F. Dignum and B. van Linder. Modelling rational agents in a dynamic environment: Putting humpty dumpty together again. In J.L. Fiadeiro and P.-Y. Schobbens, editors, *ModelAge-96*, pages 81--92, Sesimbra, Portugal, 1996.

Modelling Rational Agents in a Dynamic Environment: Putting Humpty Dumpty Together Again

F. Dignum
Fac. of Maths. & Comp. Sc., Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
E-mail: dignum@win.tue.nl, phone: +31-40-2473705, fax: +31-40-2463992

B. van Linder*
Dept. of Computer Science, Utrecht University
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
E-mail: bernd@cs.ruu.nl, phone: +31-30-2534097, fax: +31-30-2513791

Abstract

In this paper we propose a formal framework to model rational agents. We distinguish four levels where aspects of agency are situated, viz. the informational, action, motivational and social level. On these levels we consider concepts like knowledge and belief of agents, opportunities for, and results of actions that they may perform, preferences, goals, intentions and commitments, and speech acts. The language describing these concepts is a multi-modal one, and the models used to interpret this language are Kripke-style possible worlds models. Both the language and the models are defined in a rigorously formal way, while the actual semantics are only sketched. We conclude this paper with a brief comparison with related work on the formalisation of rational agents.

EXTENDED ABSTRACT

1 Introduction

The formalisation of rational agents is a topic of continuing interest in AI. Research on this subject has held the limelight ever since the pioneering work of Moore [12] in which knowledge and actions are considered. Over the years contributions have been made on both *informational* attitudes like knowledge and belief [11] and *motivational* attitudes like intentions and commitments [1, 4]. Recent developments in which various kinds of attitudes are combined include the work on agent-oriented programming [15], the Belief-Desire-Intention architecture [13] and the specification of multi-agent systems [18].

In our basic framework [6, 7] we modelled the informational attitudes of agents as well as various aspects of action by means of a theory about the *knowledge, belief* and *abilities* of agents, as well as the *opportunities* for, and the *results* of their actions. In this framework it can for instance be modelled that an agent knows that it is able to perform an action and that it knows that it is correct to perform that action to bring about some result.

*This author is partially supported by Esprit III BRWG Project No. 8319 'ModelAge'.

We subsequently dealt with the motivational attitudes of agents [4, 9]. Here we defined the concepts of *preferences*, *goals*, *intentions*, and *commitments* or *obligations*. By combining this formalisation with the basic framework it is for instance possible to model the fact that an agent prefers some situation to hold while it also knows that it is able to bring about that situation by performing a sequence of actions. Furthermore it can be modelled that after an agent commits itself to achieve a goal it is obliged to perform those actions that achieve its goal.

Finally, in [2, 3, 8, 16] we formalised communication between agents. In this theory we model both the communication itself as well as the consequences of communication. For instance, if some authorised agent gives orders to another agent to perform a certain action, the latter agent will be obliged to perform the action. Also if an authorised agent asserts a fact to another agent, the latter agent will believe this fact to be true.

In this paper we intend to bring the different fragments of the framework together in one all-embracing formal system. That is, we will define a model for the following concepts: *belief, knowledge, action, preference, goal, decision, intention, commitment, obligation* and *communication*. Following [4, 9] we base this model on dynamic logic [5], which is extended with epistemic, doxastic, temporal and deontic (motivational) operators. The semantics will be based on Kripke structures with a variety of relations imposed on the states. We characterise this integrated framework by pointing out some differences with the formalisations proposed by Cohen & Levesque [1] and Rao & Georgeff [13], respectively.

The rest of this paper is organised as follows. In Section 2 we single out the concepts that in our opinion constitute agency. As mentioned above, these concepts are situated at four different levels. In Section 3 we define the multi-modal language used to formalise the concepts described in Section 2, and the models used to interpret this language. We furthermore sketch the actual semantics, without going into too much detail. Section 4 contains a comparison with the formalisations of agency as proposed by Cohen & Levesque and Rao & Georgeff, respectively. In Section 5 we round off.

