
- $BEFORE\ \varphi\ \psi = \forall x(HAPPENS\ x;\psi?) \rightarrow$ $\exists y((y \leq x) \wedge HAPPENS\ y;\varphi?)$
- $KNOW\ i\ \varphi = \varphi \wedge BEL\ i\ \varphi$
- $COMPETENT\ i\ \varphi = (BEL\ i\ \varphi \rightarrow KNOW\ i\ \varphi)$

Definition 5.5 *(Semantics.) The semantics of the language is given by means of models of the form $\mathcal{M} = \langle \Theta, P, E, Agt, T, B, G, \Phi\rangle$, where*

- *Θ is a set (universe of discourse)*
- *P is a set of agents*
- *E is a set of primitive event types, or events, for short.*
- *$Agt \in [E \rightarrow P]$ specifies the agent of an event*
- *$T \subseteq [\mathcal{Z} \rightarrow E]$: a set of possible worlds (event sequence)*
- *$B \subseteq T \times P \times \mathcal{Z} \times T$ is the belief accessibility relation*
- *$G \subseteq T \times P \times \mathcal{Z} \times T$ is the goal accessibility relation*
- *Φ interprets predicate symbols*

Remark 5.6 *As we have a first-order base logic, a universe of discourse (or domain) Θ and an interpretation function Φ are needed. We will abuse language by writing $\Phi(p)$ if the atomic predicate formula p is true with respect to interpretation Φ. Furthermore, agents are used for interpreting agent names, and events are employed for the interpretation of atomic actions, which are only names (variables). Note that in general it is allowed to map an atomic action to a* sequence *of events. The function Agt yields the agent involved in an event. The set T is a set of possible worlds, which in this approach take the form of event sequences (courses of events). B and G are the usual belief and goal accessibility relations, here indexed by agents and time point. So, for example, $B(\sigma, i, n, \sigma')$ (which in the sequel we shall write as $\langle\sigma, n\rangle B[i]\sigma'$) denotes that the world σ and σ' are belief-related with respect to agent i and time n.*

In the following we refer to the domain D as the set $\Theta \cup P \cup E^*$, where E^* stands for the set of event sequences. We use v for a set of bindings of variables to objects, which is thus a function of type $Var \rightarrow D$, of course respecting the matching of the types of variables and values. (Note that in general we may have that $v(a) \in E^*$ for atomic action name (variable) a.) The function Agt is extended to $E^* \rightarrow 2^P$ by defining $Agt[e_1, \ldots, e_n] = \{Agt(e_i) \mid 1 \leq i \leq n\}$.

Definition 5.7 *(Interpretation of formulas.) The interpretation of formulas w.r.t. a model $\mathcal{M}$, a world σ, a time point n, and a set v of bindings of variables to objects in D. is now defined by:*

- $\mathcal{M}, \sigma, v, n \models p \Leftrightarrow \Phi(p)$ *is true, for* $p \in Pred$
- *The interpretations of the propositional connectives are defined as usual*
- $\mathcal{M}, \sigma, v, n \models \exists x \varphi \Leftrightarrow \mathcal{M}, \sigma, v\{d/x\}, n \models \varphi$ *for some* $d \in D$, *where* $v\{d/x\}$ *is a set of bindings like* v, *but with* x *bound to* d
- $\mathcal{M}, \sigma, v, n \models HAPPENS\ \alpha \Leftrightarrow$ *exists* $m \geq n$ *such that* $\mathcal{M}, \sigma, v, n[\alpha]m$ *(see the remark below)*
- $\mathcal{M}, \sigma, v, n \models DONE\ \alpha \Leftrightarrow$ *exists* $m \leq n$ *such that* $\mathcal{M}, \sigma, v, m[\alpha]n$ *(see the remark below)*
- $\mathcal{M}, \sigma, v, n \models AGT\ i\ a \Leftrightarrow Agt[v(a)] = \{v(i)\}$
- $\mathcal{M}, \sigma, v, n \models BEL\ i\ \varphi \Leftrightarrow$ *for all* σ^* *with* $\langle \sigma, n \rangle B[v(i)]\sigma^*$: $\mathcal{M}, \sigma^*, v, n \models \varphi$
- $\mathcal{M}, \sigma, v, n \models GOAL\ i\ \varphi \Leftrightarrow$ *for all* σ^* *with* $\langle \sigma, n \rangle G[v(i)]\sigma^*$: $\mathcal{M}, \sigma^*, v, n \models \varphi$
- $\mathcal{M}, \sigma, v, n \models a \leq b \Leftrightarrow v(a)$ *is an initial subsequence of* $v(b)$.

Remark 5.8 *In the above definition* $\mathcal{M}, \sigma^*, v, n[\alpha]m$, *where* $n \leq m$, *means informally that the sequence of events described by the action* α *happens between the time points* n *and* m. *(This notion can be formally defined with induction on* α *[2]. Here we only note that as to the test action we have that* $\mathcal{M}, \sigma, v, n[\varphi?]n$ *iff* $\mathcal{M}, \sigma, v, n \models \varphi$.*) With this in mind, most of the definitions above are straight-forward. The sixth item says that* $AGT\ i\ a$ *is true iff the agent involved in the event denoted by the atomic action* a *is the agent with agent name* i.

In order to have the belief and goal operators behave in the desired way Cohen and Levesque impose the following constraints on their models:

Definition 5.9 *(Constraints on models.)*

- *Consistency: relation* B *is euclidean, transitive and serial; relation* G *is serial*
- *Realism:* $G \subseteq B$: *worlds consistent with what the agent has chosen are not ruled out by his beliefs*

Definition 5.10 *(Satisfiability and validity.)*

1. *A formula $\varphi \in L_{CL}$ is* satisfiable *iff there is a model $\mathcal{M}$, world σ, value assignment v and a time point n such that $\mathcal{M}, \sigma, v, n \models \varphi$.*
2. *A formula $\varphi \in L_{CL}$ is* valid *(denoted $\models \varphi$) iff $\mathcal{M}, \sigma, v, n \models \varphi$ for all $\mathcal{M}, \sigma, v, n$.*

The constraints above has as a consequence that the following are validities in the logic:

Proposition 5.11

- $\models (BEL\ i\ \varphi \wedge BEL\ i\ (\varphi \rightarrow \psi)) \rightarrow BEL\ i\ \psi$
- $\models BEL\ i\ \varphi \rightarrow BEL\ i\ (BEL\ i\ \varphi)$
- $\models \neg BEL\ i\ \varphi \rightarrow BEL\ i\ \neg(BEL\ i\ \varphi))$
- $\models BEL\ i\ \varphi \rightarrow \neg BEL\ i\ \neg\varphi$
- $\models (GOAL\ i\ \varphi \wedge GOAL\ i\ (\varphi \rightarrow \psi)) \rightarrow GOAL\ i\ \psi$
- $\models GOAL\ i\ \varphi \rightarrow \neg GOAL\ i\ \neg\psi$
- $\models BEL\ i\ \varphi \rightarrow GOAL\ i\ \varphi$
- $\models (GOAL\ i\ \varphi \wedge BEL\ i\ (\varphi \rightarrow \psi)) \rightarrow GOAL\ i\ \psi$

Remark 5.12 *These validities are the familar ones from the previous section. (Note that $BEL\ i\ \varphi \rightarrow \neg BEL\ i\ \neg\varphi$ is equivalent with $\neg BEL\mathtt{ff}$, and likewise for GOAL.) The only remarkable new validity is the seventh one (the eighth follows from this one): if one believes something it is also a goal. This may strike one as strange. (In fact we shall encounter the converse in the approach of Rao & Georgeff below!) In Cohen & Levesque's approach it directly follows from the 'realism' constraint on the models. They explain it as follows: if an agent believes something* currently *it is not rational for him to want it to be false* currently*: "agents do not choose what they cannot change". We will see that in other approaches (like the KARO framework) the goal modality has another meaning (viz. a chosen wish to let something be true, not currently, but in the future! In this case the current validity of belief implies goal is not desirable any more.)*

As to the other operators we have, for example:

Proposition 5.13

- $\models HAPPENS\ \alpha;\beta \leftrightarrow HAPPENS\ (\alpha;(HAPPENS\ \beta)?)$
- $\models HAPPENS\ \alpha + \beta \leftrightarrow (HAPPENS\ \alpha \vee HAPPENS\ \beta)$
- $\models \varphi \leftrightarrow (DONE\ \varphi?)$
- $\models \varphi \rightarrow \diamond\varphi$
- $\models \Box(\varphi \rightarrow \psi) \wedge \diamond\varphi \rightarrow \diamond\psi$
- $\models \neg LATER\ \diamond\varphi$
- $\models \diamond\psi \wedge (BEFORE\ \varphi\ \psi) \rightarrow \diamond\varphi$

Next Cohen and Levesque define *achievement goals* by means of the primitives:

Definition 5.14 *(Achievement goals.)*

$$AGOAL\ i\ \varphi = GOAL\ i\ (LATER\ \varphi) \wedge BEL\ i\ \neg\varphi$$

Remark 5.15 *So an achievement goal is something that is desired for the future (it is a goal that it will later be true) but is currently believed to be false.*

Persistent goals are now defined as special achievement goals:

Definition 5.16 *(Persistent goals.)*

$$PGOAL\ i\ \varphi = AGOAL\ i\ \varphi\ \wedge$$

$$[BEFORE(BEL\ i\ \varphi \vee BEL\ i\ \Box\neg\varphi)\neg GOAL\ i\ (LATER\ \varphi)]$$

Remark 5.17 *So a persistent goal is an achievement goal that the agent will not give up until he thinks it has been satisfied, or until he thinks it will never be true.*

Persistent goals enjoy the following properties:

Proposition 5.18

- $\models PGOAL\ i\ \neg\varphi \rightarrow \neg PGOAL\ i\ \varphi$
- $\models PGOAL\ i\ \psi \rightarrow \diamond(BEL\ i\ \psi \vee BEL\ i\ \Box\neg\psi)$
- $\models [PGOAL\ j\ \varphi \wedge \Box COMPETENT\ j\ \varphi\ \wedge$ $\neg BEFORE(BEL\ j\ \Box\neg\varphi)\neg GOAL\ j\ (LATER\ \varphi)] \rightarrow \diamond\varphi$

Remark 5.19 *The first property says that also persistent goals are consistent. The second one states that if an agent adopts a persistent goal, eventually he must believe it to be true, or believe that it will never become true. The third is a more complicated property: it says that if someone has a persistent goal of bringing about φ, φ is within his area of competence, and, before dropping his goal, the agent will not believe φ will never occur, then eventually φ will become true.*

Finally Cohen and Levesque are able to define a notion of intention as a kind of persistent goal:

Definition 5.20 *(Intention(_to_do).)*

$$INTEND_1\ i\ \alpha = PGOAL\ i\ [(DONE\ i\ (BEL\ i\ (HAPPENS\alpha))?;\alpha]$$

Remark 5.21 *This definition states that an agent intends to do an action α iff he has the persistent goal to have done that action (successfully) after having believed that it actually was taking place (so that it was not done accidentally or unknowingly!).*

This notion of intention enjoys the following nice properties:

Proposition 5.22 *1. (screen of admissibility)*
$\models INTEND_1\ i\ \beta \wedge \Box(BEL\ i\ [DONE\ i\ \alpha \rightarrow \Box\neg DONE\ i\ \beta])$
$\rightarrow \neg INTEND_1\ i\ \alpha;\beta$

2. *('tracking' success)*
   $\models (DONE\ i\ [INTEND_1\ i\ a \wedge BEL\ i\ (HAPPENS\ i\ a)]?; b) \wedge$
   $BEL\ i\ (\neg DONE\ i\ a) \wedge \neg BEL\ i\ \Box(\neg DONE\ i\ a)$
   $\rightarrow INTEND_1\ i\ a$

Remark 5.23 *The first of these properties expresses that intention provide a "screen of admissibility" for adopting other intentions: if an agent has an intention to do β, and the agent (always) believes that doing α prevents the achievement of β, then the agent cannot have the intention to do $\alpha; \beta$, or even the intention to do α before doing β. The second states that agents "track" the success of their attempts to achieve intentions, or in other words, agents keep their intentions after failure. If an agent has the intention to do a and then does something, b, thinking it would bring about the doing of a, but he then comes to believe it did not, then, provided the agent does not think a can never be done, the agent still has the intention to perform a.*

The notion $INTEND_1$ represents an intention to do an action. Cohen and Levesque also define a notion of intention that is applicable to a state of affairs (formula):

Definition 5.24 *(Intention_to_be.)*

$$INTEND_2\ i\ \varphi = PGOAL\ i\ \exists a(DONE\ i\ [BEL\ i\ \exists b HAPPENS\ i\ b; \varphi?)$$
$$\wedge \neg GOAL\ i\ \neg HAPPENS\ i\ a; \varphi?]?; a; \varphi?)$$

Remark 5.25 *The definition of intention_to_be looks rather cumbersome. It expresses firstly that when an agent intends to bring about a certain state of affairs φ, it is committed to do a sequence of events (denoted by) a himself, after which φ holds. To avoid that this happens accidentally or unknowingly, we furthermore require that the agent believes he is about to do some action ('plan') b which will have φ as a result. Finally it is specified that prior to doing a an agent does not have as a goal a's not bringing about φ, i.e. what in fact does happen (a) is compatible with the agent's goals.*

Although the theory of Cohen and Levesque yields a very interesting account of the motivational attitudes, in particular intentions, of agents, we observe, especially from the last definitions and propositions, that one has to deal with action in a rather roundabout way, rendering the theory rather complicated. In our opinion this may primarily be due to the fact that the logic is based on a temporal rather than an action-based framework, although it is also clear that the notions that Cohen and Levesque try to capture (like the notion of $INTEND_2$, in which the agent may believe other things to happen than actually happen, while maintaining his intention), are inherently complex.

6 BDI Logic

Another formalisation of related ideas is provided by the BDI model proposed by Rao & Georgeff [31, 32] which has been very influential in the agent community, too. The model is based on (branching time) temporal logic (CTL*). Agent behaviour is modelled by tree-like structures, where each path through such a tree represents a possible 'life' of the agent. The basic logic containing temporal modalities such as "along every path in the future there is some point where" is augmented by means of 'BDI'-modalities, viz. a belief operator BEL, a desire operator GOAL and an intention operator INTEND. Thus in this model one is able to express how the beliefs, desires and intentions of an agent evolve over time (or rather over possible time lines). Formally, Rao & Georgeff's BDI-model is a formal (modal) logic with a Kripke-style semantics and a logical calculus. Rao & Georgeff were especially interested in the relationship between the BDI modalities. In their paper they discuss several such possible relations such as Belief-Goal compatibility and Goal-Intention compatibility. The former expresses that agents believe that their goals are obtainable in some future, while the latter states that the agents' intentions should be goals. Rao & Georgeff and other researchers have used their model as an inspiration for their work on the realisation of agents. The BDI model has thus given rise to BDI architectures where the elements of belief bases, goals bases and plan libraries are central (cf. [41]).

We will now go briefly into some of the formal details. (The language of) BDI logic is constructed as follows. Two types of formulas are distinguished: state formulas and path formulas. We assume some given first-order signature. Furthermore, we assume a set E of event types with typical element e. The operators $BEL, GOAL, INTEND$ have as obvious intended reading the belief, goal and intention of an agent, respectively, while $\mathsf{U}, \diamond, \mathsf{O}$ are the usual temporal operators, viz. until, eventually and next, respectively.

Definition 6.1 *(State and path formulas.)*

1. *The set of* state formulas *is the smallest closed under:*
   - *any first-order formula w.r.t. the given signature is a state formula*
   - *if φ_1 and φ_2 are state formulas then also $\neg\varphi_1, \varphi_1 \vee \varphi_2, \exists x \varphi_1(x)$ are state formulas*
   - *if e is an event type, then $succeeded(e), failed(e)$ are state formulas*
   - *if φ is a state formula, then $BEL(\varphi), GOAL(\varphi), INTEND(\varphi)$ are state formulas*
   - *if ψ is a* path formula*, then $optional(\psi)$ is a state formula*
2. *The set of* path formulas *is the smallest set closed under:*
   - *any state formula is a path formula*
   - *if ψ_1, ψ_2 are path formulas, then $\neg\psi_1, \psi_1 \vee \psi_2, \psi_1 \mathsf{U} \psi_2, \diamond\psi_1, \mathsf{O}\psi_1$ are path formulas*

Remark 6.2 *As the names suggest, a state formulas will be interpreted over a state, that is a (state of the) world at a particular point in time, while path formulas will be interpreted over / along a path of a time tree (representing the evolution of a world). In the sequel we will see how this will be done formally. Here we just give the informal readings of the operators.*

As the names suggest the operators $succeeded$ *and* $failed$ *are used to express that events have (just) succeeded and failed, respectively. As in the framework of Cohen & Levesque action-like entities should be given a place in the theory by means of additional operators. Here we see that Rao & Georgeff's approach also account for the distinction of trying an action / event and succeeding versus failing. With the latter one may think of several things: either the agent tried to do some action which failed due to circumstances in the environment. For example, for an action 'grip' to be successful there should be an object to be gripped; for a motor to be started there should be fuel, etc.; perhaps there is also some internal capacity missing needed for successful performance of an action: again for an action 'grip' to be successful the robot should have a gripper. All this is related to the infamous* qualification problem in AI, *[33].*

Next there are the modal operators for belief, goal and intend. (In the original version of BDI theory [31], desires are represented by goals, or rather a $GOAL$ *operator. In a later paper [32] the* $GOAL$ *operator was replaced by* DES *for desire. Although this is perhaps a better name for the operator, here we stick to the original to keep more in line with the approach of Cohen & Levesque.) The optional operator states that there is a future (represented by a path) where the argument of the operator holds. Finally, there are the familiar (linear-time) temporal operators, such as the 'until', 'eventually' and 'nexttime', which are to be interpreted along a linear time path.*

