F. Dignum and B. van Linder. Modelling social agents: Communication as actions. In M. Wooldridge J. Muller and N. Jennings, editors, *Intelligent Agents III (LNAI-1193)*, pages 205--218. Springer-Verlag, 1997.

Modelling Social Agents: Communication as Action

Frank Dignum* Bernd van Linder†

* Faculty of Mathematics & Computer Science, Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
dignum@win.tue.nl

Department of Computer Science, Utrecht University
P.O. Box 80.089 Utrecht, The Netherlands
bernd@cs.ruu.nl

Abstract. In this paper we present a formal framework for social agents. The social agents consist of four components: the information component (containing knowledge and belief), the action component, the motivational component (where goals, intentions, etc. play arole) and the social component (containing aspects of speech acts and relations between agents). The main aim of this work was to describe all components in a uniform way, such that it is possible to verify each component separately but also formally describe the interactions between the different components. E.g. the effect of a speech act on the believes of an agent or on the commitment to a goal it pursues.

1 Introduction

The formalization 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 [15] in which knowledge and actions are considered. Over the years contributions have been made on both *informational* attitudes like knowledge and belief [14] and *motivational* attitudes like intentions and commitments [2, 5].

In our basic framework [8, 12] 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.

In [5, 11] we dealt with the motivational attitudes of agents. In these papers we defined the concepts of *wishes*, *goals*, *intentions*, and *commitments* or *obligations*. By combining this formalization with the basic framework it is for instance possible to model the fact that an agent prefers some situation to hold and it also knows that it is able to achieve 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 bring about its goal.

Finally, in [4, 10] we formalized communication between agents. In this theory we can show 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. The examples show that the communication can change both the informational attitudes of agents as well as the behaviour of agents.

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, wish, goal, decision, intention, commitment, obligation* and *communication*. Following [5, 11] we base this model on dynamic logic [7], 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 formalizations proposed by Cohen & Levesque [2] and Rao & Georgeff [19].

In this extended abstract we will describe the different components and their relationships informally. Even though we will give a precise definition of our language and the models used to interpret this language, we only have room to sketch the actual semantics.

2 The Concepts

The concepts that we formalize can roughly be situated at four different levels: the informational level, the action level, the motivational level and the social level. We will introduce the concepts of each of these levels in the following subsections.

2.1 The Informational Level

At the informational level we consider both knowledge and belief. Many formalizations 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 [8, 24] 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 dynamic logic as described by Harel in [7]. We use $[\alpha(i)]\phi$ to indicate 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. In [8] we also used an *ability* predicate. This predicate was used to indicate that an agent has the inherent possibility to perform a certain action (at a certain place and time). So, whether an agent can perform an action depends on both its inherent ability and the external opportunity to perform the action. In the present paper the abilities of agents do not play a mayor role. Therefore we left them out in the formalization and assume (for the present) that the abilities of an agent are incorporated in the opportunities for an agent to perform an action.

Besides the OPP operator, which already has a temporal flavor 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. Note that the standard dynamic logic operator ”<>” can only be used to indicate the *possible* previous action. That is, we can use $true < \alpha >$ to indicate that the present state can have been reached by performing α in a previous state. However, to denote the *actual* previous action a new operator is needed! See [3] for a more in depth discussion of this issue. The same holds for the $NEXT$ operator for actions. We also define a more traditional $NEXT$ operator on formulas in terms of the $NEXT$ operator on events.

$$NEXT(\phi) \text{ iff } NEXT(\alpha(i)) \rightarrow [\alpha(i)]\phi$$

This means that the formula ϕ is true in all next states iff when an action $\alpha(i)$ is performed next the formula ϕ is true after the performance of $\alpha(i)$.

2.3 The Motivational Level

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

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

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

A goal is now formally defined as a wish that does not hold but is achievable:

$$Goal_i\phi \equiv W_i\phi \wedge \neg\phi \wedge Achiev_i\phi$$

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 the wish now holds, 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. Goals can either be known or unconscious goals of an agent. Most goals will be known, but we will later on see that goals can also arise from commitments and these goals might not be known explicitly.

