Modal Logics for Intelligent Agents

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

August 15, 2004

1 Introduction

Intelligent agents have become a major field of research in AI. Although there is little consensus about the precise definition of an intelligent agent, it is generally held that agents are *autonomous* pieces of hardware/software, taking initiative on behalf of a user or to satisfy some goal, more in general. Agents are often held to possess *mental attitudes*, that is they are supposed to deal with information, and act upon this, based on motivation. This generally calls for a description in terms of the agent's beliefs/knowledge, desires, goals, intentions, commitments, obligations, etc.

To describe these mental or cognitive attitudes one may fruitfully employ *modal logic*. Typically for the description of agents one needs an amalgam of modal operators/logics to cater for several of the mental attitudes as mentioned above. Moreover, it is important to note that, since agents by definition act and display behaviour, it is important to include the *dynamics* of these mental attitudes in the description. One might even maintain that the logics of some of these attitudes, such as goals and *a fortiori* desire, *per se* have little interest: they are rather weak logics without exciting properties. What makes them interesting is their dynamics: their change over time in connection with each other! So, although (modal) logics for e.g. knowledge, belief, desires etc. certainly play a role, it is also imperative to be able to specify the agent's behaviour / attitudes over time. Therefore, generally also a (modal) logic of time or action plays a role in agent specification logics.

In this section we will first spend some time on modal logics for some of the mental attitudes in isolation, after which we will turn to agent logics proper that are proposed in the literature, which typically are mixtures of these 'single-attitude' logics and contain an element of time and/or action. Our emphasis in this part lies on the logical languages and semantics of these, and less on axiomatics and metatheory.

1.1 Epistemic and doxastic logic

Epistemic logic deals with the mental attitude of knowledge while doxastic logic treats belief. These logics have become quite popular in computer science as well as artificial intelligence to describe the knowledge/belief involved in (particularly distributed) computation processes and agents. As to the former, the work of Halpern et al. [24] must be mentioned. As this is beyond the scope of the present chapter we concentrate on the role of epistemic/doxastic logic in AI, and the description of intelligent agents in particular.

The main idea behind a modal approach to knowledge/belief is that if an agent is not sure about the truth of a certain proposition p (say that it rains outside), it considers both the possibility of a situation where p holds and that where p does not hold. Formally this is captured by a Kripke model where at the state/world the agent is in, the agent considers possible alternatives (captured by the accessibility relation) where some of these satisfy p while other ones do not satisfy p. So we have a very intuitive use of modal semantics here: a formula φ is known / believed by the agent if all alternatives deemed possible by the agent (formally, all worlds accessible for the agent from the actual state) satisfy φ. Thus we have the following formal definitions.

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 1.1 *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 1.2 *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 $\mathbf{K}$ or $\mathbf{B}$. 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.*

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 1.3 *(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* $R(w, w')$

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 1.4 *(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 1.5

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

The first validity 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. The second validity expresses that knowledge is true. (One cannot honestly, truthfully and justifiably state to *know* something that is false.) The third and fourth validities 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 ([40]). The logic can beaxiomized by taking the four above validities as axioms, together with an axiomatization of classical propositional logic and the rules of modus ponens and necessitation $(\varphi/\mathbf{K}\varphi)$.

With respect to doxastic logic we obtain the following validities:

Proposition 1.6

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

Again we observe the introspection properties, but the second validity now states that an agent's belief is not inconsistent, which is weaker than the property that belief should be true. If one takes these properties as axioms completed by modus ponens, necessitation for **B** and (sufficient) classical propositional validities, one obtains the system known as **KD45**.

A natural question is whether the knowledge and belief modalities are interrelated in some meaningful way. The answer is more involved than one might suspect. In the literature (for example [32, 28, 50, 51]), several interesting possibilities for such an interaction have been investigated. In these studies it became clear that one has to be careful in putting several plausible properties together, as one might otherwise end up with undesirable properties such as the collapse of knowledge and belief!

As usual with applications of modal logic, one may wonder whether the properties one obtains are all desirable and not 'over-idealisations'. In the realm of epistemic/doxastic logic one may dispute the so-called paradoxes of *logical omniscience*. Most of these are inherent in the use of (normal) modal logic using standard Kripke semantics. For example, the basic modal property $\models \mathbf{K}(\varphi \to \psi) \to (\mathbf{K}\varphi \to \mathbf{K}\psi)$ gives rise to a Sorites-like paradox: the agent knows *all*

consequences of its knowledge, and likewise for belief. For 'finitary', resource-bounded agents this is unrealistic. One can now either take this for granted and view the modal operators for knowledge/belief as idealisations of the real thing, or one has to resort to non-standard ('non-normal') semantics (such as neighborhood semantics) to be able to avoid validities such as the above one. For a fuller treatment of this issue we refer to [40, 38].