Furthermore, the following abbreviations are defined:

Definition 6.3

1. $\Box\psi = \neg\diamond\neg\psi$ *(always)*
2. $inevitable(\psi) = \neg optional(\neg\psi)$
3. $done(e) = succeeded(e) \vee failed(e)$
4. $succeeds(e) = inevitable \bigcirc (succeeded(e))$
5. $fails(e) = inevitable \bigcirc (failed(e))$
6. $does(e) = inevitable \bigcirc (done(e))$

Remark 6.4 *The 'always' operator is the familiar one from (linear-time) temporal logic. The 'inevitability' operator expresses that its argument holds along all possible futures (paths from the current time). The 'done' operator states that an event occurs (action is done) no matter whether it is succeeding or not. The final three operators state that an event succeeds, fails, or is done iff it is inevitable (i.e. in any possible future) it is the case that at the next instance the event has succeeded, failed, or has been* *done, respectively. (so, this means that an event, succeeding or failing, is supposed to take one unit of time!)*

Definition 6.5 *(Semantics.)*
The semantics is given w.r.t. models of the form $\mathcal{M} = \langle W, E, T, \prec, \mathcal{U}, B, G, I, \Phi\rangle$, where

- *W is a set of possible worlds*
- *E is a set of primitive event types*
- *T is a set of time points*
- *$\prec$ is a binary relation on time points, which is serial, transitive and back-wards linear*
- *$\mathcal{U}$ is the universe of discourse*
- *Φ is a mapping of first-order entities to $\mathcal{U}$, for any world and time point*
- *$B, G, I \subseteq W \times T \times W$ are accessibility relations for $BEL, GOAL, INTEND$, respectively*

Remark 6.6 *The semantics of BDI logic, Rao & Georgeff-style, is rather complicated. Of course, we have possible worlds again, but as we will see below, these are not just unstructured elements, but they are each time trees, describing possible flows of time. So, we also need time points and an ordering on them. As BDI logic is based on branching time, the ordering need not be linear in the sense that all time points are related in this ordering. However, it is stipulated that the time ordering is serial (every time point has a successor in the time ordering), the ordering is transitive and backwards-linear, which means that every time point has only one direct predecessor. The accessibility relations for the 'BDI'-modalities are standard apart from the fact that they are also time-related, that is to say that worlds are (belief/goal/intend-)accessible with respect to a time point. Another way of viewing this is that – for all three modalities – for every time point there is a distinct accessibility relation between worlds.*

Next we elaborate on the structure of the possible worlds.

Definition 6.7 *(Possible worlds.)*
Possible worlds in W are assumed to be time trees*: an element $w \in W$ has the form $w = \langle T_w, A_w, S_w, F_w\rangle$ where*

- *$T_w \subseteq T$ is the set of time points in world w*
- *A_w is the restriction of the relation $\prec$ to T_w*
- *$S_w : T_w \times T_w \rightarrow E$ maps adjacent time points to (successful) events*
- *$F_w : T_w \times T_w \rightarrow E$ maps adjacent time points to (failing) events*
- *the domains of the functions S_w and F_w are disjoint*

Remark 6.8 *As announced before, a possible world itself is a time tree, a temporal structure representing possible flows of time. The definition above is just a technical one stating that the time relation within a possible* *world derives naturally from the* a priori *given relation on time points. Furthermore it is indicated by means of the functions S_w and F_w how events are associated with adjacent time points.*

Now we come to the formal interpretation of formulas on the above models. Naturally we distinguish state formulas and path formulas, since the former should be interpreted on states whereas the latter are interpreted on paths. In the sequel we use the notion of a *fullpath*: a fullpath in a world w is an *infinite* sequence of time points such that, for all i, $(t_i, t_{i+1}) \in A_w$. We denote a fullpath in w by $(w_{t0}, w_{t1}, \ldots)$.

Definition 6.9 *(Interpretation of formulas.) The interpretation of formulas w.r.t. a model $\mathcal{M} = \langle W, E, T, \prec, \mathcal{U}, B, G, I, \Phi \rangle$ is now given by:*

1. *(state formulas)*
   - $\mathcal{M}, v, w_t \models q(y_1, \ldots, y_n) \leftrightarrow (v(y_1), \ldots, v(y_n)) \in \Phi(q, w, t)$
   - $\mathcal{M}, v, w_t \models \neg\varphi \leftrightarrow \mathcal{M}, v, w_t \not\models \varphi$
   - $\mathcal{M}, v, w_t \models \varphi_1 \vee \varphi_2 \leftrightarrow \mathcal{M}, v, w_t \models \varphi_1$ *or* $\mathcal{M}, v, w_t \models \varphi_2$
   - $\mathcal{M}, v, w_t \models \exists x\varphi \leftrightarrow \mathcal{M}, v\{d/x\}, w_t \models \varphi$ *for some* $d \in \mathcal{U}$
   - $\mathcal{M}, v, w_{t0} \models optional(\psi) \leftrightarrow$ *exists fullpath* $(w_{t0}, w_{t1}, \ldots)$ *such that* $\mathcal{M}, v, (w_{t0}, w_{t1}, \ldots) \models \psi$
   - $\mathcal{M}, v, w_t \models BEL(\varphi) \leftrightarrow$ *for all* $w' \in B(w,t) : \mathcal{M}, v, w'_t \models \varphi$
   - $\mathcal{M}, v, w_t \models GOAL(\varphi) \leftrightarrow$ *for all* $w' \in G(w,t) : \mathcal{M}, v, w'_t \models \varphi$
   - $\mathcal{M}, v, w_t \models INTEND(\varphi) \leftrightarrow$ *for all* $w' \in I(w,t) : \mathcal{M}, v, w'_t \models \varphi$
   - $\mathcal{M}, v, w_t \models succeeded(e) \leftrightarrow$ *exists* $t0$ *such that* $S_w(t0, t) = e$
   - $\mathcal{M}, v, w_t \models failed(e) \leftrightarrow$ *exists* $t0$ *such that* $F_w(t0, t) = e$

   where $v\{d/x\}$ denotes the function v modified such that $v(x) = d$, and $R(w,t) = \{w' \mid R(w,t,w')\}$ for $R = B, G, I$

2. *(path formulas)*
   - $\mathcal{M}, v, (w_{t0}, w_{t1}, \ldots) \models \varphi \leftrightarrow \mathcal{M}, v, w_{t0} \models \varphi$, *for φ state formula*
   - $\mathcal{M}, v, (w_{t0}, w_{t1}, \ldots) \models \mathsf{O}\varphi \leftrightarrow \mathcal{M}, v, (w_{t1}, w_{t2}, \ldots) \models \varphi$
   - $\mathcal{M}, v, (w_{t0}, w_{t1}, \ldots) \models \diamond\varphi \leftrightarrow \mathcal{M}, v, (w_{tk}, \ldots) \models \varphi$ *for some* $k \geq 0$
   - $\mathcal{M}, v, (w_{t0}, w_{t1}, \ldots) \models \psi_1 \mathsf{U} \psi_2 \leftrightarrow$
     either there exists $k \geq 0$ such that $\mathcal{M}, v, (w_{tk}, \ldots) \models \psi_2$ and for all $0 \leq j < k$: $\mathcal{M}, v, (w_{tj}, \ldots) \models \psi_1$, or
     for all $j \geq 0$: $\mathcal{M}, v, (w_{tj}, \ldots) \models \psi_1$

Remark 6.10 *Most of the above clauses should be clear, incuding those concerning the modal operators for belief, goal and intention. The clause for the 'optional' operator expresses exactly that optionally ψ is true if ψ is true in one of the possible futures represented by fullpaths starting at the present time point. The interpretation of the temporal operators is as usual.*

Rao & Georgeff now discuss a number of properties that may be desirable to have as axioms. In the following we use α to denote so-called *O-formulas*, which are formulas that contain no positive occurrences of the '*inevitable*' operator (or negative occurrences of '*optional*") outside the scope of the modal operators $BEL, GOAL$ and $INTEND$.