2 The constituents of agency

The concepts that we consider to be essential when formalising rational agents, can roughly be situated at four different levels: the informational level, the action level, the motivational level and the social level. The informational level comprises concepts such as knowledge and belief. At the action level we consider actions, results, abilities and opportunities. At the motivational level concepts like preferences, intentions, decisions and commitments are situated. Finally, at the social level we deal with concepts such as obligations and speech acts.

2.1 The informational level

At the informational level we consider both knowledge and belief. Many formalisations have been given of these concepts and we will follow the more common approach in epistemic and doxastic logic: the formula $K_i\phi$ denotes the fact that agent i knows ϕ and $B_i\phi$ that agent i believes ϕ. Both concepts are interpreted in a Kripke-style semantics, where each of the operators is interpreted by a relation between a possible world and a set of possible worlds determining the formulas that the agent knows respectively believes. We demand knowledge

to obey an S5 axiomatisation, belief to validate a KD45 axiomatisation, and agents to believe all the things that they know.

2.2 The action level

At the action level we consider both dynamic and temporal notions. The main dynamic notion that we consider is that of actions, which we interpret as functions that map some some state of affairs into another one. Following [6, 17] we use parameterised actions to describe the event consisting of a particular agent's execution of an action. We let $\alpha(i)$ indicate that agent i performs the action α. The results of actions are modelled using concepts from dynamic logic [5]: $[\alpha(i)]\phi$ indicates that if agent i performs the action indicated by α the result will be ϕ. Note that it does not state anything about whether the action will actually be performed. So, it might for instance be used to model a statement like: 'If I jump over 2.5m high I will be the world record holder'.

Besides these formulas that indicate the results of actions we also would like to express that an agent has the reliable opportunity to perform an action. This is done through the predicate OPP: $OPP(\alpha(i))$ indicates that agent i has the opportunity to do α, i.e. the event $\alpha(i)$ will possibly take place. Besides the OPP operator, which already has a temporal flavour to it, we introduce two genuinely temporal operators: $PREV$, denoting the events that actually just took place, and the standard temporal operator $NEXT$, which indicates, in our case, which event will actually take place next. We also define a $NEXT$ operator on formulas in terms of the $NEXT$ operator on events:

$$NEXT(\phi) \equiv \forall \alpha(i) : NEXT(\alpha(i)) \rightarrow [\alpha(i)]\phi$$

2.3 The motivational level

At the motivational level we consider a variety of concepts, ranging from preferences, goals and decisions to intentions and commitments. The most fundamental of these notions is that of preferences. Formally, preferences are defined as the combination of implicit and explicit preferences, which allows us to avoid all kinds of problems that plague other formalisations of motivational attitudes. A formula ϕ is preferred by an agent i, denoted by $P_i\phi$, iff ϕ is true in all the states that the agent considers desirable, and ϕ is an element of a predefined set of (explicitly preferred) formulas.

Goals are not primitive in our framework, but instead defined in terms of preferences. Informally, a preference of agent i constitutes one of i's goals iff i knows the preference not to be brought about yet, but implementable, i.e. i knows that it has the opportunity to achieve the goal. To formalise this notion, we first introduce the operator $Achiev$. Informally, $Achiev_i\phi$ means that agent i has the opportunity to perform some action which leads to ϕ. Formally, $Achiev$ is defined by

$$Achiev_i\phi \equiv \exists \beta : [\beta(i)]\phi \wedge OPP(\beta(i))$$

A goal is now formally defined as a (known) preference, which is known not to hold but to be achievable:

$$Goal_i\phi \equiv K_iP_i\phi \wedge K_i\neg\phi \wedge K_iAchiev_i\phi$$

Note that our definition implies that there are three ways for an agent to drop one of its goals: since it no longer considers achieving the goal to be desirable, since it no longer knows

the preference not to hold, or since it is no longer certain that it can achieve the goal. This implies in particular that our agents will not indefinitely pursue impossible goals.

