F. Dignum and R. Conte. Intentional agents and goal formation. In M. Singh et.al., editor, *Intelligent Agents IV (LNAI 1365}*, pages 231-244, Springer Verlag, 1998.

Intentional Agents and Goal Formation

Frank Dignum[*] Rosaria Conte[†]

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

[†] Division of AI, Cognitive and Interaction Modelling
PSS (Project on Social Simulation)
IP/Cnr, V.LE Marx 15 - 00137 Roma, Italy
rosaria@pscs2.irmkant.rm.cnr.it

Abstract. This paper is about a fundamental aspect of intentional action, namely the process of goal formation. Existing formal theories of agents are found essentially inadequate to account for the formation of new goals and intentions of the agent; on the other hand, the formation of new goals is often viewed as an essential feature of autonomous agents. Autonomous goal-formation is described thanks to the interplay between existing (built-in) goals and new beliefs. A general rule for goal formation is then formally expressed in terms of a language (FORM) developed for treating properties of autonomous agents. More specific applications of this rule to the social domain are examined, in particular to conformity and help.

1 Introduction

In [16] an answer to the question what actually constitutes an ”agent” is given. Although the authors do not provide a comprehensive definition they do provide some characteristics of an agent. The first characteristic is that of autonomy. An agent should operate without the direct intervention of humans and have control over its actions and internal state.

Generally, this feature is integrated into existing agent models. Actually many programs are called agents simply because they exhibit some autonomy. However, Wooldridge and Jennings also mention another characteristic for agents that we find to be crucial. One should be able to ascribe intentional attitudes (like desires, goals, intentions, wants, etc.) to agents. And ,more important, these intentional attitudes should be instrumental to explaining the behaviour of the agent.

In fact the use of intentional attitudes to explain the behaviour of an agent tells us something about the complexity of the agent. If a program or machine is very simple its behaviour can be very well explained without ascribing intentional attitudes to it. However, when a program or machine is very complicated we need to use a higher level of abstraction (like intentional attitudes) to explain its behaviour.

In this paper we will only consider these so-called *intentional agents*.

[0] The research is partially supported by ESPRIT BRA 8319 ModelAge.

Of course, if the behaviour of a program is explained by ascribing intentional attitudes to it, the program should also comply to the intuitive characteristics of these intentional attitudes. Otherwise the metaphor would have no use. Therefore much research is done about the characteristics and connections of the intentional attitudes.

A weak point about formal theories of intentions and intended actions is their failure to account for the formation of intentions. This remark may look arbitrary to those readers who are familiar with formal agent theories, e.g. BDI architectures [14] in which intentions are defined as a subset of beliefs and desires (see also [5]).
However, BDI architectures and, more generally, agent theories take for granted the agents choice. Within the logic-based framework, by and large, intentions are a subset of motivations which are chosen for action. Even though Cohen and Levesque [5] and Kinny and Georgeff [12] extensively discuss the abandonment of intentions and goals they do not indicate what the agent should do when a goal is dropped. How does it get a new goal or intention? Do they have a large stack of goals from which they pop a new goal every time?

Of course in many implemented systems some type of goal formation takes place. However, it is usually extremely simple through some rule that relates a perceived external condition with a new goal like in [11] or through a received request (or command) like in [8]. In other frameworks it is something that is done by the user at the start of the system. The agent only decides which goal is active at a certain moment based on the circumstances (see e.g. [10]). We do not intend to disqualify this research, but rather advocate an extension of present research to include the matter of goal formation explicitly in the theories.

This paper does not aim to provide a complete theory of intention formation, but rather clarify some (theoretical) premises for such an ambitious task. The main claim of this paper is that a general notion of goal is a fundamental requirement for a formal theory of intentional action.

A goal will be here defined as a state of the world represented in an agent's mind, which the agent wants to become true; however, a goal should not be intended as chosen by the agent for action ([6]). An agent may have a goal without being oriented to action, neither positive nor negative.
The goal will feed a mental process of goal-dynamics [4], including goal-revision and abandonment, enabling the system to respond to the requirements of the external world by generating goals and checking the conditions and convenience for their achievement. Goals only possibly lead to intentions, and thereby to actions; at the same time, they are necessary for intentions to arise.