1.2 Deontic logic

1.2.1 Standard deontic logic

One of the first systems for deontic logic that really was a serious attempt to capture deontic reasoning was the now so-called "Old System" of Von Wright ([55]), of which a modal logic (Kripke-style) version has become known as Standard Deontic Logic (**SDL**).

The syntax of SDL is that of a propositional modal logic with a modal operator O for obligation. $O\varphi$ is read as 'φ is obligatory / obligated' or 'it ought to be the case that φ'. The modalities F and P for 'it is forbidden' and 'it is permitted', respectively, are introduced as abbreviations: $F\varphi = O\neg\varphi$ and $P\varphi = \neg F\varphi$: something is forbidden iff its negation is obligatory, and something is permitted iff it is not forbidden.

SDL has a Kripke-style modal semantics based on a set of possible worlds (, a truth assignment function of primitive propositions per possible world) and an accessibility relation associated with the O-modality. This accessibility relation points to "ideal" or "perfect deontic alternatives" of the world under consideration. The crux behind this is that in some possible world something (say φ) is obligated, if φ holds in all the perfect alternatives of this world, as indicated by the accessibility relation.

So, formally these models have the following form: $M = (S, \pi, R_O)$, where S is the set of states, π is a truth assignment function, and R_O is the deontic accessibility relation, which are assumed to be serial, i.e. for all $s \in S$ there is a $t \in S$ such that $R_O(s,t)$.

The operator O is interpreted by means of the relation R_O: $M, s \models O\varphi$ iff $M, t \models \varphi$ for all t with $R_O(s,t)$. Validity is defined as usual for modal logic. We obtain the following validities:

Proposition 1.7

- $O(\varphi \to \psi) \to (O\varphi \to O\psi)$
- $O(\varphi \wedge \psi) \leftrightarrow (O\varphi \wedge O\psi)$
- $P(\varphi \wedge \psi) \to (P\varphi \wedge P\psi)$
- $(F\varphi \vee F\psi) \to F(\varphi \wedge \psi)$
- $(O\varphi \vee O\psi) \to O(\varphi \vee \psi)$
- $P(\varphi \vee \psi) \leftrightarrow (P\varphi \vee P\psi)$
- $F(\varphi \vee \psi) \leftrightarrow (F\varphi \wedge F\psi)$

- $\neg(O\varphi \wedge O\neg\varphi)$

The first property ('K-axiom') together with modus ponens, necessitation ($\varphi/O\varphi$) and a sufficient number of axioms of propositional logic can be used to axiomatise **SDL**. (This system coincides with the system **KD** in the classification of Chellas ([12]).)

Again the question arises whether the above properties are adequate for deontic reasoning. **SDL** suffers from a number of paradoxes, again mostly inherent in the (normal) modal semantics of the operators. For example, Ross' paradox: $O\varphi \rightarrow O(\varphi \vee \psi)$: if one ought to mail the letter then one ought to mail it or burn it. This sounds peculiar, but if one interprets O as holding in ideal alternative worlds it is evidently true. What is problematic here, is that in natural language an obligation of a disjunction is normally held to be an obligation of one of the disjuncts that may be chosen arbitrarily by the agent, and this intuition is simply not captured by SDL semantics. There are also more serious paradoxes, notably those having to do with contrary-to-duty (CTD) imperatives, in which certain obligations are specified in case one is already violating another obligation. For example, one ought to refrain from killing animals. But if one kills an animal, one ought to do it gently. This kind of CTD obligations cannot be expressed adequately in **SDL**. To reason about CTDs, or more generally about *conditional* obligations of the form $O(\varphi/\psi)$, read as the obligation to φ under circumstance ψ, so-called *dyadic deontic logic* was introduced, aleady in the 60s by von Wright [56]. However, in the 90s it became apparent that a truly adequate treatment of CTDs seems to force one to enter the realm of *nonmonotonic / defeasible / preferential* reasoning, which is beyond the scope of this chapter. More about this can be found in [42] and particularly [43, 10].

1.2.2 Dynamic deontic logic

Another issue that plays a role in deontic logic is the confusion about the argument of the modal operators. In **SDL** these are propositions (and we may refer to the O-operator as being of an 'ought-to be' nature). But many examples in the literature (and indeed also the example we gave illustrating Ross' paradox) actually seem to concern actions rather than propositions. (This is already noted by e.g. Castañeda [11].) One may also try and capture this notion of 'ought-to-do' in a different logic, and this is what we do next.

DDL, introduced in [37], is a version of dynamic logic especially tuned to use as ought-to-do style deontic logic. It is based on the idea of Anderson's reduction of ought-to-be style deontic logic to alethic modal logic ([2]), but instead it reduces ought-to-do deontic logic to dynamic logic ([26]). The basic idea is very simple: some action is forbidden if doing the action leads to a state of violation. In a formula: $\hat{F}\alpha \leftrightarrow_{def} [\alpha]V$, where the dynamic logic formula $[\alpha]\varphi$ denotes that execution / performance of the action α leads (necessarily) to a state (or states) where φ holds, and V is a special atomic formula denoting violation. (We write $\hat{F}$ instead of F to indicate that this operator is of a different

(viz. 'to-do') kind than the SDL operator; likewise for the other operators in this section.)