Intentions are divided in two categories, viz. the intention to perform an action and the intention to bring about a proposition. We define the intention of an agent to perform a certain action as primitive. We relate intentions and goals in two ways. Firstly, the intention to bring about a proposition is defined as the goal to bring about that proposition. The second way is through *decisions*. An intention to perform an action is based on the decision to try to bring about a certain proposition. We assume a (total) ordering between the explicit preferences of each agent in each world. On the basis of this ordering the agent can make a decision to try to achieve the goal that has the highest preference. Because the order of the preferences may differ in each world, this does not mean that once a goal has been fixed the agent will always keep on trying to reach that goal (at least not straight away). As the result of deciding to do α, denoted by $DEC(i, \alpha)$, the agent has the intention to do α, denoted by $INT_i\alpha$. Formally, this is described by

$$OPP(DEC(i, \alpha)) \text{ iff } \exists \phi : Goal_i\phi \wedge [\alpha; \beta(i)]\phi \wedge \neg\exists\psi(P_i\psi \wedge \phi < \psi)$$

There is no direct relation between the intention to perform an action and the action that is actually performed next. We do, however, establish an indirect relation between the two through a binary implementation predicate, ranging over pairs of actions. The idea is that the formula $IMP_i(\alpha_1, \alpha_2)$ expresses that, for agent i, executing α_2 is a reasonable attempt at executing α_1. For example, if I intend to jump over 1.5m and I jump over 1.4m it can be said that I tried to implement my intention, i.e. the latter action is within the intention of performing the first action. However, if instead of jumping over 1.5m I killed a referee it can no longer be said that I performed that action with the intention of jumping over 1.5m.

Having defined the binary IMP predicate, we may now relate intended actions to the actions that are actually performed. We demand the action that is actually performed by an agent to be an attempt to perform one of its intentions. Formally, this amounts to the formula

$$(INT_i\alpha_1 \wedge NEXT(\alpha_2(i))) \rightarrow IMP_i(\alpha_1, \alpha_2)$$

being valid.

The last concept that we consider here, viz. that of commitment, links the motivational level to the social level. Agents can make commitments either to themselves (an act situated at the motivational level) or to other agents (situated at the social level). We treat the motivational kind of commitment as a special instance of the social kind. As the result of i performing a $COMMIT(i, j, \alpha)$ action the formula $O_{ij}\alpha$ becomes true (cf. [3]), i.e. by committing itself to an action, an agent i obliges itself towards j to perform the action α. The commitment is a private one if j is the same as i. Although the obligation does not ensure the actual performance of the action by the agent, it does have a practical consequence. If an agent commits itself to an action and afterwards does not perform the action a *violation* condition is registered, i.e. the state is not ideal (anymore). To model violations we introduce an operator O_{ij}, ranging over formulas, which is semantically interpreted as a (deontic) relation between states. This relation connects each world with the set of ideal worlds with respect to that world. More details about the formal semantics of this deontic operator can be found in [4].

2.4 The social level

The $COMMIT$ action described in the previous section is one of the four types of *speech acts* [14] that play a role at the social level. Speech acts deal with aspects of communication. The result of a speech act is a change in the doxastic or deontic state of an agent, or in some cases a change in the state of the world. We distinguish the following speech act types: *commitments, directions, declarations* and *assertions*. The idea underlying a direction is that of giving orders, i.e. an utterance like 'Forward, march!'. A typical example of a declaration is the utterance 'Let there be light', and a typical assertion is 'I tell you that the earth is flat'. For a speech act to be successful, the agent that utters it has to have some kind of authority: in the case of directions and assertions with the agent to which the speech act is directed, and in the case of a declaration it has to be in the power of the agent to make the declaration. For instance, generals have the authority to order soldiers, pupils believe that the earth is flat because their teacher told them so, and in the Netherlands only civil servants can declare people married. We formalise this authority relation through a binary predicate $auth$; $auth(i, j)$ means that agent i is considered an authority by agent j and $auth(i, f)$ means that it is within the authority of agent i to declare f. The speech acts are formalised as meta-actions: $DIR(i, j, \alpha)$ formalises that agent i directs agent j to perform α, $DECL(i, f)$ models the declaration of i that f holds, and $ASS(i, j, f)$ formalises the assertion of i to agent j that f holds.