The rest of the paper is organized as follows. In section 2 we will describe some related work and indicate the pros and cons of each theory. We will also indicate the open issues with regard to this work. In section 3 we will give an informal description of our theory on goal formation. In section 4 the formal basis for this theory will be sketched. We finish the paper with some conclusions in section 5.

2 Related work

The following two sub-sections discuss some formal attempts to fill the gap between intentional concepts and modalities, and the motivational ones: the BDI architecture worked out by Georgeff and his collaborators and Cohen and Levesque's theory of rational interaction.

2.1 The BDI architecture

The BDI architecture (cf. [14, 15, 9]) describes agents in terms of three primitive modalities: beliefs (B), desires, (D), and intentions (I). The temporal dimension is treated thanks to branching structures. However, the modalities for beliefs, desires, and intentions are defined as usual in terms of possible words semantics. Therefore, possible worlds are defined as belief-, desire-, and intention-accessible worlds, and are represented as time branches, or potential courses of events, which successive actions will prune.

Pros

Among the advantages of this representation, one deserves special mention: the branching structure allows the well-known phenomenon of overcommitment to be avoided. Within uni-linear models of intentional action, goals are closed under consequence; in other words, any believed consequence of one's goals is wanted also. With a branching structure, what is believed to be a consequence on one given path, but is not believed to occur in other paths, is not necessarily intended. In such a way, one can essentially represent different possible scenarios. But what is interesting is that only a subset of such parallel scenarios is represented as following from the agent's intentions. In fact, there is an implication order in the three modalities: the intention-accessible worlds represent a sub-set of both the desire-accessible worlds, and of the belief-accessible worlds.

Cons

Some considerations can be drawn from this model.

1. Somewhat counter-intuitively, intention is here defined as a primitive modality, which is not semantically decomposed into more general notions.
2. Pro-attitudes can be categorised as a subset of desires, that is, endogenous motivations. Apparently, the model rules out the possibility that intentions arise from other inputs (requests, commands, norms, etc.).
3. A selection among desires is presupposed rather than accounted for, despite the complex tests such a selection requires (concerning how likely a given desire is to be realized).
4. It is far from clear whether the model allows for reasoning about instrumental action to be expressed; how to express the notion of a sub-goal which is not yet a subsequence of an intended action ? This is a tricky issue. In fact, before an intention is formed, instrumental reasoning is usually applied even only to test whether a given desire is achievable. Therefore, either the notion of desire is stretched to cover that of goal and sub-goal, or the notion of intention is extended to embrace sub-goals, including those which will be dropped later; which is the case in the model in question?

5. It is not clear how the model in question can express the generation of new goals as means for realizing existing desires: since intention-accessible worlds are a sub-set of desire-accessible worlds, the formation of intentions looks as a mere pruning of existing desires; how is it possible in such a context to account for the generation of new intentions as means for achieving one's desires? Suppose I want to impress my boy-friend showing off on Saturday night in a new silk dress. For this reason, I decide to ingratiate myself with my mother, taking her to a Malaysian restaurant, in order to be able to borrow, later, her beautiful silk dress. In what sense can we say that the intention which I finally execute is a subset of my desire?

To sum up, at least four problems are left open by the model in question:

1. how are realistic desires selected?
2. how are realistic desires selected for action?
3. how are sub-goals generated out of existing desires?
4. how are sub-goals generated out of obligations and other external inputs?

2.2 Goals as chosen desires

One of the most influential theories of intentions, at least in the area of Multi- Agent Systems was developed by Cohen and Levesque (from now on, C&L) ([5]), on the track of Bratman's analysis ([1, 2]).