Formally, we say that the meaning of action α is captured by an accessibility relation $R_\alpha \subseteq S \times S$ associated with α, where S is the set of possible worlds. This relation R_α describes exactly what possible moves (state transitions) are induced by performance of the action α: $R_\alpha(s,t)$ says that from s one can get into state t by performing α. (In concurrency semantics and process algebra this is often specified by a so-called (labeled) transition system which enables one to derive (all) transitions of the kind $s \rightarrow_\alpha t$, which in fact defines the relation R_α for all possible actions α.) Now the formal meaning of the formula $[\alpha]\varphi$ is given by: $[\alpha]\varphi$ is true in a state (possible world) s iff all states t with $R_\alpha(s,t)$ satisfy φ. This then provides the formal definition of the $\hat{F}$-operator, as given above. In the sequel we will also employ the dual $\langle\alpha\rangle$ of $[\alpha]$: $\langle\alpha\rangle\varphi$ is true in s iff there is some state t satisfying φ such that $R_\alpha(s,t)$.

The other deontic modalities are derivatives of $\hat{F}$: permission is not-forbidden ($\hat{P}\alpha \leftrightarrow \neg\hat{F}\alpha$), and obligation is forbidden-not-to ($\hat{O}\alpha \leftrightarrow \hat{F}\bar{\alpha}$), where $\bar{\alpha}$ has the meaning of "not-α". The formal semantics of this negated action is non-trivial, especially in case one considers composite actions (cf. [37, 53, 18, 52]). In these papers we considered connectives for composing non-atomic actions, such as '$\cup$' (choice, the dynamic analogue of disjunction in a static setting), '&' (parallel, the analogue of conjunction), '$-$' (non-performance, the analogue of negation), and ';' (sequential composition, which has no analogue in a static setting). Without giving a formal semantics here (see the papers mentioned above for that), the meaning of these are as follows: $\alpha_1 \cup \alpha_2$ expresses a choice between α_1 and α_2 (this—roughly—corresponds to taking $R_{\alpha_1\cup\alpha_2}$ as the set-theoretic union of R_{α_1} and R_{α_2}), $\alpha_1 \& \alpha_2$ a parallel performance of α_1 and α_2 (this amounts to more or less taking $R_{\alpha_1\&\alpha_2}$ to be the intersection of R_{α_1} and R_{α_2}), $\bar{\alpha}$ (we will also write $-\alpha$) the non-performance of α, as stated above (it more or less amounts to taking $R_{\bar{\alpha}}$ to be some complement of R_α, but see also the discussion below), and $\alpha_1;\alpha_2$ the performance of α_1 followed by that of α_2. For a full account of the semantics of particularly negated actions we refer to [37, 18, 19, 21].

With this semantics the following formulas are valid:

Proposition 1.8

- $[\alpha](\varphi \rightarrow \psi) \rightarrow ([\alpha]\varphi \rightarrow [\alpha]\psi)$
- $[\alpha;\beta]\varphi \leftrightarrow [\alpha][\beta]\varphi$
- $[\alpha \cup \beta]\varphi \leftrightarrow ([\alpha]\varphi \wedge [\beta]\varphi)$
- $[\alpha]\varphi \rightarrow [\alpha\&\beta]\varphi$
- $[-(\alpha;\beta)]\varphi \leftrightarrow ([-\alpha]\varphi \wedge [\alpha][-\beta]\varphi)$
- $[-\alpha]\varphi \rightarrow [-(\alpha \cup \beta)]\varphi$
- $[-(\alpha\&\beta)]\varphi \leftrightarrow ([-\alpha]\varphi \wedge [-\beta]\varphi)$
- $\hat{F}\alpha \leftrightarrow [\alpha]V$

- $\hat{P}\alpha \leftrightarrow \neg\hat{F}\alpha (\leftrightarrow \langle\alpha\rangle\neg V)$
- $\hat{O}\alpha \leftrightarrow \hat{F}(-\alpha) (\leftrightarrow [-\alpha]V)$
- $\hat{O}(\alpha;\beta) \leftrightarrow (\hat{O}\alpha \wedge [\alpha]\hat{O}\beta)$
- $\hat{P}(\alpha;\beta) \leftrightarrow \langle\alpha\rangle\hat{P}\beta$
- $\hat{F}(\alpha;\beta) \leftrightarrow [\alpha]\hat{F}\beta$
- $\hat{O}(\alpha\&\beta) \leftrightarrow (\hat{O}\alpha \wedge \hat{O}\beta)$
- $\hat{P}(\alpha\&\beta) \rightarrow (\hat{P}\alpha \wedge \hat{P}\beta)$
- $(\hat{F}\alpha \vee \hat{F}\beta) \rightarrow \hat{F}(\alpha\&\beta)$
- $(\hat{O}\alpha \vee \hat{O}\beta) \rightarrow \hat{O}(\alpha \cup \beta)$
- $\hat{P}(\alpha \cup \beta) \leftrightarrow (\hat{P}\alpha \vee \hat{P}\beta)$
- $\hat{F}(\alpha \cup \beta) \leftrightarrow (\hat{F}\alpha \wedge \hat{F}\beta)$