1. $GOAL(\alpha) \rightarrow BEL(\alpha)$
2. $INTEND(\alpha) \rightarrow GOAL(\alpha)$
3. $INTEND(does(e)) \rightarrow does(e)$
4. $INTEND(\varphi) \rightarrow BEL(INTEND(\varphi))$
5. $GOAL(\varphi) \rightarrow BEL(GOAL(\varphi))$
6. $INTEND(\varphi) \rightarrow GOAL(INTEND(\varphi))$
7. $done(e) \rightarrow BEL(done(e))$
8. $INTEND(\varphi) \rightarrow inevitable \diamond (\neg INTEND(\varphi))$

Remark 6.11 *In order to render these formulas validities further constraints should be put on the models, since in the general setting above these are not yet valid. For reasons of space we will not enter into the details here; for these the reader is referred to [31, 32, 40].*

Looking at the first formula above it is intriguing to observe that Rao & Georgeff seem to propose the converse of a validity in the logic of Cohen & Levesque. This may seem rather puzzling. However, although there is definitely something strange about this, it should be kept in mind that, first of all, the formula $GOAL(\alpha) \rightarrow BEL(\alpha)$ is only proposed as a desired validity for certain formulas (viz. O-formulas) and not for all formulas, and also that the framework here (based on branching time) is quite different from that in Cohen & Levesque (which is based on linear time). Moreover, the very formulas for which the validity is wanted (the O-formulas) are typical branching-time formulas: they allow positive occurrence of 'optional' outside the scope of the doxastic and motivational modalities, thus they typically may express properties that hold along a branch (and not along all branches)! Of course, the very fact that here we have a formula as a proposed validity that is the converse of one proposed by Cohen & Levesque raises the question whether the notions of belief and goal that are modelled in both approaches are the same. I believe they are not, but it is very hard to put the exact differences into words. I invite the reader to ponder about this further. In any case Rao & Georgeff try to make the formula above (which they call 'belief-goal compatibility') plausible by considering a typical O-formula α of the form $optional(\psi)$, and then note that if it is a goal that something is optional (true in some future) then it should also be believed that it is optional (true in some future). This, indeed, sounds plausible in the sense that a rational and realistic agent would adhere to it. But also objective (nonmodal) formulas are O-formulas, and whether this is also plausible for these formulas I'm not sure. Perhaps for objective formulas we could say that if something is a goal now *it must coincide with a belief* now *(thus resulting in the validity $GOAL(p) \leftrightarrow BEL(p)$ for* *objective formulas p), since nothing can be done about it anymore (there is no future left in which we could work on it). This would reconcile for this class of formulas the approaches of Rao & Georgeff and Cohen & Levesque. For objective formulas goals trivialize to beliefs.*

The second formula is a similar one to the first. This one is called goal-intention compatibilty, and is defended by Rao & Georgeff by stating that if an optionality is intended it should also be wished (a goal in their terms). So, Rao & Georgeff have a kind of selection filter in mind: intentions (or rather intended options) are filtered / selected goals (or rather goal (wished) options), and goal options are selected believed options. If one views it this way, it looks rather close to Cohen & Levesque's Intention is choice (chosen / selected wishes) with commitment, or loosely, wishes that are committed to. Here the commitment acts as a filter.

The third one says that the agent really does the primitive actions that s/he intends to do. This means that if one adopts this as an axiom the agent is not allowed to do something else (first). (In my opinion this is rather strict on the agent, since it may well be that postponing its intention for a while is also an option.) On the other hand, as Rao & Georgeff say, the agent may also do things that are not intended since the converse does not hold. And also nothing is said about the intention to do complex actions.

The fourth, fifth and seventh express that the agent is conscious of its intentions, goals and what primitive action he has done in the sense that he believes what he intends, has as a goal and what primitive action he has just done.

The sixth one says something like that intentions are really wished for: if something is an intention then it is a goal that it is an intention.

The eighth formula states that intentions will inevitably (in every possible future) be dropped eventually, so there is no infinite deferral of its intentions. This leaves open, whether the intention will be fulfilled eventually, or will be given up for other reasons. Below we will discuss several possibilities of giving up intentions according to different types of commitment an agent may have.

BDI-logical expressions can be used to characterize different types of agents. Rao & Georgeff mention the following possibilities:

1. (blindly committed agent) $INTEND(\mathit{inevitable} \diamond \varphi) \rightarrow \mathit{inevitable}(INTEND(\mathit{inevitable} \diamond \varphi)\mathsf{U}BEL(\varphi))$
2. (single-minded committed agent) $INTEND(\mathit{inevitable} \diamond \varphi) \rightarrow \mathit{inevitable}(INTEND(\mathit{inevitable} \diamond \varphi)\mathsf{U}(BEL(\varphi) \vee \neg BEL(\mathit{optional} \diamond \varphi)))$
3. (open minded committed agent) $INTEND(\mathit{inevitable} \diamond \varphi) \rightarrow \mathit{inevitable}(INTEND(\mathit{inevitable} \diamond \varphi)\mathsf{U}(BEL(\varphi) \vee \neg GOAL(\mathit{optional} \diamond \varphi)))$

Remark 6.12 *A blindly committed agent maintains his intentions to inevitably obtaining eventually something until he actually believes that that* *something has been fulfilled. A single-minded committed agent is somewhat more flexible: he maintains his intention until he believes he has achieved it* or he does not believe that it can be reached (it is still an option in some future) anymore. *Finally, the open minded committed agent is even more flexible: he can also drop his intention if it is not a goal (desire) anymore.*

Rao & Georgeff are then able to obtain results under which conditions the various types of committed agents will reach their intentions. For example, for a blindly committed agent it holds that under the assumption of the axioms we have discussed earlier that:

$$INTEND(inevitable(\diamond\varphi)) \rightarrow inevitable(\diamond BEL(\varphi))$$

expressing that if the agent intends to eventually obtain φ it will inevitably eventually believe that it has succeeded in achieving φ.

7 KARO Logic

In this section we turn to our own formalisation of BDI-like notions, viz. the KARO formalism, in which *action* rather than time, together with knowledge / belief, is the primary concept, on which other agent notions are built. The KARO framework has been developed in a number of papers (e.g. [25, 26, 18, 28]) as well as the thesis of Van Linder ([24]).

The KARO formalism is an amalgam of dynamic logic and epistemic / doxastic logic, augmented with several additional (modal) operators in order to deal with the motivational aspects of agents. So, besides operators for knowledge (**K**), belief (**B**) and action ($[\alpha]$, "after performance of α it holds that"), there are additional operators for ability (**A**) and desires (**D**).

Assume a set $\mathcal{A}$ of atomic actions and a set $\mathcal{P}$ of atomic propositions.

Definition 7.1 *(Language.) The language $\mathcal{L}_{KARO}$ of KARO-formulas is given by the BNF grammar:*

$$\begin{array}{lcl} \varphi & ::= & p(\in \mathcal{P}) \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \ldots \\ & & \mathbf{K}\varphi \mid \mathbf{B}\varphi \mid \mathbf{D}\varphi \mid [\alpha]\varphi \mid \mathbf{A}\alpha \end{array}$$

$$\begin{array}{lcl} \alpha & ::= & a(\in \mathcal{A}) \mid \alpha_1;\alpha_2 \mid \varphi? \mid \\ & & \texttt{if } \varphi \texttt{ then } \alpha_1 \texttt{ else } \alpha_2 \texttt{ fi} \mid \\ & & \texttt{while } \varphi \texttt{ do } \alpha \texttt{ od} \end{array}$$

Here the formulas generated by the second (α) part are referred to as actions (or rather action expressions).

Remark 7.2 *Thus formulas are built by means of the familiar propositional connectives and the modal operators for knowledge, belief, desire, action and ability. Actions are the familiar ones from imperative programming: atomic ones, tests and sequential composition, conditional and repetition.*

Definition 7.3 *(KARO models.)*

1. *The semantics of the knowledge, belief and desires operators is given by means of Kripke structures of the following form:* $\mathcal{M} = \langle W, \vartheta, R_K, R_B, R_D \rangle$*, where*
   - W *is a non-empty set of states (or worlds)*
   - ϑ *is a truth assignment function per state*
   - R_K, R_B, R_D *are accessibility relations for interpreting the modal operators* $\mathbf{K}, \mathbf{B}, \mathbf{D}$*. The relation* R_K *is assumed to be an equivalence relation, while the relation* R_B *is assumed to be euclidean, transitive and serial. Futhermore we assume that* $R_B \subseteq R_K$*. (No special constraints are assumed for the relations* R_D*.)*
2. *The semantics of actions is given by means of structures of type* $\langle \Sigma, \{R_a \mid a \in \mathcal{A}\}, \mathcal{C}, Ag \rangle$*, where*
   - Σ *is the set of possible model/state pairs (i.e. models of the above form, together with a state appearing in that model)*
   - R_a $(a \in \mathcal{A})$ *are relations on* Σ *encoding the behaviour of atomic actions*
   - $\mathcal{C}$ *is a function that gives the set of actions that the agent is able to do per model/state pair*
   - Ag *is a function that yields the set of actions that the agent is committed to (the agent's 'agenda') per model/state pair.*

Remark 7.4 *We observe familiar elements in the structures for the operators for knowledge, belief, and desire. Actions are modelled as model/state pair transformers to emphasize their influence on the mental state (that is, the complex of knowledge, belief and desires) of the agent rather than just the state of the world. Both (cap)abilities and commitments are given by functions that yield the relevant information per model / state pair.*

Definition 7.5 *(Interpretation of formulas.) In order to determine whether a formula* $\varphi \in \mathcal{L}$ *is true in a model/state pair* (M, w) *(if so, we write* $(M, w) \models \varphi$*), we stipulate:*

- $\mathcal{M}, w \models p$ *iff* $\vartheta(w)(p) = \mathrm{true}$*, for* $p \in \mathcal{P}$
- *The logical connectives are interpreted as usual.*
- $\mathcal{M}, w \models \mathbf{K}\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $R_K(w, w')$
- $\mathcal{M}, w \models \mathbf{B}\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $R_B(w, w')$
- $\mathcal{M}, w \models \mathbf{D}\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $R_D(w, w')$
- $\mathcal{M}, w \models [\alpha]\varphi$ *iff* $\mathcal{M}', w' \models \varphi$ *for all* $\mathcal{M}', w'$ *with* $R_\alpha((\mathcal{M}, w), (\mathcal{M}', w'))$

- $\mathcal{M}, w \models \mathbf{A}\alpha$ *iff* $\alpha \in \mathcal{C}(\mathcal{M}, w)$[3]
- $\mathcal{M}, w \models \mathbf{Com}(\alpha)$ *iff* $\alpha \in Ag(\mathcal{M}, w)$[4]

Here R_α is defined as usual in dynamic logic by induction from the basic case R_a (cf. e.g. [14, 24, 18], but now on model/state pairs rather than just states). Likewise the function $\mathcal{C}$ is lifted to sets of complex actions ([24, 18]).