This theory is aimed at modelling the ”rational” properties of action. Intelligent, autonomous, rational agents are designed so as to be capable of producing and dropping intentions under given conditions. But which conditions are relevant for intentions formation and discharge? The authors developed an incremental view of intentions such that, at any step in a goal-driven process leading to intentions, agents are bound to decide, on the grounds of some relevant criterion, whether to keep to or abandon their goals. The language appears as a first-order language with operators for mental attitudes and action. They introduced two modalities for beliefs and goals,

(BEL x p) and (GOAL x p)

defined according to the possible worlds semantics, and therefore through accessibility relations. They implemented two modalities for action

(HAPPENS e) and (DONE a)

expressing, respectively, events taking place in the world independent of the agents' actions and occurrence of actions. Finally, time is represented as an infinite sequence of events.

In such a model, a goal is defined as a belief-compatible desire. (In other words, agents cannot have goals which they believe to be unachievable.) Many notions can be constructed on the grounds of these primitive modalities plus the operators $\Box$ for ”always” and $\Diamond$ for ”eventually”. Among others, the notion of achievement goal,

$$\text{(A-GOAL x p)} \equiv \text{(BEL x } \neg\text{p)} \wedge \text{(GOAL x (LATER p))}$$

where

$$(\text{LATER p}) \equiv \neg p \wedge \Diamond p$$

that is, x has an achievement goal p if x believes that p is not true now but wants x to become eventually true.

Indeed, in this model, an achievement goal is not yet an intention. The process of transforming a goal into an intention has been partially modelled. For example, Cohen and Levesque's theory of persistent goals is an account of a relevant aspect of this process; persistent goals are those that the agents believe to be neither realised nor unachievable:

$$\begin{aligned}(\text{P-GOAL x p}) \equiv\ & \text{A-GOAL(x p)} \wedge \\ & [\text{BEFORE((BEL x p)} \vee \text{(BEL x } \neg\Diamond\text{p))} \\ & \quad \neg\text{(GOAL x (LATER p))}]\end{aligned}$$

in words, x has a persistent goal p if before giving up trying to achieve it, x believes that p is true or will never be true.

However, even though goals are defined by the authors as both chosen desires and primitives, the process of their choice has not been addressed so far by them.

Finally, Cohen and Levesque have proposed the notion of relativised goal where a goal can be dropped when the ”circumstances” relative to which the goal has been formed have changed:

$$\begin{aligned}(\text{P-R-GOAL x p q}) \equiv\ & \text{A-GOAL(x p)} \wedge \\ & [\text{BEFORE((BEL x } \neg\text{q)} \vee \text{(BEL x p)} \vee \text{(BEL x } \neg\Diamond\text{p))} \\ & \quad \neg\text{(GOAL x (LATER p))}]\end{aligned}$$

x has a goal p relativised to q, when x has an achievement goal p, and before ceasing to have p as an achievement goal, x believes either that p is realised or unachievable or that the reason q (for the goal p) does not hold. Essentially, this means that x has p as long as and because he believes that q. Finally, Cohen & Levesque indicate how intentions can be defined as special types of goals, taking action expressions as arguments. We will not describe this step here, because it is not relevant for this paper and it contains many intricate details.

Pros

The above model

1. is somewhat more explicit and richer than what is allowed by the preceding ones;
2. allows for an incremental view of intentions, seen as a special case of goals;
3. sheds some light on the issue of the formation of intentions

Cons

But there are some questions still unsolved:

1. how are desires chosen? Which are the mechanisms responsible for this choice?
2. Should goals be necessarily seen as a subset of desires? In other words, can a goal arise from a non-motivational input? We will get back to this issue in the next section.
3. Some aspects of goal-processing are overlooked, for example, the temporary interruption of goals.
4. Finally, no mechanism of goal-generation is provided: relativised goals can be seen as the outputs, the results of goal-generation, but do not indicate its building blocks.

3 The formation of goals

As can be seen from the previous section, much work has been done about goal achievement. However, not much has been said about goal formation. In this section we describe the rationale behind this process and indicate some rules that could be used for this process.

3.1 Rationale of goal formation

To describe the process of goal formation we will divide the goal formation rules into three categories.

The first category of rules only works on concrete goals. They are used to construct plans that the agent will try to execute. All agents will be endowed with these type of rules. However, they do not really change the behaviour of the agent. Furthermore, if a goal is found unachievable the agent will keep working only if it has some alternatives to reach its overall goal. However, if the overall goal is no longer achievable the agent will come to a halt.