3 A sketch of a formalisation

In this section we precisely define the language that we use to formally represent the concepts described in the previous section, and the models that are used to interpret this language. We will not go into too much detail with regard to the actual semantics, but try to provide the reader with an intuitive grasp for the formal details without actually mentioning them.

The language that we use is a multi-modal, propositional language, based on three denumerable, pairwise disjoint sets: Π, representing the propositional symbols, Ag representing agents, and At containing atomic action expressions. The language $FORM$ is defined in four stages. Starting with a set of propositional formulas ($PFORM$), we define the action- and meta-action expressions, after which $FORM$ can be defined.

Definition 1 *The language $PFORM$ of propositional formulas is defined to be the smallest set closed under:*

1. $\Pi \subseteq PFORM$
2. $f, f_1, f_2 \in PFORM \Longrightarrow \neg f, f_1 \wedge f_2 \in PFORM$

The set Act of regular action expressions is built up from the set At of atomic action expressions using the operators ; (sequential composition), + (nondeterministic composition), & (parallel composition), and $^-$(action negation). The constant actions **any** and **fail** denote 'don't care what happens' and 'failure' respectively.

Definition 2 *The set Act of action expressions is defined to be the smallest set closed under:*

1. $At \cup \{\mathbf{any}, \mathbf{fail}\} \subseteq Act$

2. $\alpha_1, \alpha_2 \in Act \Longrightarrow \alpha_1;\alpha_2, \alpha_1 + \alpha_2, \alpha_1 \& \alpha_2, \overline{\alpha_1} \in Act$

The set $MAct$ of general action expressions contains the regular actions and all of the special meta-actions informally described in the previous section. For simplicity we restrict ourselves to closing the set $MAct$ under sequential composition.

Definition 3 *The set $MAct$ of general action expressions is defined to be the smallest set closed under:*

1. $Act \subseteq MAct$
2. $\alpha \in Act, i, j \in Ag \Longrightarrow DEC(i,\alpha), COMMIT(i,j,\alpha), DIR(i,j,\alpha) \in MAct$
3. $f \in PFORM, i, j \in Ag \Longrightarrow DECL(i,f), ASS(i,j,f) \in MAct$
4. $\gamma\alpha_1, \gamma\alpha_2 \in MAct \Longrightarrow \gamma\alpha_1;\gamma\alpha_2 \in MAct$

The complete language $FORM$ is now defined to contain all the constructs informally described in the previous section, i.e. there are operators representing informational attitudes, motivational attitudes, aspects of actions, and the social traffic between agents.

Definition 4 *The language $FORM$ of formulas is defined to be the smallest set closed under:*

1. $PFORM \subseteq FORM$
2. $\phi, \phi_1, \phi_2 \in FORM \Longrightarrow \neg\phi, \phi_1 \wedge \phi_2, \phi_1 \prec \phi_2 \in FORM$
3. $\phi \in FORM, i \in Ag \Longrightarrow K_i\phi, B_i\phi \in FORM$
4. $\gamma\alpha \in MAct, i \in Ag, \phi \in FORM \Longrightarrow [\gamma\alpha(i)]\phi \in FORM$
5. $\alpha \in Act, \phi \in FORM \Longrightarrow$
   $PREV(\alpha(i)), OPP(\alpha(i)), NEXT(\alpha(i)), NEXT(\phi) \in FORM$
6. $f \in PFORM, \phi \in FORM, i, j \in Ag, \alpha, \alpha_1, \alpha_2 \in Act \Longrightarrow$
   $P_i\phi, Achiev_i\phi, INT_i\alpha, IMP_i(\alpha_1,\alpha_2), O_{ij}\phi, O_{ij}\alpha, auth(i,j), auth(i,f) \in FORM$

The models used to interpret $FORM$ are based on Kripke-style possible worlds models, i.e. the backbone of these models is given by a set Σ of states, and a valuation π on propositional symbols relative to a state. Various relations and functions on these states are used to interpret the different (modal) operators. These relations and functions can roughly be classified in four parts, dealing with the informational level, the action level, the motivational level and the social level, respectively. We assume tt and ff to denote the truth values 'true' and 'false', respectively.