Remark 7.6 *We observe the by now familiar clauses for knowledge, belief and desire. The action modality gets a similar interpretation: something (necessarily) holds after the performance / execution of action α if it holds in all the situations that are accessible from the current one by doing the action α. The only thing which is a bit nonstandard is that, as stated above, a situation is characterised here as a model / state pair. The interpretations of the ability and commitment operators are rather trivial in this setting (but see the footnotes): an action is enabled (or rather: the agent is able to do the action) if it is indicated so by the function C, and, likewise, an agent is committed to an action α if it is recorded so in the agent's agenda.*

Furthermore, we will make use of the following syntactic abbreviations serving as auxiliary operators:

Definition 7.7

- *(dual)* $\langle\alpha\rangle\varphi = \neg[\alpha]\neg\varphi$, *expressing that the agent has the opportunity to perform α resulting in a state where φ holds.*
- *(opportunity)* $\mathbf{O}\alpha = \langle\alpha\rangle\mathtt{tt}$, *i.e., an agent has the opportunity to do an action iff there is a successor state w.r.t. the R_α-relation;*
- *(practical possibility)* $\mathbf{P}(\alpha, \varphi) = \mathbf{A}\alpha \wedge \mathbf{O}\alpha \wedge \langle\alpha\rangle\varphi$, *i.e., an agent has the practical possibility to do an action with result φ iff it is both able and has the opportunity to do that action and the result of actually doing that action leads to a state where φ holds;*
- *(can)* $\mathbf{Can}(\alpha, \varphi) = \mathbf{KP}(\alpha, \varphi)$, *i.e., an agent can do an action with a certain result iff it knows it has the practical possibilty to do so;*
- *(realisability)* $\Diamond\varphi = \exists a_1, \ldots, a_n \mathbf{P}(a_1; \ldots; a_n, \varphi)$[5], *i.e., a state property φ is realisable iff there is a finite sequence of atomic actions of which the agent has the practical possibility to perform it with the result φ;*
- *(goal)* $\mathbf{G}\varphi = \neg\varphi \wedge \mathbf{D}\varphi \wedge \Diamond\varphi$, *i.e., a goal is a formula that is not (yet) satisfied, but desired and realisable.*[6]

[3] In [19] we have shown that the ability operator can alternatively defined by means of a second accessibility relation for actions, in a way analogous to the opportunity operator below.

[4] The agenda is assumed to be closed under certain conditions such as taking 'prefixes' of actions (representing initial computations). Details are omitted here, but can be found in [28].

[5] We abuse our language here slightly, since strictly speaking we do not have quantification in our object language. See [28] for a proper definition.

[6] In fact, here we simplify matters slightly. In [28] we also stipulate that a goal should be explicitly selected somehow from the desires it has, which is modelled in that paper by means of an additional modal operator. Here we leave this out for simplicity's sake.

- *(possible intend)* $\mathbf{I}(\alpha, \varphi) = \mathbf{Can}(\alpha, \varphi) \wedge \mathbf{KG}\varphi$, *i.e., an agent (possibly) intends an action with a certain result iff the agent can do the action with that result and it moreover knows that this result is one of its goals.*

Remark 7.8

- *The dual of the (box-type) action modality expresses that there is at least a resulting state where a formula φ holds. It is important to note that in the context of* deterministic *actions, i.e. actions that have at most one successor state, this means that the* only *state satisfies φ, and is thus in this particular case a stronger assertion than its dual formula $[\alpha]\varphi$, which merely states that if there are any successor states they will (all) statisfy φ. Note also that if atomic actions are assumed to be deterministic all actions including the complex ones will be deterministic.*
- *Opportunity to do an action is modelled by having at least one successor state according to the accessibility relation associated with the action.*
- *Practical possibility to to an action with a certain result is modelled as having both ability and opportunity to do the action with the appropriate result. Note that $\mathbf{O}\alpha$ in the formula $\mathbf{A}\alpha \wedge \mathbf{O}\alpha \wedge \langle\alpha\rangle\varphi$ is actually redundant since it already follows from $\langle\alpha\rangle\varphi$. However, to stress the opportunity aspect it is added.*
- *The Can predicate applied to an action and formula expresses that the agent is 'conscious' of its practical possibility to do the action resulting in a state where the formula holds.*
- *A formula φ is realisable if there is a 'plan' consisting of (a sequence of) atomic actions of which the agent has the practical possibility to do them with φ as a result.*
- *A formula φ is a goal in the KARO framework if it is not true yet, but desired and realisable in the above meaning, that is, there is a plan of which the agent has the practical possibility to realise it with φ as a result.*
- *An agent is said to (possibly) intend an action α with result φ if he Can do this (knows that he has the practical possibility to do so), and, moreover, knows that φ is a goal.*

In order to manipulate both knowledge / belief and motivational matters special actions `revise`, `commit` and `uncommit` are added to the language. (We assume that we cannot nest these operators. So, e.g., `commit` (`uncommit`α) is not a well-formed action expression. For a proper definition of the language the reader is referred to [28].) The semantics of these are again given as model/state transformers (We only do this here in a very abstract manner, viewing the accessibility relations associated with these actions as functions. For further details we refer to e.g. [24, 18, 28]):

Definition 7.9 *(Accessibility of revise, commit and uncommit actions.)*

1. *$R_{\mathtt{revise}\varphi}(\mathcal{M}, w) = update\_belief(\varphi, (\mathcal{M}, w))$.*
2. *$R_{\mathtt{commit}\alpha}(\mathcal{M}, w) = update\_agenda^{+}(\alpha, (\mathcal{M}, w))$, if $\mathcal{M}, w \models \mathbf{I}(\alpha, \varphi)$ for some φ, otherwise $R_{\mathtt{commit}\alpha}(\mathcal{M}, w) = \emptyset$ (indicating failure of the commit action).*
3. *$R_{\mathtt{uncommit}\alpha}(\mathcal{M}, w) = update\_agenda^{-}(\alpha, (\mathcal{M}, w))$, if $\mathcal{M}, w \models \mathbf{Com}(\alpha)$, otherwise $R_{\mathtt{uncommit}\alpha}(\mathcal{M}, w) = \emptyset$ (indicating failure of the uncommit action);*
4. *$\mathtt{uncommit}\alpha \in \mathcal{C}(\mathcal{M}, w)$ iff $\mathcal{M}, w \models \neg\mathbf{I}(\alpha, \varphi)$ for all formulas φ, that is, an agent is able to uncommit to an action if it is not intended to do it (any longer) for any purpose.*

Remark 7.10 *Here $update\_belief$, $update\_agenda^{+}$ and $update\_agenda^{-}$ are functions that update the agent's belief and agenda (by adding or removing an action), respectively. Details are omitted here, but essentially these actions are model/state transformers again, representing a change of the mental state of the agent (regarding beliefs and commitments, respectively). The $update\_belief(\varphi,, (\mathcal{M}, w))$ function changes the model $\mathcal{M}$ in such a way that the agent's belief is updated with the formula φ, while $update\_agenda^{+}(\alpha, (\mathcal{M}, w))$ changes the model $\mathcal{M}$ such that α is added to the agenda, and like wise for the $update\_agenda^{-}$ function, but now with respect to removing an action from the agenda. The formal definitions can be found in [25, 26] and [28]. The* `revise` *operator can be used to cater for revisions due to observations and communication with other agents, which we will not go into further here (see [26]).*

The interpretation of formulas containing revise and (un)commit actions is now done using the accessibility relations above. One can now define validity as usual with respect to the KARO-models. One then obtains the following validities (of course, in order to be able to verify these one should use the proper model and not the abstraction we have presented here.) Besides the familiar properties from epistemic / doxastic logic, typical properties of this framework, called the KARO logic, include (cf. [25, 28]):