Usually, the goals will therefore be dropped temporarily in circumstances when the agent believes they are not achievable. Somehow they should be re-activated later on, because otherwise the agent would soon run out of goals to pursue and would come to a halt.

Also the overall goals should be some type of maintenance goals. Otherwise the agent would stop as soon as the goal was reached. Usually there are some external events that change the state such that the goal is no longer achieved and the agent has to start pursuing it again. One can think of the search robots on the WWW that have to maintain all information about all Web sites. They have to start again everytime updates are made to the Web to be of real utility. (Of course, to be practical they only operate at certain intervals).

We do not consider agents that are allowed only this kind of goal formation very interesting. They are not flexible and cannot change their behaviour over time under changing circumstances. Actually the goal of the agent is nothing more then the expected result of executing the main program of the agent (one or more times).

The second category of rules is that of ”production rules”. These production rules can be endogeneous like the ones used for planning. However, we talk about production rules in the case new concrete goals are generated from the overall built-in goals that are abstract. That is, the built-in goal is pursued through the achievement of one or more concrete goals. Often the built-in goal is not really achievable but can be approximated through the concrete goals. E.g. the goal of assembling all information about living creatures on the world. This goal is translated into a number of concrete goals like an extensive search on the WWW, in libraries, etc.

Usually the built-in goals are not made explicit and the goal formation takes place when the agent is implemented. Sometimes, goals can also be added (by the user). In this case the production rules are used to make the agent reactive to its environment. E.g. a user might tell the agent what it wants and this desire is translated into a goal of the agent. Reactive databases can be seen as having this kind of goals. They trigger certain updates on the occurrence of some event. The result of the updates can be seen as the goal generated by the event.

Most agents have these kind of rules and this type of simple goal formation. Usually, however, it is hidden within the implementation and not reflected in the agent architectures or theories.

The last category of rules is that of ”instrumental” reasoning. These rules are what we consider to be the only ”real” goal formation rules. They form new goals on the basis of a (built-in) goal and some instrumental beliefs. The instrumental beliefs indicate that (and how) a certain condition contributes to achieving the original goal, which can be seen as the reason for the new goal. E.g. the built-in goal can be to obey the law, or avoid the punishment. The instrumental belief is that driving slower than the speed limit is instrumental for obeying the law. Together with the rule the new goal of driving slower than the speed limit is derived.

The connection between the two goals can be of different kinds as we will show in the next subsection. However, the new goal always contributes in some way to achieve the old goal.

From the example above it can be seen that the goals that are derived in this way do not necessarily have to be desires by themselves (I might prefer to drive very fast), which sets them apart from the goals discussed elsewhere.

Using this instrumental reasoning one can start from very abstract, unachievable built-in goals to form goals that might be achievable and at least contribute to the achievement of the built-in goals. E.g. the built-in goal is to be a good person. In order to be a good person one should obey the law. Therefore one should not speed (above the existing limits). Therefore I should not push the gaspedal too much.

Agents that contain instrumental goal formation rules are the most flexible ones. They can start with very abstract goals which will surely remain stable throughout their existence. The concrete goals that they will pursue follow from the instrumental beliefs they have. These beliefs can be changed due to the circumstances or through learning (in any way).

Before we give a formalization for the goal formation process described above we want to make a final remark about the use of the above rules. In the complete process of goal formation all rules above will be used. First the instrumental ones, then the production rules and finally the planning rules. However, besides these rules more is needed to form the final set of goals to be pursued at a certain moment. From one built-in goal many concrete goals may be generated of which many will be alternatives of each other and even may be inconsistent together. Therefore we say that the goals, that are generated with the goal formation rules, are *candidate* goals. Some selection process will choose the actual goals that are pursued by the agent. This selection is based on some type of preference ordering which might be induced by the capabilities of the agent, the costs to reach a goal, an ordering of the built-in goals and the strength of the instrumental beliefs. We will not go into this selection process in this paper but note its importance for further (practical) research.

3.2 The formal rules

We will now describe the goal formation rules in a logic roughly based on [7].