Modal action logics that contain action negation (complement) operators have been studied by several authors, for instance [47, 25, 5]. From these studies, in particular [5], it has become clear that there are several ways to define action negation, particularly in the context of the operators ';' and intersection. The choices made in **DDL** above were motivated mainly by the desirability of the validities concerning the deontic operators above.

There have been proposed several dynamic deontic logics that could be viewed as some kind of refinement of the original logic as presented here. These concern -amongst other ones- issues of context (pertaining to the kind of complement / negation again) [19], the exact way action bring about violations [20], and of a more refined view of action than just input-output relations [36, 9].

1.3 BDI logic

BDI logic as proposed by Rao & Georgeff [44] came about after the groundbreaking work on Bratman [4] on the philosophy of intelligent (human) agents. In this work Bratman made a case for the notion of *intention* besides belief and desire, to describe the behaviour of rational agents. Intentions force the agent to commit to certain desires and to really 'go for them'. So focus of attention is an important aspect here, which also enables the agent to monitor how s/he is doing and take measures if things go wrong. Rao & Georgeff stress that in the case of resource-bounded agents it is imperative to focus on desires / goals and make choices. This was also observed by Cohen & Levesque [13], who tried to formalize the notion of intention in a linear-time temporal logic in terms of the notion of a (persistent) goal.

Here we follow Rao & Georgeff who use a branching-time temporal logic framework to give a formal-logical account of BDI theory. BDI logic has influenced many researchers (including Rao & Georgeff themselves) to think about

architectures of agent-based systems in order to realize these systems. Rao & Georgeff's BDI logic is more liberal than that of Cohen & Levesque in the sense that they *a priori* regard each of the three attitudes of belief, desire and intention as primitive: they introduce separate modal operators for belief, desire and intention, and then study possible relations between them.

(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 1.9 *(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*

State formulas are interpreted over a state, that is a (state of the) world at a particular point in time, while path formulas are interpreted over 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.

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. This is related to the well-known *qualification problem in AI*, [48].

Next there are the modal operators for belief, goal and intend. (In the original version of BDI theory [44], desires are represented by goals, or rather a GOAL operator. In a later paper [45] the $GOAL$ operator was replaced by DES for desire.) 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 1.10

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

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 1.11 *(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 backwards 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*

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 1.12 *(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
- $S_w : T_w \times T_w \to E$ *maps adjacent time points to (successful) events*
- $F_w : T_w \times T_w \to E$ *maps adjacent time points to (failing) events*
- *the domains of the functions* S_w *and* F_w *are disjoint*

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 $(w_{t0}, w_{t1}, \ldots)$, and define $fullpaths(w)$ as the set of all fullpaths occurring in world w (i.e. all fullpaths that start somewhere in the time tree w).

Definition 1.13 *(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 $j \geq 0 :\ \mathcal{M}, v, (w_{tj}, \ldots) \models \psi_1$

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. $INTEND(\varphi) \rightarrow inevitable \diamond (\neg INTEND(\varphi))$

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 only consider the first two. (More can be found in [44, 45, 54].) In order to define constraints on the models such that these two become valid, we introduce the relation $\lhd$ on worlds, as follows: $w'' \lhd w' \Leftrightarrow fullpaths(w'') \subseteq fullpaths(w')$. So $w'' \lhd w'$ means that there the world (time tree) w'' represents less choices than w'.

Now we define the *B-G condition* as the property that the following holds:

$$\forall w' \in B(w,t) \exists w'' \in G(w,t) : w'' \lhd w'$$

Informally, this condition says that for any belief accessible world there is a goal accessible world that contains less choices. It is now easy to show the following proposition.

Proposition 1.14 *Let $\mathcal{BG}$ be the class of models of the above form that satisfy the B-G condition. Then: $\mathcal{BG} \models GOAL(\alpha) \rightarrow BEL(\alpha)$ for O-formulas α.*

Similarly one can define the *G-I condition* as

$$\forall w' \in G(w,t) \exists w'' \in I(w,t) : w'' \lhd w'$$

and obtain:

Proposition 1.15 *Let $\mathcal{GI}$ be the class of models of the above form that satisfy the G-I condition. Then: $\mathcal{GI} \models INTEND(\alpha) \rightarrow GOAL(\alpha)$ for O-formulas α.*

Let us now consider the properties deemed desirable by Rao & Georgeff again. Actually the first one is rather controversial. (It is in fact the inverse implication that Cohen & Levesque had in their framework, although admittedly that framework is quite different from Rao & Georgeff's because of the different temporal model – linear time instead of branching time, so that it is not completely fair to compare formulas...!) Rao & Georgeff try to render the formula concerned (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 is debatable.