Definition 5 *A model Mo for $FORM$ from the set CMo is a structure $(\Sigma, \pi, I, A, M, S)$ where*

1. *Σ is a non-empty set of states and $\pi : \Sigma \times \Pi \rightarrow \{tt, ff\}$.*
2. *$I = (Rk, Rb)$ with $Rk : Ag \rightarrow \wp(\Sigma \times \Sigma)$ denoting the epistemic alternatives of agents and $Rb : Ag \times \Sigma \rightarrow \wp(\Sigma)$ denoting the doxastic alternatives.*

3. $A = (Sf, Mf, Ropp, Rprev, Rnext)$ *with* $Sf : Ag \times Act \times \Sigma \to \wp(\Sigma)$ *yielding the interpretation of regular actions,* $Mf : Ag \times MAct \times (CMo \times \Sigma) \to (CMo \times \Sigma)$ *yielding the interpretation of meta-actions,* $Ropp : Ag \times \Sigma \to \wp(Act)$ *denoting opportunities,* $Rprev : Ag \times \Sigma \to Act$ *yielding the action that has been performed last and* $Rnext : Ag \times \Sigma \to Act$ *yielding the action that will be performed next.*

4. $M = (Rp, Rep, <, Ri, Ria, Ro)$ *with* $Rp : Ag \times \Sigma \to \wp(\Sigma)$ *denoting implicit preferences,* $Rep : Ag \times \Sigma \to \wp(FORM)$ *yielding explicit preferences,* $< \subseteq FORM \times FORM$ *which is a preference relation on preferences,* $Ri : Ag \times \Sigma \to \wp(Act)$ *denoting intended actions,* $Ria : Ag \times \Sigma \times \wp(Act \times Act)$ *denoting implementation relations between actions and* $Ro : Ag \times Ag \times \wp(\Sigma \times \Sigma)$ *denoting obligations.*

5. $S = (Auth)$ *with* $Auth : Ag \times (Ag \cup PFORM) \to \{tt, ff\}$ *yielding authority relations.*

such that the following constraints are validated:

1. $Rk(i)$ *is an equivalence relation for all* i*, and* $Rb(i,s) \neq \emptyset$*,* $Rb(i,s) \subseteq \{s' \mid (s,s') \in Rk(i)\}$ *and* $(s,s') \in Rk(i) \Longrightarrow Rb(i,s) = Rb(i,s')$*, which ensures that knowledge validates an S5 axiomatisation and belief obeys a KD45 axiomatisation, while agents indeed believe all things they know.*

2. Sf *yields the state-transition interpretation for regular actions. This function satisfies the usual constraints ensuring an adequate interpretation of composite actions in terms of their constituents. The function* Mf *models the model-transforming interpretation of meta-actions. Below we elaborate on the definition of* Mf *for the meta-actions introduced in the previous section.*

3. $Rnext(i,s) \in Ropp(i,s) \subseteq \{\alpha \mid Sf(i,\alpha,s) \neq \emptyset\}$*, which ensures that opportunities are a subset of the actions that are possible by virtue of the circumstances and that the next action performed is an opportunity, and* $Rprev(i,s) = \alpha \Longleftrightarrow \alpha \in Ropp(i,s')$ *for some* s' *with* $s \in Sf(i,\alpha,s')$*, which relates previously executed actions to past opportunities.*

4. $Ri(i,s) \subseteq \{\alpha \mid Sf(i,\alpha,s) \neq \emptyset\}$ *and for all* $s \in \Sigma$ *some* $s' \in \Sigma$ *exists with* $(s,s') \in Ro$*.*

The complete semantics contains an algebraic interpretation of action expresses, based on the action semantics of Meyer [10]. In this abstract we will refrain from the algebraic interpretation of actions and instead interpret actions as functions on states of affairs. For the meta-actions the state-transition interpretation is not adequate, because meta-actions do not change states, but *relations between states*. For instance, in the case of an assertion, the effect is to change the doxastic state of the receiving agent j and nothing else. To formalise this behaviour, we interpret meta-actions as *model-transforming functions*. In the case of an assertion to j, the resulting model will differ from the starting model only in $Rb(j)$, which is the relation embodying j's doxastic states.