The agents that we consider will be held to have beliefs that conform to the KD45 axiomatisation. We denote the fact that agent x believes q by $BEL_x(q)$.

As in [7], goals are expressed in a conditional form. There are several reasons for this choice. The most important is that the goals that are generated usually depend on the original goal. When the original goal is dropped, the derived goals are also dropped. This looks like the relative goals as defined in [5]. However, there is one major difference. The goals do not disappear but just become unapplicable while the condition upon which they depend is false.

This feature makes it possible to drop goals temporarily and resume the achievement later on. This is not possible with the relative goals described in [5].

We say that an agent x has a candidate goal p in situation q, denoted by $C - GOAL_x(p|q)$, if p is true in all states that the agent x considers desirable when it believes q is true and p is not true in the current state.

An agent x has a goal p in situation q, denoted by $GOAL_x(p|q)$, if p is a candidate goal and x intends to achieve p.

The idea is that an agent can have many candidate goals out of which one or more goals are selected that the agent actually wants to achieve. E.g. to reach work the agent can go by bus or by car. So, from the goal to reach work he might generate two candidate goals: reach the bus and reach the car. Now the agent uses some selection mechanism to determine which of the two candidate goals will become the actual goal.

We say that p is instrumental for q, denoted by $INSTR(p, q)$, if achieving p contributes to achieving q.

This notion of instrumentality can be seen as a generalization of the idea of subgoals. Somehow agent x does not have a plan for q (and cannot directly construct one), but achieving p is a step towards achieving q.

With the above notions we can now define the general goal generation rule:

$$GOAL_x(q|r) \ \wedge \ BEL_x(INSTR(p, q)) \ \supset \ C - GOAL_x(p|GOAL_x(q|r) \wedge r)$$

I.e. if agent x has a goal q as long as r is true and it believes that p is instrumental in achieving q then agent x has a candidate goal p as long as it has the goal q and r is true. Of course we can also generate a new candidate goal from an existing one:

$$C{-}GOAL_x(q|r) \wedge BEL_x(INSTR(p, q)) \ \supset \ C{-}GOAL_x(p|C{-}GOAL_x(q|r){\wedge}r)$$

The above rules are completely endogeneous goal formation rules if the beliefs about the instrumentality are given beforehand and do not change. In that case all goals are generated from the built-in goals. The instrumentality rules can then be seen as plan formation rules. An advantage is that the rules are explicit and can be questioned.

However, the goal generation rules can also be used to react to the environment. To effect this, the agent should have some beliefs about the benefits of reacting to other agents. I.e. it should have a theory about how the generation of a goal in response to an event contributes to some overall goal of itself.

We will start with two types of conformity behaviour.

3.3 Goal generation through conformity

The first type of conformity is called *behavioural conformity*. In this case the action of another agent is seen as example for one's own behaviour. This means that whenever an action of agent Y is perceived agent X generates the goal to perform this action as well. To achieve this goal generation we need the following formulas to be true for the agent:

$$GOAL_x(be\_like(x,y)|true)$$
$$BEL_x[DONE(y,\alpha) \supset INSTR(DONE(x,\alpha), be\_like(x,y))]$$

Now, with the goal generation rule, we can derive

$$C-GOAL_x(DONE(x,\alpha)|GOAL_x(be\_like(x,y)))$$

whenever $BEL_x(DONE(y,\alpha))$

Of course we could also describe the action conformity by ascribing the following conditional goal to the agent:

$$GOAL_x(DONE(x,\alpha)|DONE(y,\alpha))$$

This is also correct but not very flexible. This goal only disappears when y does not perform any action anymore or some explicit goal retraction action takes away the goal.
In the first definition of action conformity the reason of the conformity is given by the (built-in) goal ”to be like y”. In this case we made it unconditional, but it could be defined depending on circumstances, like this goal being beneficial for x. The action conformity would then stop whenever it is no longer beneficial for x.
The action conformity will also stop if the agent no longer believes that the conformity is instrumental to the goal of being like y. I.e. we have made the belief of agent x in action conformity and its purpose explicit.