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 reach a certain state is defined as the goal to reach that state. 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 wishes 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 wishes 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 $INT_i\alpha$. The above is described formally by

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

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 fulfill 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 not be said anymore 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

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

being valid.

The last concept that we consider at the motivational level is that of commitment. This concept is also part of the social level if the commitment is made towards another agent. As the result of i performing a $COMMIT(i, j, \alpha)$ action the formula $O_{ij}\alpha$ becomes true. (See [4] for more details.) I.e. by committing itself to an action, an agent i obliges itself towards j to perform the action α. The commitment can be a private one if j is the same as i. In that case the result is an obligation of the agent towards itself to perform the action. Although the obligation does not ensure the actual performance of the action by the agent, it does have two practical consequences. 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). The registration of the violation is done through the introduction of a deontic relation between the worlds. 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 [5].

Secondly, an obligation to perform an action leads to the goal of having performed the action. Formally this is achieved with the following formula:

$$O_{ij}\alpha \rightarrow W_i(PREV(\alpha(i)))$$

Note that this is sufficient to create a goal, because $PREV(\alpha(i))$ does not hold presently (the action is not performed yet when the obligation arises) and it is achievable (by performing the action $\alpha(i)$.

We only consider sincere agents and therefore we assume that an agent can only commit itself to actions that it intends to do eventually, i.e. intention provides a precondition for commitment.

2.4 The Social Level

The $COMMIT$ described in the previous section is one of the four types of *speech acts* [20] that play a role at the social level. Speech acts are used to communicate between agents. 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 ‘Pay the bill before next week’. A typical example of a declaration is the utterance ‘Herewith you are granted permission to access the database’, and a typical assertion is ‘I tell you that the earth is flat’. In particular for directions and declarations the agent uttering the statement should have some kind of authorisation for the speech act to have any effect. We formalize this authority relation through a binary predicate $auth$; $auth(i, \alpha)$ means that agent i is authorised to perform α. The speech acts are formalized as meta-actions: $DIR(i, j, \alpha)$ formalizes that agent i directs agent j to perform α, $DECL(i, f)$ models the declaration of i that f holds, and $ASS(i, j, f)$ formalizes the assertion of i to agent j that f holds.

A directive from agent i to agent j to perform α results in an obligation of j towards i to perform that action *if* agent i was authorised to give the order in the first place. These authorisations can come into existence in several ways. First there may be a hierarchical relation between the two agents which is fixed for a period of time (e.g. manager-employee relations). This automatically incurs the authorisation for some directives (e.g. the manager can order the employee to perform some job). Another way to create authorisations is by agent j giving an explicit or implicit authorisation to i to give him some directives. For example, when agent i orders a product from agent j it implicitly gives the authorisation to agent j for ordering i to pay for the product (after delivery).

The following axiom holds for creating authorisations:

$$[DECL(i, auth(j, DIR(j, i, \alpha(i))))]auth(j, DIR(j, i, \alpha(i)))$$

which states that an agent i can create authorisations for an agent j concerning actions that i has to perform.

The following formulas hold for the effects of commitments, orders and declaratives:

- $[COMMIT(i,j,\alpha)(i)]O_{ij}\alpha$
- $auth(i, DIR(i,j,\alpha)(i)) \rightarrow [DIR(i,j,\alpha)(i)]O_{ji}\alpha$
- $auth(i, DECL(i,f)(i)) \rightarrow [DECL(i,f)(i)]f$

That is, no precondition has to hold for a commitment to result in an obligation afterwards.

A directive from agent i results in an obligation of agent j (towards i) if agent i was authorised to give the order.

Finally a declaration can change the state of the world if the agent making the declaration is authorised to do so. (This is the only speech act that has a direct effect on the states other than a change of the mental attitudes of the agents!).