Proposition 7.11

1. $\models \mathbf{O}(\alpha;\beta) \leftrightarrow \langle\alpha\rangle\mathbf{O}\beta$
2. $\models \mathbf{Can}(\alpha;\beta, \varphi) \leftrightarrow \mathbf{Can}(\alpha, \mathbf{P}(\beta, \varphi))$
3. $\models [\mathtt{revise}\varphi]\mathbf{B}\varphi$
4. $\models \mathbf{K}\neg\varphi \leftrightarrow [\mathtt{revise}\varphi]\mathbf{B}\mathtt{ff}$
5. $\models \mathbf{K}(\varphi \leftrightarrow \psi) \rightarrow ([\mathtt{revise}\varphi]\mathbf{B}\chi \leftrightarrow [\mathtt{revise}\psi]\mathbf{B}\chi)$
6. $\models \mathbf{I}(\alpha, \varphi) \rightarrow \langle\mathtt{commit}\alpha\rangle\mathbf{Com}(\alpha)$
7. $\models \mathbf{I}(\alpha, \varphi) \rightarrow \neg\mathbf{A}\mathtt{uncommit}(\alpha)$
8. $\models \mathbf{Com}(\alpha) \rightarrow \langle\mathtt{uncommit}(\alpha)\rangle\neg\mathbf{Com}(\alpha)$
9. $\models \mathbf{Com}(\alpha) \wedge \neg\mathbf{Can}(\alpha, \top) \rightarrow \mathbf{Can}(\mathtt{uncommit}(\alpha), \neg\mathbf{Com}(\alpha))$
10. $\models \mathbf{Com}(\alpha) \rightarrow \mathbf{K}\mathbf{Com}(\alpha)$

11. $\models \mathbf{Com}(\alpha_1;\alpha_2) \rightarrow \mathbf{Com}(\alpha_1) \wedge \mathbf{K}[\alpha_1]\mathbf{Com}(\alpha_2)$

12. $\models \mathbf{Com}(\texttt{if } \varphi \texttt{ then } \alpha_1 \texttt{ else } \alpha_2 \texttt{ fi}) \wedge \mathbf{K}\varphi \rightarrow \mathbf{Com}(\varphi?;\alpha_1)$

13. $\models \mathbf{Com}(\texttt{if } \varphi \texttt{ then } \alpha_1 \texttt{ else } \alpha_2 \texttt{ fi}) \wedge \mathbf{K}\neg\varphi \rightarrow \mathbf{Com}(\neg\varphi?;\alpha_2)$

14. $\models \mathbf{Com}(\texttt{while } \varphi \texttt{ do } \alpha \texttt{ od}) \wedge \mathbf{K}\varphi \rightarrow \mathbf{Com}((\varphi?;\alpha);\texttt{while } \varphi \texttt{ do } \alpha \texttt{ od})$

Remark 7.12 *The first of these properties says that having the opportunity to do a sequential composition of two actions amounts to having the opportunity of doing the first action first and then having the opportunity to do the second. The second states that an agent that* can *do a sequential composition of two actions with result φ iff the agent can do the first actions resulting in a state where it has the practical possibility to do the second with φ as result. The third expresses that a revision with φ results in a belief of φ. The fourth states that the revision with φ results in inconsistent belief iff the agent knows $\neg\varphi$ for certain. The fifth expresses that revisions with formulas that are known to be equivalent have identical results. The sixth asserts that if an agent possibly intends to do α with some result φ, it has the opportunity to commit to α with result that it is committed to α (i.e. α is put into its agenda). The seventh says that if an agent intends to do α with a certain purpose, then it is unable to uncommit to it (so, if it is committed to α it has to perservere in it). The eighth property says that if an agent is committed to an action and it has the opportunity to uncommit to it with as result that indeed the commitment is removed. The ninth says that whenever an agent is committed to an action that is no longer known to be practically possible, it knows that it can undo this impossible commitment. The tenth property states that commitments are known to the agent. The last four properties have to do with commitments to complex actions. For instance, the eleventh says that if an agent is committed to a sequential composition of two actions then it is committed to the first one, and it knows that after doing the first action it will be committed to the second action.*

8 Logics for Multi-Agent Systems

8.1 Multi-agent epistemic logic

In previous section we have concentrated mainly on single agents and how to describe them. Of course, if multiple agents are around, things become both more complicated as well as more interesting. To start with, with respect to the epistemic (doxastic) aspect, one can introduce epistemic (doxastic) operators for every agent, resulting in a multi-modal logic, called $\mathbf{S5}_n$. Models for this logic are inherently less simple and elegant as those for the single agent case (cf. [27]). So then one has indexed operators $\mathbf{K}_i$ and $\mathbf{B}_i$ for agent i's knowledge and belief, respectively.

But one can go on and define knowledge operators that involve a group of agents in some way. This gives rise to the notions of common and (distributed) group knowledge.

The simplest notion is that of 'everybody knows', often denoted by the operator $\mathbf{E_K}$. But one can also add an operator $\mathbf{C_K}$ for 'common knowledge', which is much more powerful. The language is the same as epistemic logic, only now extended with the clause:

Definition 8.1 *(multi-agent epistemic logic.)*

- *if φ is a multi-agent epistemic formula, then $\mathbf{E_K}\varphi$ and $\mathbf{C_K}\varphi$ are multi-agent epistemic formulas*

For the interpretation we use the following models:

Definition 8.2 *Models for n-agent epistemic logic are Kripke structures of the form $\mathcal{M} = \langle W, \vartheta, R_1, \ldots, R_n, R_E, R_C\rangle$, where:*

- *W is a non-empty set of states (or worlds)*
- *ϑ is a truth assignment function per state*
- *The R_i are accessibility relations on W for interpreting the modal operators $\mathbf{K}_i$, assumed to be equivalence relations*
- *$R_E = \bigcup_i R_i$*
- *$R_C = R_E^*$, the reflexive transitive closure of R_E*

Definition 8.3 *(Interpretation of multi-agent epistemic formulas.) In order to determine whether an multi-agent epistemic formula is true in a model/state pair $\mathcal{M}, w$ ($\mathcal{M}, w \models \varphi$), we stipulate:*

- *$\mathcal{M}, w \models p$ iff $\vartheta(w)(p) = \text{true}$, for $p \in \mathcal{P}$*
- *The logical connectives are interpreted as usual.*
- *$\mathcal{M}, w \models \mathbf{K}_i\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all w' with $R_i(w, w')$*
- *$\mathcal{M}, w \models \mathbf{E_K}\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all w' with $R_E(w, w')$*
- *$\mathcal{M}, w \models \mathbf{C_K}\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all w' with $R_C(w, w')$*

Using the analogous notion of validity as for single-agent epistemic logic, we obtain:

Proposition 8.4

- $\models \mathbf{E_K}\varphi \leftrightarrow \mathbf{K}_1\varphi \wedge \ldots \wedge \mathbf{K}_n\varphi$
- $\models \mathbf{C_K}(\varphi \rightarrow \psi) \rightarrow (\mathbf{C_K}\varphi \rightarrow \mathbf{C_K}\psi)$
- $\models \mathbf{C_K}\varphi \rightarrow \varphi$
- $\models \mathbf{C_K}\varphi \rightarrow \mathbf{C_K}\mathbf{C_K}\varphi$
- $\models \neg\mathbf{C_K}\varphi \rightarrow \mathbf{C_K}\neg\mathbf{C_K}\varphi$
- $\models \mathbf{C_K}\varphi \rightarrow \mathbf{E_K}\mathbf{C_K}\varphi$
- $\models \mathbf{C_K}(\varphi \rightarrow \mathbf{E_K}\varphi) \rightarrow (\varphi \rightarrow \mathbf{C_K}\varphi)$

Remark 8.5 *The first statement of this proposition shows that the 'everybody knows' modality is indeed what its name suggests. The next four says that common knowledge has at least the properties of knowledge: closed under implication, it is true, and enjoys the introspective properties. The sixth property says that common knowledge is known by everybody. The last is a kind of induction principle: the premise gives the condition under which one can 'upgrade the truth of φ to common knowledge of φ; this premise expresses that it is common knowledge that the truth of φ is known by everybody.*

As to multi-agent doxastic logic one can look at similar notions of 'everybody believes' and common belief. One can introduce operators $\mathbf{E_B}$ and $\mathbf{C_B}$ for these notions:

Definition 8.6 *(multi-agent doxastic logic.)*

- *if φ is a multi-agent doxastic formula, then $\mathbf{E_B}\varphi$ and $\mathbf{C_B}\varphi$ are multi-agent doxastic formulas*

For the interpretation we use the following models:

Definition 8.7 *Models for n-agent epistemic logic are Kripke structures of the form $\mathcal{M} = \langle W, \vartheta, R_1, \ldots, R_n, R_F, R_D \rangle$, where:*

