Modeling the mental activity of an autonomous agent: an implementation based on intentions and persuasions

Pietro Baroni, Daniela Fogli, Giovanni Guida
Università di Brescia - DEA
Via Branze 38, I - 25123 Brescia, Italy
e-mail: {baroni, guida, fogli}@ing.unibs.it

Silvano Mussi
CILEA
Via Sanzio 4, I - 20090 Segrate (MI), Italy
e-mail: mussi@icil64.cilea.it

Abstract

In this paper, we propose a novel approach to the design of intelligent autonomous agents, based on an original model of mental activity. In particular, we introduce the concept of active mental entity, as a new way of representing mental attitudes such as intentions and persuasions. The internal architecture of an agent is thus understood as a distributed system whose reasoning activity is determined by the interactions among active mental entities. A detailed description of the structure and operation of an agent is provided. The implementation of the proposed paradigm is then illustrated and some performance examples are presented.

1 Introduction

The capability of autonomous behavior in a complex and unpredictable environment is a crucial property for intelligent agents, as it has been often stressed in the literature (see for example [Maes 95][Wooldridge and Jennings 95]). One of the major issues in order to actually achieve a fully autonomous behavior is the design of a proper agent architecture. The main approaches to agent architecture design include the reactive approach and the deliberative approach. *Reactive* agents [Agre and Chapman 87][Brooks 91][Maes 95] are built according to a paradigm called *behavior-based*, since their architecture is understood as a hierarchy of task-accomplishing behaviors whose selection is determined by a fixed set of *stimulus-response* rules. Agent operation is relatively simple and guarantees an high level of reactivity; however, it is completely driven by the external stimuli rather than by its (implicit) internal goals. In a sense, the agent is at the mercy of the stimuli: its behavior depends entirely on them. On the other side, the *deliberative* approach [Georgeff and Lansky 87][Bratman et al. 88][Pollack 92] is based on the assumption that the explicit representation of mental activity should provide agents with high-level reasoning capabilities. In particular, agents are assumed to possess a set of *goals* and to produce and execute plans in order to fulfil such goals. External stimuli give rise to *beliefs* and agent operation includes specific mechanisms devoted to revise plans according to beliefs. In this approach, the behavior of an agent is fully regulated by its mental activity.

While several theoretical works provide a sound logical foundation for modeling agent mental activity (see for example [Cohen and Levesque 90] [Rao and Georgeff 91]), there is still a significant gap, however, between theoretical proposals and actual implementations of agents endowed with mental attitudes, as explicitly pointed out for example in [Wooldridge and Jennings 95]. In this paper, starting from the basic ideas presented in [Cohen and Levesque 90], we propose an original approach to modeling agent mental activity which provides a sound basis for a practical implementation.

2 Active mental entities

Cohen and Levesque's theory [Cohen and Levesque 90] is based on the notions of belief and goal. However, some pragmatic questions, concerning both the representation and the manipulation of such mental attitudes, arise when an actual implementation has to be realized:

- How should beliefs and goals be represented?
- When, how and why do beliefs arise?
- Which is the mechanism ruling belief revision?
- When, how and why do goals arise?
- How is commitment to goals realized?
- Which is the mechanism ruling the selection and revision of action plans for pursuing adopted goals?

A typical answer to these questions is provided by a family of agent architectures [Georgeff and Lansky 87][Bratman et al. 88][Pollack 92][Jennings 95] implementing the widely adopted BDI theory. Within such architectures, mental attitudes are conceived as data structures which are manipulated by a specific centralized process that constitutes the kernel of the agent. The design of such centralized mechanism is however very complicated and not transparent enough to guarantee a clear understanding of the issues listed above.

In order to overcome these difficulties, we start from an original point of view concerning the model of the mental activity of an agent. Our main ideas can be summarized as follows:

(i) the mental processes occurring inside an agent are the result of the cooperation and conflict between a set mental attitudes;
(ii) since the interactions between attitudes are expected to produce a globally intelligent behavior, it is essential that mental attitudes are provided with individual and independent operation capabilities;
(iii) as a consequence of (i) and (ii), we represent mental attitudes as *active mental entities*, i.e. entities that can freely interact in a distributed context according to their own nature;
(iv) active mental entities give existence to dynamic processes that can be created and disposed when necessary.