Assertions can be used to transfer beliefs from one agent to another. Often these dialogues follow a protocol in which agent i first states some belief after which agent j can accept it, reject it or give a counter argument. At this place we do not want to describe the rules governing this types of protocols but only the messages. However, it is possible to describe the different moves that are allowed in these protocols. Every reply of j can be described by another assertion as follows:

- $[ASS(i,j,f)(i)][ASS(j,i,f)(j)]K_j B_i f \wedge K_i B_j f$ (accept)
- $[ASS(i,j,f)(i)][ASS(j,i,\neg f)(j)]K_j B_i f \wedge K_i B_j \neg f$ (reject)
- $[ASS(i,j,f)(i)][ASS(j,i,\neg f \wedge g)(j)]K_j B_i f \wedge K_i B_j \neg f \wedge g$ (counter)

In this case we model the acceptance of j of a statement made by i as an assertion by j of the same statement. In practice this speech act can, of course, be abbreviated by a special ACCEPT message. In the same way a rejection of a statement is modelled by asserting the opposite of the original statement. This can, in practice, be abbreviated to a REJECT message.

Note that agent j does not automatically believe what agent i tells him. It only knows that agent i believes it (we assume sincere agents). The only way to directly transfer a belief is when agent i is authorised to make a statement:

$$auth(i, ASS(i,j,f)(i)) \wedge B_i f \rightarrow [ASS(i,j,f)(i)]B_j f$$

Usually this situation arises when agent j first requested some information from i. Such a request for information (modelled by a directive without authorisation) gives an implicit authorisation on the assertions that form the answer to the request.

3 A Sketch of a Formalization

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.

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

Definition 1. 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 these meta-actions it is not always clear whether they can be performed in parallel or what the result is of taking the negation of a meta-action. This area needs a more thorough study in the future. For simplicity, we restrict ourselves in this paper to closing the set $MAct$ under sequential composition.

Definition 2. 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. $\gamma\alpha_1, \gamma\alpha_2 \in MAct \Longrightarrow \gamma\alpha_1; \gamma\alpha_2 \in MAct$

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

Definition 3. The language $FORM$ of formulas is defined to be the smallest set closed under:

1. $PFORM \subseteq FORM$
2. $\phi, \phi_1, \phi_2 \in FORM \Longrightarrow \neg\phi, \phi_1 \wedge \phi_2 \in FORM$
3. $\phi \in FORM, i \in Ag \Longrightarrow K_i\phi, B_i\phi \in FORM$
4. $\gamma\alpha \in MAct, \phi \in FORM \Longrightarrow [\gamma\alpha]\phi \in FORM$
5. $\psi, \phi \in FORM, i, j \in Ag, \Longrightarrow [DECL(i, \psi)]\phi, [ASS(i, j, \psi)]\phi \in FORM$
6. $[\gamma\alpha]\phi, [\gamma\beta]\psi, \theta \in FORM \Rightarrow [\gamma\alpha; \gamma\beta]\theta \in FORM$
7. $\alpha \in Act, \phi \in FORM \Longrightarrow PREV(\alpha), OPP(\alpha), NEXT(\phi) \in FORM$
8. $\phi, \psi \in FORM, i, j \in Ag, \alpha, \alpha_1, \alpha_2 \in Act \Longrightarrow W_i\phi, \psi <_i \phi, INT_i\alpha,$
   $IMP_i(\alpha_1, \alpha_2), O_{ij}(\alpha), auth(i, \alpha) \in FORM$

The models used to interpret $FORM$ are based on Kripke-style possible worlds models. That is, 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 various (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 4. 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 $\pi : \Sigma \times \Pi \to \{tt, ff\}$.
2. $I = (Rk, Rb)$ with $Rk : Ag \to \wp(\Sigma \times \Sigma)$ denoting the epistemic alternatives of agents and $Rb : Ag \times \Sigma \to \wp(\Sigma)$ 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 wishes, $Rep : Ag \times \Sigma \to \wp(FORM)$ yielding explicit wishes, $<\subseteq Ag \times \Sigma \to FORM \times FORM$ which is a preference relation on wishes, $Ri : Ag \times \Sigma \to \wp(Act)$ denoting intended actions, $Ria : Ag \times \Sigma \to \wp(Act) \times \wp(Act)$ denoting implementation relations between actions and $Ro : Ag \times Ag \to \wp(\Sigma \times \Sigma)$ denoting obligations.
5. $S = (Auth)$ with $Auth : Ag \times \wp(MAct) \to \{tt, ff\}$ yielding authorisations

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-action. 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. Furthermore, $Rprev(i, s) = \alpha$ iff $\alpha \in Ropp(i, s')$ 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 semantics of action expresses, based on the action semantics of Meyer [13]. In this abstract we will abstract 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 they change relations between states. For instance, in the case of an assertion, the effect is to change the doxastic state of the receiving agent, and nothing else. To formalize this behaviour, we interpret meta-actions as model-transforming functions. In the case of an assertion, the resulting model will differ from the starting model in the doxastic accessibility relation of the receiving agent.