- *W is a non-empty set of states (or worlds)*
- *ϑ is a truth assignment function per state*
- *The R_i are accessibility relations on W for interpreting the modal operators $\mathbf{B}_i$, assumed to be serial, transitive and euclidean relations*
- *$R_F = \bigcup_i R_i$*
- *$R_D = R_F^+$, the (nonreflexive) transitive closure of R_F*

Remark 8.8 *Note that the accessibility relation for common belief is the* non*reflexive closure of R_F, contrary to that for common knowledge. This has to do with the fact that common belief needs not to be true!*

Definition 8.9 *(Interpretation of multi-agent doxastic formulas.) In order to determine whether an multi-agent epistemic formula is true in a model/state pair $\mathcal{M}, w$ ($\mathcal{M}, w \models \varphi$), we stipulate:*

- *... (as usual)*
- *$\mathcal{M}, w \models \mathbf{B}_i\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all w' with $R_i(w, w')$*
- *$\mathcal{M}, w \models \mathbf{E_B}\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all w' with $R_F(w, w')$*
- *$\mathcal{M}, w \models \mathbf{C_B}\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all w' with $R_D(w, w')$*

Now we obtain a similar set of properties for common belief (cf. [21]):

Proposition 8.10

- $\models \mathbf{E_B}\varphi \leftrightarrow \mathbf{B}_1\varphi \wedge \ldots \wedge \mathbf{B}_n\varphi$
- $\models \mathbf{C_B}(\varphi \rightarrow \psi) \rightarrow (\mathbf{C_B}\varphi \rightarrow \mathbf{C_B}\psi)$
- $\models \mathbf{C_B}\varphi \rightarrow \mathbf{E_B}\varphi$

- $\models \mathbf{C_B}\varphi \rightarrow \mathbf{C_B}\mathbf{C_B}\varphi$
- $\models \neg\mathbf{C_B}\varphi \rightarrow \mathbf{C_B}\neg\mathbf{C_B}\varphi$
- $\models \mathbf{C_B}\varphi \rightarrow \mathbf{E_B}\mathbf{C_B}\varphi$
- $\models \mathbf{C_B}(\varphi \rightarrow \mathbf{E_B}\varphi) \rightarrow (\mathbf{E_B}\varphi \rightarrow \mathbf{C_B}\varphi)$

Remark 8.11 *Note the differences dus to the fact that common belief is not based on a reflexive accessibility relation.*

8.2 Multi-agent BDI logic

Also with respect to the other modalities one may consider multi-agent aspects. In this subsection we focus on the notion of collective or joint intentions. We follow ideas from [6] (but we give a slightly different though equivalent presentation of definitions). We now assume that we have belief and intention opertors $\mathbf{B}_i, \mathbf{I}_i$ for every agent $1 \leq i \leq n$.

First we enrich the language of multi-agent doxastic with operators $\mathbf{E_I}$ (everybody intends) and $\mathbf{M_I}$ (mutual intention). (We call this a multi-agent BDI logic, although multi-agent BI logic would be a more adequate name, since we leave out the modality of desire / goal.)

Definition 8.12 *(multi-agent BDI logic.) Multi-agent BDI logic is obtained by taking the (analogous clauses of) multi-agent doxastic logic of the previous subsection extended with the clauses:*

- *if φ is a multi-agent BDI formula, then so is $\mathbf{I}_i\varphi$ for every $1 \leq i \leq n$).*
- *if φ is a multi-agent BDI formula, then $\mathbf{E_I}\varphi$ and $\mathbf{M_I}\varphi$ are multi-agent BDI formulas*

The language thus obtained is interpreted on slightly enhanced models.

Definition 8.13 *Models for n-agent BDI logic are Kripke structures of the form $\mathcal{M} = \langle W, \vartheta, R_1, \ldots, R_n, R_F, R_D, S_1, \ldots, S_n, S_F, S_D \rangle$, where:*

- *W is a non-empty set of states (or worlds)*
- *ϑ is a truth assignment function per state*
- *The R_i are accessibility relations on W for interpreting the modal operators $\mathbf{B}_i$, assumed to be serial, transitive and euclidean relations, while the S_i are accessibility relations on W for interpreting the modal operators $\mathbf{I}_i$, assumed to be serial relations.*
- *$R_F = \bigcup_i R_i$ and $S_F = \bigcup_i S_i$*
- *$R_D = R_F^+$ and $S_D = S_F^+$, the (nonreflexive) transitive closure of R_F and S_F, respectively.*

Definition 8.14 *(Interpretation of multi-agent BDI formulas.) In order to determine whether an multi-agent epistemic formula is true in a model/state pair $\mathcal{M}, w$ ($\mathcal{M}, w \models \varphi$), we stipulate:*

- *... (as before)*

- $\mathcal{M}, w \models \mathbf{I}_i\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $S_i(w, w')$
- $\mathcal{M}, w \models \mathbf{E_I}\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $S_F(w, w')$
- $\mathcal{M}, w \models \mathbf{M_I}\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $S_D(w, w')$

Hence we get similar properties for mutual intention as we had for common belief (but of course no introspective properties):

Proposition 8.15

- $\models \mathbf{E_I}\varphi \leftrightarrow \mathbf{I}_1\varphi \wedge \ldots \wedge \mathbf{I}_n\varphi$
- $\models \mathbf{M_I}(\varphi \rightarrow \psi) \rightarrow (\mathbf{M_I}\varphi \rightarrow \mathbf{M_I}\psi)$
- $\models \mathbf{M_I}\varphi \rightarrow \mathbf{E_I}\varphi$
- $\models \mathbf{M_I}\varphi \rightarrow \mathbf{E_I}\mathbf{M_I}\varphi$
- $\models \mathbf{M_I}(\varphi \rightarrow \mathbf{E_I}\varphi) \rightarrow (\mathbf{E_I}\varphi \rightarrow \mathbf{M_I}\varphi)$

Remark 8.16 *We see that E-intentions ('everybody intends') and mutual intentions are defined in a way completely analogous with E-beliefs ('everybody believes') and common beliefs, respectively.*

Finally we define the notion of *collective intention* ($\mathbf{C_I}$) as follows:

Definition 8.17

- $\mathbf{C_I}\varphi \;=\; \mathbf{M_I}\varphi \wedge \mathbf{C_B}\mathbf{M_I}\varphi$

Remark 8.18 *This definition states that collective intentions are mutual intentions that are moreover* mutually believed *to be so.*

In this context we also mention the work of Singh [35] where multi-agent intentions are studied. An interesting distinction is made between *exodeictic* and *endodeictic* intentions of groups, where the former is 'pointing outward' (intention of the group as viewed by others) while the latter is 'pointing inward' (intention as viewed by the group itself). Technically Singh uses modal operators for intentions and commitments, and bases group intentions on the accessibility relations for the individual ones, where exodeictic and endodeictic intentions are treated in a different way, amounting to the following. A team exodeictically intends φ iff φ holds on all paths that satisfy the exodeictic intentions of the individual members of the team and satisfy the team structure requirements (as to commitments and coordination of interactions), while a team endodeictically intends φ iff φ holds on all paths that satisfy the endodeictic intentions of the individual members of the team, satisfy the team structure requirements, *and require that the members are committed to the team in bringing about that* φ.

Although a logic of common goals, intentions and commitments is important, such a logic generally does not say much about how these come about. These typically come about in a (social) process (e.g. by negotiation), and therefore notions like goal formation are procedural rather than declarative of nature. For instance, in Dignum and Conte [3] a sketch is given how goal formation comes about, using a logical framework which is BDI/KARO-like (based on an action logic) extended with operators for instrumentality (to talk about subgoals) and obligations (for the normative aspect).

8.3 Further developments: cooperation and normative behaviour

If one considers 'societies of agents' obviously also other notions become important besides mere multi-agent extensions of BDI-notions. For instance, one can investigate hoe communication takes place in such a system, and how this affects the mental states of the agents in the system. This in turn, is important for synchronisation, coordination and cooperation in the system. There has also been done some work on this. For example, Dignum and Van Linder [4] have extended the KARO framework to deal with speech acts. Moreover, in societies it may be important to consider norms, obligations and permissions as a way to control societal behaviour. By introducing *deontic* notions such as obligation, permission and prohibition, Dignum *et al.* [5] try to make the logical framework sufficientl expressive to treat this issue. In particular they consider how agents may take norms and obligations into account when deliberating its intentions (by means of a modified 'BDI loop').

9 Conclusion

In this paper we have reviewed a number of logical theories for describing intelligent agents. First we looked at single agents and considered logics for informational and motivational attitudes such as belief, desires and intentions. In particular we have discussed the theories of Cohen & Levesque, Rao & Georgeff and the KARO framework of Van Linder *et al.* Finally we turned to multi-agent logics and discussed multi-agent epistemic and doxastic logic, multi-agent BDI logic and some developments with respect to even more expressive logics in which one may express some more advanced societal issues such as norms, obligations and communication by speech acts.

Acknowledgements The author likes to thank Marco Wiering for reading a preliminary draft of this paper and suggestions for improvement. Also the author is grateful to Rineke Verbrugge for valuable input for this paper.