Definition 6 *The binary relation* $\models$ *between an element of* $FORM$ *and a pair consisting of a model* Mo *in* CMo *and a state* s *in* Mo *is for propositional symbols, conjunctions and negations defined as usual. Epistemic formulas* $K_i\phi$ *and doxastic formulas* $B_i\phi$ *are interpreted as necessity operators over* Rk *and* Rb *respectively. For the other formulas* $\models$ *is defined as follows:*

$Mo, s \models [\alpha(i)]\phi \iff Mo, s' \models \phi$ *for all* $s' \in Sf(i, \alpha, s)$
$Mo, s \models [\gamma\alpha(i)]\phi \iff Mo', s' \models \phi$ *for all* $Mo', s' \in Mf(i, \alpha, Mo, s)$
$Mo, s \models PREV(\alpha(i)) \iff \alpha \in Rprev(i, s)$
$Mo, s \models OPP(\alpha(i)) \iff \alpha \in Ropp(i, s)$
$Mo, s \models NEXT(\alpha(i)) \iff \alpha \in Rnext(i, s)$
$Mo, s \models NEXT(\phi) \iff Mo, s \models NEXT(\alpha(i)) \rightarrow [\alpha(i)]\phi$ *for all* $\alpha(i)$
$Mo, s \models P_i\phi \iff Mo, s' \models \phi$ *for all* $s' \in Rp(i, s)$ *and* $\phi \in Rep(i, s)$
$Mo, s \models \phi_1 \prec \phi_2 \iff \phi_1 < \phi_2$
$Mo, s \models Achiev_i\phi \iff Mo, s \models [\beta(i)]\phi \wedge OPP(\beta(i))$ *for some* β
$Mo, s \models INT_i\alpha \iff \alpha \in Ri(i, s)$
$Mo, s \models IMP_i(\alpha_1, \alpha_2) \iff (\alpha_1, \alpha_2) \in Ria(i, s)$
$Mo, s \models O_{ij}\phi \iff Mo, s' \models \phi$ *for all* s' *with* $(s, s') \in Ro(i, j)$
$Mo, s \models O_{ij}\alpha \iff Mo, s \models [\mathbf{any}(i)]O_{ij}(PREV(\alpha(i)))$
$Mo, s \models auth(i, x) \iff Auth(i, x) = tt$ *for* $x \in Ag \cup PFORM$

The functions interpreting the special meta-actions are described below in terms of the preconditions and the postconditions for execution of the actions. The precondition describes for which models the model-transforming function has the desired effect and the postcondition describes the model yielded by the application of the meta-action.

DEC The precondition for execution of $DEC(i, \alpha)$ is that for some $\phi \in FORM$, $Goal_i\phi \wedge [\alpha; \beta(i)]\phi$ holds, for some $\alpha, \beta \in Act$ and furthermore no ψ exists such that $P_i\psi$ and $\phi \prec \psi$ hold. Thus agents may only decide to intend to do those actions that fulfil some most preferred goal. As the result of execution of $DEC(i, \alpha)$ the model is changed in such a way that $INT_i\alpha$ holds in the resulting model.

COMMIT Since our agents are assumed to be sincere, having the intention to do α is a precondition for execution of $COMMIT(i, j, \alpha)$ by agent i. The effect of the commitment is that the model is changed in such a way that $O_{ij}\alpha$ holds afterwards.

DIR The preconditions for execution of $DIR(i, j, \alpha)$ by i are given by $auth(i, j)$. This implies that agent i should have the authority over j before it can order it around. The effect of such an action is that j is committed to i to perform α, which is implemented in a way similar to the implementation of the $COMMIT$ action.