The second type of conformity is called *goal conformity*. This is a more autonomous type of conformity, where actions are not copied but the reason of the actions (the goal) is adopted. The way goals are generated for goal conformity can be described (like above), with the following two formulas:

$$GOAL_x(be\_like(x,y)|true)$$
$$BEL_x[GOAL_y(p|r) \wedge r \supset INSTR(p, be\_like(x,y))]$$

And again with the general goal generation rule we can derive:

$$C-GOAL_x(p|GOAL_x(be\_like(x,y)))$$

whenever $BEL_x(GOAL_y(p|r) \wedge r)$.

From the above we see that an agent x might keep its goal to be like y, but changes its belief about how to achieve this. In goal conformity the agent believes that adopting the goal of the other agent contributes to being like the other agent.

3.4 Generation of goals through adoption

Goals can also be generated through a process of adoption of norms or goals.
The idea of goal adoption is similar to that of goal conformation. However, the goal is that the other agent obtains its goal. We can write the formulas needed for this type of goal generation as follows:

$$GOAL_x(help(x,y)|true)$$
$$BEL_x[GOAL_y(p|r) \wedge r \supset INSTR(OBT_y(p), help(x,y))]$$

where

$$OBT_y(p) \;\equiv\; p \;\wedge\; GOAL_y(p|q)UNTILp$$

So, agent y obtained p if p is true and it had the goal p until the moment p became true. Here $qUNTILp$ is defined as follows:

$$qUNTILp \equiv \ll \alpha \gg p \rightarrow (\forall \beta : (\beta \neq \gamma; \alpha; \delta) \rightarrow [\beta]q)$$

In this definition we translate the temporal operator UNTIL into dynamic logic. The antecedent states that p becomes true only after performing some sequence of actions α. It follows from this antecedent that q should be true until α has been performed completely. That is, it should be true after each sequence of actions that does not include α. We can now derive the following candidate goal for x:

$$C - GOAL_x(OBT_y(p)|GOAL_x(help(x, y)))$$

whenever $BEL_x(GOAL_y(p|r) \wedge r)$.
Notice that the goal for x is not p itself like with goal conformity, but the goal is that agent y obtains the goal p.

The goal generation mechanisms described in this section are more general and flexible than the production rules that are often encountered in agents. For instance the goal conformation could be realized through the following axiom:

$$C - GOAL_x(p|GOAL_y(p))$$

However, in this case agent x would adopt every goal of agent y without discrimination. In our goal generation system agent x might decide to adopt some goals of y, but not all of them, because they will not all serve the goal of helping y. Also the general goal (of helping y) could be made conditional upon the circumstances of agent x. E.g. whether it has enough resources, etc.

4 A sketch of a formalisation

Space limitations prevent us from incorporating the goal formation rules into a complete agent model. The language that we present in this section, however, is based on a language to model complete agents, including actions, communication and norms (see [7]). We will not present the complete range of concepts, because they would distract from the formalisation of the concepts used in this paper. On the other hand, we admit that the concepts used in this paper could be formalised in a simpler way. However, this formalisation could then not so easily be extended to a theory for all aspects of the agents.

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 three stages. Starting with a set of propositional formulas ($PFORM$), we define the 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 **any** and **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 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 and aspects of actions. In this paper we leave out the communication aspects due to a lack of space.

Definition 2. The language $FORM$ of formulas is defined to be the smallest set closed under:

1. $PFORM \subseteq FORM$
2. $\phi, \phi_1, \phi_2 \in FORM \Longrightarrow \neg\phi, \phi_1 \wedge \phi_2, INSTR(\phi_1, \phi_2) \in FORM$
3. $\phi \in FORM, i \in Ag \Longrightarrow BEL_i(\phi), GOAL_i(\phi), C - GOAL_i(\phi) \in FORM$
4. $\alpha \in Act, \phi \in FORM \Longrightarrow [\alpha]\phi, \ll \alpha \gg \phi \in FORM$
5. $\alpha \in Act, \phi \in FORM \Longrightarrow DONE(\alpha) \in FORM$
6. $\phi, \psi \in FORM, i \in Ag, \alpha, \alpha_1, \alpha_2 \in Act \Longrightarrow INT_i\alpha, O_i(\alpha) \in FORM$

The last rule is only included to show that intentions and norms are also important concepts for agents and related to the goals.
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 component, the action component, the motivational component and the social component, respectively. We assume tt and ff to denote the truth values ‘true’ and ‘false’, respectively.