References and Further Reading

[1] M.E. Bratman, *Intentions, Plans, and Practical Reason*, Harvard University Press, Massachusetts, 1987.

[2] P.R. Cohen & H.J. Levesque, Intention is Choice with Commitment, *Artificial Intelligence* 42(3), 1990, pp. 213–261.

[3] F. Dignum & R. Conte, Intentional Agents and Goal Formation, in: *Intelligent Agents IV: Agents Theories, Architectures, and Languages (ATAL-97)* (M.P. Singh, A. Rao & M.J. Wooldridge, eds.), Springer, Berlin, 1998, pp. 231–244.

[4] F. Dignum & B. van Linder, Modelling Social Agents: Communication as Actions, in: *Intelligent Agents III: Agents Theories, Architectures, and Languages (ATAL-96)* (M. Wooldridge, J. Müller & N. Jennings, eds.), Springer, Berlin, 1997, pp. 205–218.

[5] F. Dignum, D. Morley, E.A. Sonenberg & L. Cavedon, Towards Socially Sophisticated BDI Agents, in: *Proc. 4th Int. Conf. on Multi-Agent Systems (ICMAS-2000)*, Boston, MA, 2000, pp. 111–118.

[6] B. Dunin-Kęplicz & R. Verbrugge, Collective Intentions, *Fundamenta Informaticae*, 2002, to appear.

[7] R.M. van Eijk, Programming Languages for Agent Communication, PhD. Thesis, Utrecht University, Utrecht, 2000.

[8] R.M. van Eijk, F.S. de Boer, W. van der Hoek & J.-J. Ch. Meyer, Operational Semantics for Agent Communication Languages, in: *Issues in Agent Communication: Proc. of First Workshop on Agent Communication Languages* (F. Dignum & B. Chaib-draa, eds.), Springer, Heidelberg, 2000, to appear.

[9] R. Fagin, J.Y. Halpern, Y. Moses & M.Y. Vardi, *Reasoning about Knowledge*, The MIT Press, Cambridge, Massachusetts, 1995.

[10] T. Finin, D. McKay & R. Fritzen, An Overview of KQML: A Knowledge Query and Manipulation Language, Techn. Report, U. of Maryland, CS Dept, 1992.

[11] M. Fischer, A Survey of Concurrent METATEM – The language and Its Applications, in: *Temporal Logic – Proc. of the First Int. Conf.* (D.M. Gabbay & H.J. Ohlbach, eds.), LNAI 827, Springer, Berlin, 1994, pp. 480–505.

[12] Foundation for Intelligent Physical Agents, FIPA'97 Specification, Part 2 - Agent Communication Language, 1997.

[13] G. de Giacomo, Y. Lespérance & H. Levesque, ConGolog, a Concurrent Programming Language Based on the Situation Calculus, *Artificial Intelligence* 121 (1,2), 2000, pp. 109–169.

[14] D. Harel, Dynamic Logic, in: D. Gabbay & F. Guenthner (eds.), *Handbook of Philosophical Logic, Vol. II*, Reidel, Dordrecht/Boston, 1984, pp. 497–604.

[15] K.V. Hindriks, Agent Programming Languages Programming with Mental Models), PhD. Thesis, Utrecht University, Utrecht, 2001.

[16] K.V. Hindriks, F.S. de Boer, W. van der Hoek & J.-J. Ch. Meyer, Agent Programming in 3APL, *Int. J. of Autonomous Agents and Multi-Agent Systems* 2(4), 1999, pp.357–401.

[17] W. van der Hoek, Systems for Knowledge and Belief, *Journal of Logic and Computation* 3(2), 1993, pp. 173–195.

[18] W. van der Hoek, B. van Linder & J.-J. Ch. Meyer, An Integrated Modal Approach to Rational Agents, in: M. Wooldridge & A. Rao (eds.), *Foundations of Rational Agency*, Applied Logic Series 14, Kluwer, Dordrecht, 1998, pp. 133–168.

[19] W. van der Hoek, J.-J. Ch. Meyer & J.W. van Schagen, Formalizing Potential of Agents: The KARO Framework Revisited, in: *Formalizing the Dynamics of Information* (M. Faller, S. Kaufmann & M. Pauly, eds.), CSLI Publications, (CSLI Lect. Notes 91), Stanford, 2000, pp. 51-67.

[20] N.R. Jennings & M.J. Wooldridge, *Agent technology: Foundations, Applications, and Markets*, Springer, Berlin, 1997.

[21] S. Kraus & D. Lehmann, Knowledge, Belief and Time, in: L. Kott (ed.), *Proceedings of the 13th Int. Colloquium on Automata, Languages and Programming*, Rennes, LNCS 226, Springer, Berlin, 1986.

[22] H.J. Levesque, P.R. Cohen & J.T. Nunes, On Acting Together, in: *Proc. Nat. Conf. on Artif. Intell.*, 1990, pp. 94-99.

[23] H.J. Levesque, R. Reiter, Y. Lespérance, F. Lin & R.B. Scherl, GOLOG: A Logic Programming Language for Dynamic Domains, *J. of Logic Programming* 31, 1997, pp. 59–84.

[24] B. van Linder, Modal Logics for Rational agents, PhD. Thesis, Utrecht University, 1996.

[25] B. van Linder, W. van der Hoek & J.-J. Ch. Meyer, Actions that Make You Change Your Mind: Belief Revision in an Agent-Oriented Setting, in: *Knowledge and Belief in Philosophy and Artificial Intelligence* (A. Laux & H. Wansing, eds.), Akademie Verlag, Berlin, 1995, pp. 103–146.

[26] B. van Linder, W. van der Hoek & J.-J. Ch. Meyer, Seeing is Believing (And So Are Hearing and Jumping), *Journal of Logic, Language and Information* 6, 1997, pp. 33–61.

[27] J.-J. Ch. Meyer & W. van der Hoek, *Epistemic Logic for AI and Computer Science*, Cambridge Tracts in Theoretical Computer Science 41, Cambridge University Press, 1995.

[28] J.-J. Ch. Meyer, W. van der Hoek & B. van Linder, A Logical Approach to the Dynamics of Commitments, *Artificial Intelligence* 113, 1999, 1–40.

[29] J.P. Müller, A Cooperation Model for Autonomous Agents, in: *Intelligent Agents III* (J.P. Müller, M. Wooldridge & N.R. Jennings, eds.), LNAI 1193, Springer, Berlin, 1997, pp. 245–260.

[30] A.S. Rao, AgentSpeak(L): BDI Agents Speak Out in a Logical Computable Language, in: *Agents Breaking Away* (W. van der Velde & J. Perram, eds.), LNAI 1038, Springer, Berlin, 1996, pp. 42–55.

[31] A.S. Rao & M.P. Georgeff, Modeling rational agents within a BDI-architecture, in *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)* (J. Allen, R. Fikes & E. Sandewall, eds.), Morgan Kaufmann, 1991, pp. 473–484.

[32] A.S. Rao & M.P. Georgeff, Decision Procedures for BDI Logics, *J. of Logic and Computation* 8(3), 1998, pp. 293–344.

[33] E. Sandewall & Y. Shoham, Nonmonotonic Temporal Reasoning, in: *Handbook of Logic in Artificial Intelligence and Logic Programming Vol.4 (Epistemic and Temporal Reasoning)* (D.M. Gabbay, C.J. Hogger & J.A. Robinson, eds.), Oxford University Press, Oxford, 1994.

[34] M.P. Singh, *Multi-Agent Systems: A Theoretical Framework for Intentions, Know-How, and Communications*, Springer, Heidelberg, 1994.

[35] M.P. Singh, The Intentions of Teams: Team Structure, Endodeixis, and Exodeixis, in: *Proc. 13th Eur. Conf. on Artif. Intell. (ECAI'98)* (H. Prade, ed.), Wiley, Chichester, 1998, pp. 303–307.

[36] M.P. Singh, A.S. Rao & M.P. Georgeff, Formal Methods in DAI: Logic-Based Representation and Reasoning, in: Multiagent Systems (G. Weiss, ed.), The MIT Press, Cambridge, MA, 1999, pp. 331–376.

[37] Y. Shoham, Agent-Oriented Programming, *Artificial Intelligence* 60(1), 1993, pp. 51–92.

[38] F. Voorbraak, The Logic of Objective Knowledge and Rational Belief, in: J. van Eijck (ed.), *Logics in AI (Proceedings of JELIA '90)*, LNCS 478, Springer, 1991, pp. 499–516.

[39] F. Voorbraak, *As Far as I Know: Epistemic Logic and Uncertainty*, PhD Thesis, Utrecht University, Utrecht, 1993.

[40] M.J. Wooldridge, *Reasoning about Rational Agents*, The MIT Press, Cambridge, MA, 2000.

[41] M.J. Wooldridge & N.R. Jennings (eds.), *Intelligent Agents*, Springer, Berlin, 1995.