DECL The action $DECL(i, f)$ has as precondition that i is authorised to declare f, i.e. $auth(i, f)$ holds. Execution of an action $DECL(i, f)$ in a certain state of a model will be a modification of the valuation π such that f is true in all the resulting states of the resulting models.

ASS The precondition for $ASS(i, j, f)$ is that $auth(i, j)$ holds. Furthermore i is demanded to believe f, i.e. $B_i f$ should hold. This implies in particular that agents are not allowed to lie, i.e. spread around rumours that they themselves do not even believe. As the result of executing $ASS(i, j, f)$ by i in some state s, the model under consideration is modified such that $Rb(j, s)$ contains only states in which f is true, which indeed implies that $B_j f$ holds in s in the resulting model.

In the following proposition we summarise the descriptions of the meta-actions as given above in a formal way. Even though this proposition is not to be proved using the (sketchy) semantics presented above, it does hold, and can be proved to do so, in the full semantics.

Proposition 1 *For all $i, j \in Ag$, $\alpha \in Act$ and $f \in PFORM$ we have:*

- $\models INT_i\alpha \rightarrow [COMMIT(i,j,\alpha)(i)]O_{ij}\alpha$
- $\models auth(i,j) \rightarrow [DIR(i,j,\alpha)(i)]O_{ji}\alpha$
- $\models auth(i,f) \rightarrow [DECL(i,f)(i)]f$
- $\models auth(i,j) \wedge B_i f \rightarrow [ASS(i,j,f)(i)]B_j f$

4 Related approaches

In this section we briefly indicate the main differences between our approach and two standard approaches to model rational agents, viz. the framework proposed by Cohen & Levesque [1] and the BDI-architecture of Rao & Georgeff [13].

The main difference between our approach and the one of Cohen & Levesque is that they define intentions in terms of goals and beliefs. We agree with this approach when it concerns intentions on propositions. However, in contrast with Cohen & Levesque we do not take a goal to be a primitive notion. Because they take a goal to be primitive, they have to define different types of goals in order to define the persistence of a goal, the achievability of a goal, etc. All these properties are direct consequences of our definition of a goal in terms of preferences and achievabilities. Furthermore, whereas we define the intention to perform an action as primitive, Cohen & Levesque define the intention to perform an action as the goal to reach a state where that action has been performed. Although both types of intentions of Cohen & Levesque are based on the notions of goals and beliefs, the relation between the intention to reach a certain state and the intention to perform an action is not clear; in fact these notions seem to be unrelated. However, it seems desirable that the intention to reach a certain goal induces the intention to perform an action which helps to reach that goal. In our approach this relation is established through the notion of decisions. A goal may induce a decision, which on its turn induces the intention to perform an action. Another relation that is unclear in the theory of Cohen & Levesque is that between an intended action and the action that is actually performed. The only relation given is that the intended action should be the same as the action whose goal it is to be performed, which in itself does not mean anything for the actual course of events. In our approach we introduce the notion of an intention relation between actions, which introduces a loose, and intuitively acceptable, coupling between intended actions and the actions that are actually performed.

The last point also shows one of the main differences between our framework and that of Rao & Georgeff. In their framework it holds that if an agent intends to perform an action it will also actually perform the action. To avoid making the intention operator into a temporal operator they introduce the notion of a successful performance of an action and a failed performance of an action. However, the relation between a successful performed event and an event that failed to be performed is unclear. Can this be any other event? Can it include the event itself? At present the best one can say if an event has been performed (either successful or failed) is that *some* event has been performed.

A last, rather important, point of difference between our framework and the other two is the fact that we also include the social level, which we consider essential when formalising (multiple) agents, but is only briefly mentioned by Cohen & Levesque, and not considered at all by Rao & Georgeff.