Definition 3. A model Mo for $FORM$ from the set CMo is a structure $(\Sigma, \pi, Rb, A, M, Dw)$ where

1. Σ is a non-empty set of states and $\pi : \Sigma \times \Pi \rightarrow \{tt, ff\}$.
2. $Rb : Ag \times \Sigma \rightarrow \wp(\Sigma)$ denotes the doxastic alternatives.
3. $A = (Sf, Rprev)$ with $Sf : Ag \times Act \times \Sigma \rightarrow \wp(\Sigma)$ yielding the interpretation of actions and $Rprev : Ag \times \Sigma \rightarrow Act$ yielding the action that has been performed la st.
4. $M = (Ri, Ro, G, CG)$ with $Ri : Ag \times \Sigma \rightarrow \wp(Act)$ denoting intended actions, $Ro : Ag \rightarrow \wp(\Sigma \times \Sigma)$ denoting obligations, $CG : Ag \times \Sigma \rightarrow \wp(\Sigma)$ denoting candidate goals and $G : Ag \times \Sigma \rightarrow \wp(\Sigma)$ denoting goals.
5. $Dw : \Sigma \times \Sigma \rightarrow Integers$ yields the ”distance” between two worlds. We do not define this any further, but one can think of how many propositions have a different truth value in both worlds together with the difference in beliefs of the agents, etc.

such that the following constraints are validated:

1. $Rb(i, s) \neq \emptyset$, which ensures that belief obeys a KD45 axiomatisation.

2. Sf yields the global state-transition interpretation for regular actions. This function satisfies the usual constraints ensuring an adequate interpretation of composite actions in terms of their constituents.
3. $Ri(i,s) \subseteq \{\alpha \mid Sf(i,\alpha,s) \neq \emptyset\}$ which means that only actions that are possible can be intended.
4. For all $s \in \Sigma$ some $s' \in \Sigma$ exists with $(s,s') \in Ro$, which ensures that all obligations are also permitted (the D-axiom holds).
5. $G(i,s) \subseteq CG(i,s)$

The complete semantics contains an algebraic semantics of action expresses, based on the action semantics of Meyer [13]. In this paper we will abstract from the algebraic interpretation of actions and instead interpret actions as functions on states of affairs.

Definition 4. 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. Doxastic formulas $BEL_i\phi$ are interpreted as a necessity operator over Rb respectively. For the other formulas $\models$ is defined as follows:

$$
\begin{array}{ll}
Mo,s \models C-GOAL_i(\phi|\psi) & \Longleftrightarrow \text{If } Mo,s \models BEL_i(\psi) \\
 & \quad\ \ \text{then } Mo,s' \models \phi \text{ for all } s' \in CG(i,s) \\
Mo,s \models GOAL_i(\phi|\psi) & \Longleftrightarrow \text{If } Mo,s \models BEL_i(\psi) \\
 & \quad\ \ \text{then } Mo,s' \models \phi \text{ for all } s' \in G(i,s) \\
Mo,s \models [\alpha(i)]\phi & \Longleftrightarrow Mo,s' \models \phi \text{ for all } s' \in Sf(i,\alpha,s) \\
Mo,s \models \ll \alpha(i) \gg \phi & \Longleftrightarrow Mo,s' \models \phi \text{ for all } s' \in Sf(i,\alpha,s) \\
 & \quad\ \ \text{and } Mo,s'' \not\models \phi \text{ for all } s'' \in Sf(i,\beta,s) \\
 & \quad\ \ \text{and } \forall\beta : \alpha = \beta;\gamma \\
Mo,s \models DONE(\alpha(i)) & \Longleftrightarrow \alpha \in Rprev(i,s) \\
Mo,s \models INSTR(\phi,\psi) & \Longleftrightarrow \text{for all } s', s'' : \pi(s',\phi) = tt \wedge \pi(s'',\psi) = tt \\
 & \quad\ \ \text{then } Dw(s',s'') \leq Dw(s,s'') \\
Mo,s \models INT_i\alpha & \Longleftrightarrow \alpha \in Ri(i,s) \\
Mo,s \models O_i(\phi) & \Longleftrightarrow Mo,s' \models \phi \text{ for all } s' \text{ with } (s,s') \in Ro(i) \\
Mo,s \models O_i(\alpha) & \Longleftrightarrow Mo,s \models [\mathbf{any}(i)]O_i(PREV(\alpha(i)))
\end{array}
$$