Definition 5. 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:

$$\begin{array}{ll}
Mo, s \models [\alpha(i)]\phi & \iff Mo, s' \models \phi \text{ for all } s' \in Sf(i, \alpha, s) \\
Mo, s \models [\gamma\alpha(i)]\phi & \iff Mo', s' \models \phi \text{ for all } Mo', s' \in Mf(i, \alpha, Mo, s) \\
Mo, s \models PREV(\alpha(i)) & \iff \alpha \in Rprev(i, s) \\
Mo, s \models OPP(\alpha(i)) & \iff \alpha \in Ropp(i, s) \\
Mo, s \models NEXT(\alpha(i)) & \iff \alpha(i) \in Rnext(i, s) \\
Mo, s \models W_i\phi & \iff Mo, s' \models \phi \text{ for all } s' \in Rp(i, s) \text{ and } \phi \in Rep(i, s) \\
Mo, s \models \psi <_i \phi & \iff (\psi, \phi) \in < (i, s) \\
Mo, s \models INT_i\alpha & \iff \alpha \in Ri(i, s) \\
Mo, s \models IMP_i(\alpha_1, \alpha_2) & \iff (\alpha_1, \alpha_2) \in Ria(i, s) \\
Mo, s \models O_{ij}(\phi) & \iff Mo, s' \models \phi \text{ for all } s' \text{ with } (s, s') \in Ro(i, j) \\
Mo, s \models O_{ij}(\alpha) & \iff Mo, s \models [\mathbf{any}(i)]O_{ij}(PREV(\alpha(i))) \\
Mos, \models auth(i, \alpha) & \iff Auth(i, \alpha, s) = tt
\end{array}$$

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 on 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 $\phi \in FORM$, $Goal_i\phi \wedge [\alpha(i); \beta(i)]\phi$ holds, for some $\alpha(i), \beta(i) \in Act$ and furthermore no ψ exists such that $P_i\psi$ and $\phi < \psi$ hold. Thus agents may only decide to intend to do those actions that fulfill 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.

$COMMIT$ There are no preconditions 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 $auth(i, DIR(i, j, \alpha))$. 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. (Only some civil servants can declare people to be married in the Netherlands). 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. Note that whenever f is inconsistent no model results.

ASS The precondition for $ASS(i, j, f)$ is that i is demanded to believe f, i.e. $B_i f$ should hold. This implies in particular that agents are not allowed to gossip, i.e. spread around rumors that they themselves do not even believe. As the result of executing $ASS(i, j, f)$ by i in some state s, two cases arise. If $auth(i, ASS(i, j, f))$ holds then 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. Regardless whether i is authorised to make the assertion the model under consideration is modified such that $Rk(j, s)$ contains only states s' such that $Rb(i, s')$ contains only states in which f is true. This implies that $K_j B_i f$ holds in the resulting model.

4 Related Approaches

In this section we very briefly indicate the main differences between our approach and three other approaches to model rational agents, viz. the framework proposed by Cohen & Levesque [2], the BDI-framework of Rao & Georgeff [19] and the theoretical framework for multi-agent systems of M. Singh [21]. After that we will shortly compare our approach with the other work in this volume on communicating agents.