5 Conclusions

In this extended abstract we presented an informal overview and a sketchy formalisation of the concepts that we consider essential to model rational agents. In our very flexible and highly expressive framework we propose a variety of concepts, which are roughly situated at four different levels: the informational level, where knowledge and belief are considered, the action level, where we consider various aspects of action, the motivational level, where we dealt with concepts like preferences, goals, intentions, commitments and obligations, and the social level, which is concerned with the social traffic between agents and where we formalised various kinds of speech acts. We characterised our framework by pointing out the differences with the frameworks of Cohen & Levesque and Rao & Georgeff.

Future work will mainly be concerned with incorporating Meyer's action semantics in the semantic framework sketched here, and filling in other missing details. We will furthermore focus on the possible combinations of different concepts, and the (unwarranted) consequences of such combinations. A nice example of an unwarranted consequence of combining epistemic and deontic concepts is the problem that if one ought to know that one's partner commits adultery then it follows that this partner ought to commit adultery, which seems to be a highly undesirable consequence.

References

[1] P. Cohen and H. Levesque. Intention is choice with commitment. *Artificial Intelligence*(42):213-261, 1990.

[2] F. Dignum, H. Weigand. Communication and Deontic Logic. In R. Wieringa and R. Feenstra, eds, *Information systems, correctness and reusability*, pages 241-260, World Scientific, London, 1995.

[3] F. Dignum and H. Weigand. Modelling communication between cooperative systems. In J. Iivari et al., *Advanced information systems engineering*, pages 140-153, Springer, 1995.

[4] F. Dignum, J.-J.Ch. Meyer, R. Wieringa and R. Kuiper. A modal approach to intentions, commitments and obligations: intention plus commitment yields obligation. In *Proceedings of the DEON'96 Workshop on deontic logic in computer science*, Lisbon, Jan. 1996.

[5] D. Harel. First Order Dynamic Logic. LNCS 68, Springer, 1979.

[6] W. van der Hoek, B. van Linder and J.-J.Ch. Meyer. A logic of capabilities. In Nerode and Matiyasevich, eds, *Proceedings of LFCS'94*, LNCS 813, pages 366-378.

[7] W. van der Hoek, B. van Linder and J.-J.Ch. Meyer. Using Modal Logic to Model Rational Agents. In *Proceedings of the First International Workshop of Decentralized Intelligent Multi Agent Systems (DIMAS'95)*, pages 215–224.

[8] B. van Linder, W. van der Hoek and J.-J.Ch. Meyer. Communicating rational agents. In Nebel and Dreschler-Fisher, eds, *Proceedings of KI'95*, LNCS 861, pages 202–213.

[9] B. van Linder, W. van der Hoek and J.-J.Ch. Meyer. How to motivate your agents. On Formalising Preferences, Goals and Commitments. This volume.

[10] J.-J.Ch. Meyer. A different approach to deontic logic. In *Notre Dame Journal of Formal Logic*(29):109–136, 1988.

[11] J.-J.Ch. Meyer and W. van der Hoek. *Epistemic Logic for AI and computer science*, CUP, 1995.

[12] R. Moore. A formal theory of knowledge and action. In J. Hobbs and R. Moore, eds, *Formal theories of the commonsense world*, pages 319-358, Ablex Publ. Comp., 1985.

[13] A.S. Rao and M.P. Georgeff. Modeling rational agents within a BDI-architecture. In J. Allen et al., eds, *Proceedings of KR'91*, pages 473-484, Morgan Kaufmann, 1991.

[14] J.R. Searle. Speech Acts. CUP, 1969.

[15] Y. Shoham. Agent Oriented Programming. *Artificial Intelligence*(60):51–92, 1993.

[16] H. Weigand, E. Verharen and F. Dignum. Integrated semantics for information and communication systems. In R. Meersman and L. Mark, eds, *Proceedings working conference on data semantics DS-6*. To be published.

[17] R. Wieringa, J.-J.Ch. Meyer and H. Weigand. Specifying dynamic and deontic integrity constraints. *Data & knowledge engineering*(4):157-189, 1989.

[18] M. Wooldridge. *The logical modelling of computational multi-agent systems*. Ph.D. thesis, UMIST, Manchester, October 1994.