5 Conclusions

In this paper some models of intentional action have been discussed. The main drawback of these models is their tendency to see goals only as a subset of endogenous motivations. This leads to building goals into the agent. This goals are therefore fixed and the agent cannot change its goal in response to events in the environment.

We have provided a general mechanism for goal generation. On the basis of some very general, abstract, built-in goals other goals can be generated thanks to some beliefs on instrumentality. This step of goal generation is a first step towards the forming of intentions. In this paper we have not shown the complete process of intention formation. Some intuitions about the rest of the process can be found in [3].

We have sketched a formalisation of a goal generation mechanism. Some examples of goal generation (through goal conformation, etc.) were given. Space constraints did not allow many details to be fully examined. Also other interesting goal generation rules, like goal generation through norm adoption could for this reason not be included. In a further study the fomalisation will be extended to all stages of intention formation. Finally, an implementation of the mechanism described is planned for future work.

References

1. Bratman, M.E. Intentions, Plans, and Practical Reason. Harvard University Press, 1987.
2. Bratman, M.E. What is intention? In P.R Cohen, J. Morgan, M.A. Pollack (eds), *Intentions in Communication*, 401-15. Cambridge, MA: MIT Press, 1990.
3. Castelfranchi, C. Commitments: From individual intentions to groups and organizations. *Proc. of the 1st International Conference on Multi-Agent Systems, ICMAS-95*, San Francisco, CA. Menlo Park, CA: AAAI Press/ The MIT Press, 1995.
4. Castelfranchi, C. Reasons: Belief support and goal-dynamics. *Journal of Mathware & Soft Computing*, 3(1-2), 233-247, 1996.
5. Cohen, P.R. and H.J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(3), 213-261, 1990.
6. Conte, R. and Castelfranchi, C. Cognitive and social action, London, UCL Press, 1995.
7. F. Dignum. Social interactions of autonomous agents; private and global views on communication. In A. Cesta and P-Y. Schobbens, editors, *Proceedings of the 4th ModelAge workshop on formal models of agents*, pages 99–114, Siena, Italy, 1997.
8. F. Dignum, E. Verharen and S. Bos. Implementation of a Cooperative Agent Architecture based on the Language-Action Perspective. In *this volume*.
9. Georgeff, M.P. and Rao, A.S. The semantics of Intention Maintenance for Rational Agents. *Proceedings of the International Joint Conference of Artificial Intelligence*, 1995.
10. M. d'Inverno, D. Kinny, M. Luck and M. Wooldridge. A Formal Specification of dMARS. In *this volume*.
11. C. Jung and K. Fischer. A Layered Agent Calculus with Concurrent, Continuous Processes. In *this volume*.
12. D. Kinny and M. Georgeff. Commitment and Effectiveness of Situated Agents. In *Proceedings International Joint Conference on Artificial Intel ligence*, Sydney, Australia, pages 82-88.
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. Rao, A.S. and M.P. Georgeff Modelling rational agents within a BDI architecture. In J. Allen, R. Fikes, E. Sandewall (eds), *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, 473-485. San Mateo, Kaufmann, 1991.
15. Rao, A.S. and Georgeff, M.P. A model-theoretic approach to the verification of situated reasoning systems. *Proceedings of the 13th International Conference of Artificial Intelligence, IJCAI-93*, Chambery, France, 1993.
16. Wooldridge, M.J. and Jennings, N.R. Agent theories, architectures, and languages: A survey, 1994.