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. 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). 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) $INTEND(\mathit{inevitable} \diamond \varphi) \rightarrow \mathit{inevitable}(INTEND(\mathit{inevitable} \diamond \varphi) \mathsf{U} (BEL(\varphi) \vee \neg GOAL(\mathit{optional} \diamond \varphi)))$

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 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: $INTEND(\mathit{inevitable}(\diamond\varphi)) \rightarrow \mathit{inevitable}(\diamond BEL(\varphi))$ expressing that if the agent intends to eventually obtain φ it will inevitably eventually believe that it has succeeded in achieving φ.

In a series of papers [7, 6, 8] Van der Torre *et al.* have extended the BDI framework to what they call the BOID framework dealing with the Beliefs, Obligations, Intentions and Desires of agents. Although the language of BOID contains operators for belief (B), obligation (O), intention (I) and desire (D),

and thus looks like an amalgam of BDI and deontic logic, BOID logic is not really a modal logic in the proper sense. The operators are not interpreted by means of accessibility relations in Kripke structures. Instead, a default logic [46] is employed and a BOID agent is specified by a number of default rules involving the BOID notions/operators, together with a priority relation on these rules. The form of these rules is $X_1 \hookrightarrow X_2$, where X_1, X_2 typically are expressions of the form Bφ, Oφ, Iφ, or Dφ. The main concern is which (consistent) extensions are yielded representing how the beliefs, obligations, intentions and desires can be combined (consistently) taking the priority on rules into account.

Another extension with a similar philosophy in mind of incorporating social notions into the BDI framework was proposed by Dignum *et al.* [22, 16] . This framework is called B-DOING, and treats Beliefs, Desires, Intentions, Norms and Goals. This approach is more like a normal modal logic, although a number of extra elements is added. As to the deontic aspect, this framework is built on dyadic (conditional) obligations. Logically, the most important addition is the incorporation of two operators: $N^z(p|q)$ and $O^x_{ab}(p|q)$, with as intended meanings "it is a norm of the society / organisation z that p should be true when q is true" and "when q is true, individual a is obliged to b that p should be true, where z is the organisation/society that is responsible for enforcing the penalty", respectively. Formally, to give an interpretation to these operators a possible world semantics is employed that is rather involved. The upshot of this semantics is that $N^z(p|q)$ ($N^z(p|q)$) holds if $p \wedge q$ worlds are preferred to $\neg p \wedge q$ ones and the (maximally) preferred q worlds satisfy p, where the preference relation on worlds is induced by (associated with) the norms and obligations in organisation z, respectively.

1.4 KARO logic

In this section we turn to 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. [34, 35, 29, 41]) as well as the thesis of Van Linder ([33]).

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

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

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

Here the formulas generated by the second (α) part are referred to as actions (action expressions). 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 1.17 *(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.*

Knowledge, belief, and desire are modeled by accessibility relations on worlds, as usual. 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 1.18 *(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* $\vartheta(w)(p) = \text{true}$, *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* $\alpha \in \mathcal{C}(\mathcal{M}, w)$[1]
- $\mathcal{M}, w \models \mathbf{Com}(\alpha)$ *iff* $\alpha \in Ag(\mathcal{M}, w)$[2]

Here R_α is defined as usual in dynamic logic by induction from the basic case R_a (cf. e.g. [27, 33, 29], but now on model/state pairs rather than just states). Likewise the function $\mathcal{C}$ is lifted to sets of complex actions ([33, 29]).

Knowledge, belief and desire are interpreted as modal operators, as usual. 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 slightly nonstandard is that 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 1.19

- *(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;*

[1] In [30] 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.

[2] 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 [41].

- *(realisability)* $\Diamond\varphi = \exists a_1, \ldots, a_n \mathbf{P}(a_1; \ldots; a_n, \varphi)$[3]*, 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.*[4]

- *(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 1.20

- *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.*

[3] We abuse our language here slightly, since strictly speaking we do not have quantification in our object language. See [41] for a proper definition.

[4] In fact, here we simplify matters slightly. In [41] 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.