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, we do not take a goal to be a primitive notion as is the case in the approach of Cohen & Levesque. Because they take a goal to be a primitive notion 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. The distinction between goals and preferences allows for a bigger flexibility than possible in the approach of Cohen & Levesque. 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 it is not clear what is the relation between the intention to reach a certain state and the intention to perform an action; 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 is needed to reach that goal. In our approach this relation is established through the notion of decisions. A goal can induce a decision. The decision then induces the intention to perform an action. Another relation that remains unclear in the theory of Cohen & Levesque is that between an intended action and the action that is actually performed. The only relation they give 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 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. They weaken this assumption in [9], where they investigate different strategies with respect to commitment to a goal. An agent can be single minded when it always performs its intended actions and open minded when it discards its goals on the basis of new information it receives. This work, however, is of a very practical nature and does not have a theoretical counterpart. Therefore it remains unclear how this would be reflected in the BDI framework.

In the BDI framework they avoided making the intention operator into a temporal operator by introducing 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, but is only briefly mentioned by Cohen & Levesque, and not considered at all by Rao & Georgeff.

The social level is treated in the theory of Singh, in the form of communication between agents. The framework of Singh is based on CTL^*, a branching-time logic. The main difference with our framework is that he does not incorporate deontic relations between agents. It means that the success of a speech act is not defined in the same way as we do. A directive is successful if it results in the hearer intending to perform the action that it was ordered to perform. This is much stronger than in our framework. Of course, an obligation can lead to a goal, but only if the hearer places the directive above the plan it is currently performing.

How does our work compare with the other current work on communicating agents? From the papers in this volume, only Bretier & Sadek [1] formally define the effects of communicative actions. However, this theory is limited to effects on the believes and intentions of agents. We have shown that the concept of obligation also forms an important ingredient of the effects of some speech acts. That is, directives and commissives always result in obligations. It must be remarked that the application for which the theory in [1] is used does not need these illocutions.

Another difference between our work and that of [1] is that our theory is geared to communication between agents while their theory is geared towards the management of human-agent dialogue. The same holds for the theory of Traum [22]. Although many aspects are compatible, there are also some differences. One of the main differences is that in the communication between humans and agents the detection of the illocution of messages is not trivial and important to steer the dialogue in a natural way. Agents will use a formal language in which the illocution of each message is clear.

Very interesting in the work of Traum is that he also recognizes the importance of obligations between the participants in a dialogue. Besides mutual knowledge and believe this is an important relation, because it indicates the expectations that the participants have of each other.

Finally the work of Noriega & Sierra [17] comes very close in spirit to our work. They also try to give one formal framework for the different components of communicating agents. Their theory is more flexible than ours in that they allow different agents to use different languages and even the different components might use different inference rules. The relation between the components is given by so-called bridge rules. We assume a uniform language for all agents and components. This provides for an automatic integration of the components. The main point in which the theory in [17] is lacking compared to ours is the semantics of the speech acts themselves and their effects.

5 Conclusions

In this paper we presented an informal overview and a sketchy formalization 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 preferences, goals, intentions, etc., and the social level, which is concerned with the social traffic between agents.

The resulting multi-modal logic is quite complex. However, we want to make two remarks about the logical formalism. First, it is not our aim to build an automated theorem prover that can prove theorems in this very rich logic. The use of a logical formalism gives the opportunity to automatically generate the logical effects of a sequence of steps in a protocol. These could be subsequently implemented in a more efficient formalism. The logical description, however, can be used as a very general and precise specification of that implementation.

Secondly, the use of logic forces a very precise formal description of the communication. It is very important that this is realized when the communication protocols are automatized. (As is the aim in communication between agents). If the communication is automatic it becomes very important to know the exact effects of the messages. What is the knowledge of each agent and what are its obligations (resp. expectations).

We admit that the logical formulas get very complicated and are not very readable. However, it is easy to define suitable abbreviations for standard formulas. At least, working this way, it is clear what these abbreviations mean exactly!

In subsequent work we want to show how communication protocols that are used in more practical work like [16, 18] can be given a formal semantics in our framework. For the basic illocutions used in the ADEPT system [18] this has been done in [6] and has led to the discovery that the seemingly simple ACCEPT message has unexpected results.

Also we want to define an agent architecture for communicating agents that adhere to our theory. Some groundwork in this respect has been done in [23]. Currently we are working on an implementation of this framework to test the ideas.