- *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 [41].) 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. [33, 29, 41]):

Definition 1.21 *(Accessibility of revise, commit and uncommit actions.)*

1. $R_{\texttt{revise}\varphi}(\mathcal{M}, w) = update\_belief(\varphi, (\mathcal{M}, w))$.
2. $R_{\texttt{commit}\alpha}(\mathcal{M}, w) = update\_agenda^{+}(\alpha, (\mathcal{M}, w))$, *if* $\mathcal{M}, w \models \mathbf{I}(\alpha, \varphi)$ *for some* φ, *otherwise* $R_{\texttt{commit}\alpha}(\mathcal{M}, w) = \emptyset$ *(indicating failure of the commit action).*
3. $R_{\texttt{uncommit}\alpha}(\mathcal{M}, w) = update\_agenda^{-}(\alpha, (\mathcal{M}, w))$, *if* $\mathcal{M}, w \models \mathbf{Com}(\alpha)$, *otherwise* $R_{\texttt{uncommit}\alpha}(\mathcal{M}, w) = \emptyset$ *(indicating failure of the uncommit action);*
4. $\texttt{uncommit}\alpha \in \mathcal{C}(\mathcal{M}, w)$ *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.*

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 [34, 35] and [41]. 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 [35]).

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. [34, 41]):

Proposition 1.22

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. $\models [\texttt{revise}\varphi]\mathbf{B}\varphi$
4. $\models \mathbf{K}\neg\varphi \leftrightarrow [\texttt{revise}\varphi]\mathbf{B}\texttt{ff}$
5. $\models \mathbf{K}(\varphi \leftrightarrow \psi) \rightarrow ([\texttt{revise}\varphi]\mathbf{B}\chi \leftrightarrow [\texttt{revise}\psi]\mathbf{B}\chi)$
6. $\models \mathbf{I}(\alpha,\varphi) \rightarrow \langle\texttt{commit}\alpha\rangle\mathbf{Com}(\alpha)$
7. $\models \mathbf{I}(\alpha,\varphi) \rightarrow \neg\mathbf{A}\texttt{uncommit}(\alpha)$
8. $\models \mathbf{Com}(\alpha) \rightarrow \langle\texttt{uncommit}(\alpha)\rangle\neg\mathbf{Com}(\alpha)$
9. $\models \mathbf{Com}(\alpha) \wedge \neg\mathbf{Can}(\alpha,\top) \rightarrow \mathbf{Can}(\texttt{uncommit}(\alpha),\neg\mathbf{Com}(\alpha))$
10. $\models \mathbf{Com}(\alpha) \rightarrow \mathbf{KCom}(\alpha)$
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. $\models \mathbf{Com}(\texttt{while } \varphi \texttt{ do } \alpha \texttt{ od}) \wedge \mathbf{K}\varphi \rightarrow \mathbf{Com}((\varphi?;\alpha);\texttt{while } \varphi \texttt{ do } \alpha \texttt{ od})$

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.

The KARO framework has been extended in various ways. In [30] we have given an account of abilities based on dynamic logic (like we did already for results and opportunities). In [31] we considered *automated reasoning* (viz. resolution) methods for (a fragment of) KARO. Furthermore, Dignum and Van Linder [17] have extended it to deal with *speech acts*. Aldewereld *et al.* [1] have extended KARO with multi-agent notions such as *joint* beliefs, actions, goals and commitments. (We will return to this briefly in the next section.) Finally we mention that KARO can also be employed beyond the realm of rational agents: in [39] it is indicated how the framework may be used to describe *emotional* aspects of agency.

1.5 Multi-agent logics

In the previous sections 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. In this subsection we will look at two generalisations of single-agent logics to multi-agent logics, viz. multi-agent epistemic logic and multi-agent BDI logic.

1.5.1 Multi-agent epistemic logic

In a multi-agent setting one can extend a single-agent framework in several ways. 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. [40]). 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 $\mathbf{E}_\mathbf{K}$. But one can also add an operator $\mathbf{C}_\mathbf{K}$ for 'common knowledge', which is much more powerful. The language is the same as epistemic logic, only now extended with the clause:

Definition 1.23 *(multi-agent epistemic logic.)*

- *if φ is a multi-agent epistemic formula, then $\mathbf{E}_\mathbf{K}\varphi$ and $\mathbf{C}_\mathbf{K}\varphi$ are multi-agent epistemic formulas*

For the interpretation we use the following models:

Definition 1.24 *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 1.25 *(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* $R_i(w, w')$
- $\mathcal{M}, w \models \mathbf{E}_\mathbf{K}\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $R_E(w, w')$
- $\mathcal{M}, w \models \mathbf{C}_\mathbf{K}\varphi$ *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 1.26

- $\models \mathbf{E}_\mathbf{K}\varphi \leftrightarrow \mathbf{K}_1\varphi \wedge \ldots \wedge \mathbf{K}_n\varphi$
- $\models \mathbf{C}_\mathbf{K}(\varphi \to \psi) \to (\mathbf{C}_\mathbf{K}\varphi \to \mathbf{C}_\mathbf{K}\psi)$
- $\models \mathbf{C}_\mathbf{K}\varphi \to \varphi$
- $\models \mathbf{C}_\mathbf{K}\varphi \to \mathbf{C}_\mathbf{K}\mathbf{C}_\mathbf{K}\varphi$
- $\models \neg\mathbf{C}_\mathbf{K}\varphi \to \mathbf{C}_\mathbf{K}\neg\mathbf{C}_\mathbf{K}\varphi$
- $\models \mathbf{C}_\mathbf{K}\varphi \to \mathbf{E}_\mathbf{K}\mathbf{C}_\mathbf{K}\varphi$
- $\models \mathbf{C}_\mathbf{K}(\varphi \to \mathbf{E}_\mathbf{K}\varphi) \to (\varphi \to \mathbf{C}_\mathbf{K}\varphi)$

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 $\mathbf{E}_\mathbf{B}$ and $\mathbf{C}_\mathbf{B}$ for these notions:

Definition 1.27 *(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 1.28 *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*

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 1.29 *(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. [32]):

Proposition 1.30

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

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

1.5.2 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 [23] (but we give a slightly different but 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 1.31 *(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 1.32 *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 1.33 *(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:*

- *$\mathcal{M}, w \models \mathbf{I}_i\varphi$ 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 1.34

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

We see that E-intentions ('everybody intends') and multual intentions are defined in a way completely analogous with E-beliefs ('everybody believes') and common beliefs, respectively. Next we define the notion of *collective intention* ($\mathbf{C_I}$) as follows:

Definition 1.35

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

This definition states that collective intentions are those formulas that are mutually believed and of which this mutual belef is a common belief amongst all agents in the system.

Finally, we mention here that in the literature there is also other work on BDI-like logics for multi-agent systems where we encounter such notions as joint intentions, joint goals and joint commitments, mostly coined in the setting of how to specify teamwork. Seminal work was done by Cohen & Levesque [14]. This work was a major influence on our own multi-agent version of KARO [1]. An important complication in a notion of joint goal involves that of persistence of the goal: where in the single agent case the agent pursues its goal until it believes it has achieved it or believes it can never be achieved, in the context of multiple agents, the agent that realizes this, has to inform the others of the team about it so that the group / team as a whole will believe that this is the case and may drop the goal. Also the work of Singh [49] must be mentioned here, where 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).

References

[1] H.M. Aldewereld, W. van der Hoek & J.-J.Ch. Meyer, Rational Teams: Logical Aspects of Multi-Agent Systems. In: B. Dunin-Keplicz & R. Verbrugge (eds.), *Proc. FAMAS'03*, 2003, pp. 35–52, Warsaw, Poland; revised version to appear in *Fundamenta Informaticae*.

[2] A.R. Anderson, A Reduction of Deontic Logic to Alethic Modal Logic, *Mind* 67, 1958, pp. 100–103.

[3] L. Åqvist, Deontic Logic, in: *Handbook of Philosophical Logic, Vol. II* (D.M. Gabbay & F. Guenthner, eds.), Reidel, Dordrecht, 1984, pp. 605–714.

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

[5] J. Broersen, Action Negation and Alternative Reductions for Dynamic Deontic Logics, *J. of Applied Logic* 2(1), 2004, pp. 153–168.

[6] J. Broersen, M. Dastani, J. Hulstijn & L. van der Torre, The BOID Architecture: Conflicts between Beliefs, Obligations, Intentions and Desires, in: *Proc. 5th Int. Conf. on Autonomous Agents (AA2001)*, ACM Press, 2001, pp. 9–16.

[7] J. Broersen, M. Dastani, & L. van der Torre, Resolving Conflicts between Beliefs, Obligations, Intentions and Desires, in: *Proc. ECSQARU'01* (S. Benferhat & Ph. Besnard, eds.), LNAI 2143, 2001, Springer, Berlin, pp. 568–579.

[8] J. Broersen, M. Dastani, Z. Huang, J. Hulstijn & L. van der Torre, Goal Generation in the BOID Architecture, *Cognitive Science Quarterly* 2(3, 4), 2002, pp. 428–447.

[9] J. Broersen, R.J. Wieringa & J.-J. Ch. Meyer, A Fixed-Point Characterization of a Deontic Logic of REgular Action, *Fundamenta Informaticae* 48(2,3), 2001, pp. 107-128.

[10] J. Carmo & A.J.I. Jones, Deontic Logic and Contrary-to-Duties, in: *Handbook of Philosophical Logic (2nd edition)* (D. Gabbay & F. Guenthner, eds.) Vol. 8:, Kluwer, Dordrecht, 2003.

[11] H.-N. Castañeda, The Paradoxes of Deontic Logic: The Simplest Solution to All of Them in One Fell Swoop, in: *New Studies in Deontic Logic* (R. Hilpinen, ed.), Reidel, Dordrecht, 1981, pp. 37–85.

[12] B.F. Chellas, *Modal Logic: An Introduction*, Cambridge University Press, Cambridge / London, 1980.

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

[14] P. Cohen & H. Levesque, Teamwork, *Nous* 24(4), 1991, pp. 487–512.

[15] M. Dastani & L. van der Torre, Programming BOID Agents: a deliberation language for conflicts between mental attitudes and plans, in: *Proc. AAMAS'04*, 2004.