Acknowledgements

We would like to thank the anonymous referees for their useful remarks and the workshop participants whose remarks contributed to this greatly improved final version of the paper.

References

1. P. Bretier and M. D. Sadek. A rational agent as the kernel of a cooperative spoken dialogue system: Implementing a logical theory of interaction. In J. P. Müller, M. J. Wooldridge, and

N. R. Jennings, editors, *Intelligent Agents III — Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages (ATAL-96)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg, 1996. In this volume.

2. P. Cohen and H. Levesque. Intention is choice with commitment. *Artificial Intelligence*, vol.42, pages 213-261, 1990.
3. F. Dignum. Using Transactions in Integrity Constraints: Looking forward or backwards, what is the difference? In *First International Workshop on Applied Logic: Logic at Work*, Amsterdam, 1992.
4. F. Dignum and H. Weigand. Modelling communication between cooperative systems In J. Iivari et al., *Advanced information systems engineering*, pages 140-153, Springer, 1995.
5. F. Dignum, J.-J.Ch. Meyer, R. Wieringa and R. Kuiper. A modal approach to intentions, commitments and obligations: intention plus commitment yields obligation. *DEON'96 Workshop on deontic logic in computer science*, Lisbon, Jan. 1996.
6. F. Dignum. Social Interactions of Autonomous Agents; Private and Global Views on Communication. Submitted to ModelAge'97 workshop.
7. D. Harel. First Order Dynamic Logic. LNCS 68 Springer, 1979.
8. 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.
9. D. Kinny and M. Georgeff. Commitment and Effectiveness of Situated Agents. In *Proceedings International Joint Conference on Artificial Intelligence*, Sydney, Australia, pages 82-88.
10. 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.
11. B. van Linder, W. van der Hoek and J.-J.Ch. Meyer. How to motivate your agents. On making promises that you can keep. In Wooldridge, Müller and Tambe, eds, *Intelligent Agents II*, LNCS 1037, pages 17–32.
12. B. van Linder *Modal Logics for Rational Agents*, PhD Thesis, Utrecht University, 1996.
13. J.-J.Ch. Meyer. A different approach to deontic logic. In *Notre Dame Journal of Formal Logic*, vol.29, pages 109–136, 1988.
14. J.-J.Ch. Meyer and W. van der Hoek. *Epistemic Logic for AI and computer science*, CUP, 1995.
15. 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.
16. J. P. Müller. A cooperation model for autonomous agents. In J. P. Müller, M. J. Wooldridge, and N. R. Jennings, editors, *Intelligent Agents III — Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages (ATAL-96)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg, 1996. In this volume.
17. P. Noriega and C. Sierra. Towards layered dialogical agents. In J. P. Müller, M. J. Wooldridge, and N. R. Jennings, editors, *Intelligent Agents III — Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages (ATAL-96)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg, 1996. In this volume.
18. T. J. Norman, N. R. Jennings, P. Faratin, and E. H. Mamdani. Designing and implementing a multi-agent architecture for business process management. In J. P. Müller, M. J. Wooldridge, and N. R. Jennings, editors, *Intelligent Agents III — Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages (ATAL-96)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg, 1996. In this volume.
19. 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.
20. J.R. Searle. Speech Acts. CUP, 1969.
21. M. Singh. Multiagent Systems. LNAI 799, Springer-Verlag, 1994.

22. D. R. Traum. A reactive-deliberative model of dialogue agency. In J. P. Müller, M. J. Wooldridge, and N. R. Jennings, editors, *Intelligent Agents III — Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages (ATAL-96)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg, 1996. In this volume.
23. E. Verhagen, F. Dignum and H. Weigand. A Language/Action Perspective on Cooperative Information Agents. In F. Dignum at al., eds, *Communication Modeling - The Language/Action Perspective (LAP-96)*, electronic Workshops in Computing, Springer-Verlag, London, 1996.
24. R. Wieringa, J.-J.Ch. Meyer and H. Weigand. Specifying dynamic and deontic integrity constraints. *Data & knowledge engineering*, vol.4, pages 157-189, 1989.

This article was processed using the LaTeX macro package with LLNCS style