[16] F. Dignum, D. Kinny & E. Sonenberg, From Desires, Obligations and Norms to Goals, *Cognitive Science Quarterly* 2(3-4), 2002, pp. 407-430.

[17] 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.

[18] F.P.M. Dignum & J.-J.Ch. Meyer, Negations of Transactions and Their Use in the Specification of Dynamic and Deontic Integrity Constraints, in: *Semantics for Concurrency, Leicester 1990*, (M.Z. Kwiatkowska, M.W. Shields & R.M. Thomas, eds.), Springer-Verlag, London/Berlin, 1990, pp. 61-80.

[19] F. Dignum, J.-J. Ch. Meyer & R.J. Wieringa, Contextual Permission: A Solution to the Free Choice Paradox, in Proc. 2nd Int. Workshop on Deontic Logic in Computer Science (DEON'94), (A.J. Jones & M. Sergot, eds.), Tano A.S., Oslo, 1994, pp. 107-135.

[20] F. Dignum, J.-J. Ch. Meyer & R.J. Wieringa, A Dynamic Logic for Reasoning about Sub-Ideal States, in *Proc. ECAI'94 Workshop "Artificial Normative Reasoning"* (J. Breuker, ed.), Amsterdam, 1994, pp. 79-92.

[21] F.P.M. Dignum, J.-J. Ch. Meyer & R.J. Wieringa, Free Choice and Contextually Permitted Actions, *Studia Logica* 57(1), pp. 193-220, 1996.

[22] 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.

[23] B. Dunin-Kęplicz & R. Verbrugge, Collective Intentions, *Fundamenta Informaticae* 51(3) (2002), pp. 271–295.

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

[25] G. Gargov & S. Passy, A Note on Boolean Modal Logic, in: *Mathematical Logic, Proc. of Heyting'88* (P. Petkov, ed.), Plenum Press, 1990, pp. 311-321.

[26] D. Harel, *First-Order Dynamic Logic*, Springer-verlag, Berlin, 1979.

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

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

[29] 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.

[30] 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.

[31] U. Hustadt, C. Dixon, R.A. Schmidt, M. Fisher, J.-J. Ch. Meyer & W. van der Hoek, Verification within the KARO Agent Theory, in: (Proc. First Goddard Workshop on) Formal Approaches to Agent-Based Systems (FAABS 2000) (Rash, J.L. and Rouff, C.A.and Truszkowski, W.and Gordon, D. and Hinchey, M.G. eds.),, LNAI 1871, Springer, Berlin/Heidelberg, 2001, pp. 33–47.

[32] 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.

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

[34] 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.

[35] 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.

[36] R. van der Meyden, The Dynamic Logic of Permission, *J. of Logic and Computation* 6(3), 1996, pp. 465-479.

[37] J.-J.Ch. Meyer, A Different Approach to Deontic Logic: Deontic Logic Viewed as a Variant of Dynamic Logic, *Notre Dame J. of Formal Logic* 29(1), (1988), pp. 109–136.

[38] J.-J. Ch. Meyer, Modal Epistemic and Doxastic Logic, in: *Handbook of Philosophical Logic (2nd edition)* (D. Gabbay & F. Guenthner, eds.) Vol. 10, Kluwer, Dordrecht, 2003, pp. 1-38.

[39] J.-J. Ch. Meyer, Reasoning about Emotional Agents, in: *Proc. ECAI 2004*, IOS Press, 2004.

[40] 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.

[41] 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.

[42] J.-J. Ch. Meyer, R.J. Wieringa & F.P.M. Dignum, The Role of Deontic Logic in the Specification of Information Systems, in: Logics for Databases and Information Systems (J. Chomicki & G. Saake, eds.), Kluwer, Boston/Dordrecht, 1998, pp. 71-115.

[43] D. Nute, *Defeasible Deontic Logic*, Kluwer, Dordrecht, 1997.

[44] 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.

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

[46] R. Reiter, A Logic for Default Reasoning, *Artificial Intelligence* 13, 1980, pp. 81-132.

[47] M. de Rijke, A System of Dynamic Modal Logic, *J. of Philosophical Logic* 27, 1998, pp. 109–142.

[48] 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.

[49] 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.

[50] 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.

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

[52] R.J. Wieringa & J.-J. Ch. Meyer, Actors, Actions, and Initiative in Normative System Specification, *Annals of Mathematics and Artificial Intelligence* 7, 1993, pp. 289–346.

[53] R.J. Wieringa, J.-J.Ch. Meyer & H. Weigand, Specifying Dynamic and Deontic Constraints, *Data & Knowledge Engineering* 4(2), 1989, pp. 157-190.

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

[55] G.H. von Wright, Deontic Logic, *Mind* 60, 1951, pp. 1-15.

[56] G.H. von Wright, A New System of Deontic Logic, *Danish Yearbook of Philosophy* 1, 1964.