In this paper we focus on two classes of active mental entities, namely *persuasions* and *intentions*, which are enough to show the potential of our approach.

2.1 Persuasion

In our model, the concept of persuasion is related to that of belief. A belief is, in general, conceived as a data structure resulting from a perceptual or reasoning activity. We say instead that an agent has a persuasion when it is interested in knowing the truth value of a given proposition, i.e. it is interested in forming a belief about a given proposition that represents a specific fact whose truth or falsity affects agent operation. Informally, we may state that a *persuasion* is an active mental entity definitely committed to find and verify elements that can support the association of a truth value to a given proposition of interest. More precisely, we can define a *persuasion* as a four-tuple P = <S, V, J, M>, where:

- S is the *subject* of the persuasion, namely a proposition whose truth value is of interest for the agent;
- V is the *truth value* of S, namely the currently most believed truth value of the subject;
- J is the list of *justifications* of the persuasion, namely a list of information items from which the association of the truth value V to the subject S can be directly derived through a single inference step. According to the nature of the information it is based upon, a justification has an associated *justification type* which may assume one of the following three values:
  - *knowledge*, i.e. knowledge available inside the agent;
  - *sense data*, i.e. data collected from the external environment;
  - *related proposition*, i.e. another proposition from which V can be derived, but which, differently from knowledge and sense data, needs in turn to be justified.
- M is a set of *methods* used by the persuasion in its operation.

The set of methods M is in turn a five-tuple M = <GM, SM, EM, RM, DM> where:

- GM is a *generation method*, i.e. a mechanism for the generation of plans capable of finding justifications supporting the association of a truth value to the subject S of a persuasion; a *plan* is understood here as a sequence of *tasks* to be performed to achieve a stated goal, namely the construction of a justification;
- SM is a *selection method*, i.e. a mechanism for the selection of a plan;
- EM is an *execution method*, i.e. a mechanism for the execution of the selected plan;
- RM is a *revision method*, i.e. a mechanism for the revision of the currently believed truth value assigned to the subject of the persuasion when external conditions affecting it change;
- DM is a *debate method*, i.e. a mechanism for carrying out a debate with conflicting persuasions; a *debate* is understood here as sequence of message exchanges according to a suitable debate protocol.

A persuasion is active for the period of time during which its subject is considered interesting for an agent: by exploiting its methods, it is initially committed to determine the truth value of the subject, then to monitor whether any relevant change occurs that could affect the currently believed truth value.

2.2 Intention

According to [Cohen and Levesque 90], an intention can be understood “as a composite concept specifying what the agent has chosen and how the agent is committed to that choice”. This implies that an intention can be modeled as a kind of persistent goal. We make a step further in this direction by stating, informally, that an *intention* is an active entity definitely committed to pursue a persistent goal.

More precisely, we can define *intention* as a three-tuple I = <S, C, M>, where:

- S is the *subject* of the intention, namely a proposition about a state of the world representing the achievement, i.e. the goal, of the intention;
- C is the *validity condition* of the intention, namely a proposition enabling the intention to be active;
- M is a set of *methods* used by the intention in its operation.

The set of methods M is in turn a five-tuple M = <GM, SM, EM, RM, CM> where:

- GM is a *generation method*, i.e. a mechanism for the generation of plans capable of achieving the subject of the intention; a *plan* is understood here as a sequence of *tasks* to be performed to achieve a stated goal, namely the subject of an intention; a plan has an associated *applicability condition* representing the condition under which the plan can be performed;

- SM is a *selection method*, i.e. a mechanism for the selection of one of the generated plans;
- EM is an *execution method*, i.e. a mechanism for the execution of the selected plan;
- RM is a *revision method*, i.e. a mechanism for the revision of the currently active plan when external conditions change, namely when the applicability condition of the currently executed plan is no more believed to be true;
- CM is a *conflict resolution method*, i.e. a mechanism for solving the possible conflicts between different intentions.

Note that, since goals can be related either to permanent internal needs of the agent or to transient needs, related to the actual satisfaction of the permanent needs in a specific context, we can distinguish two kinds of intentions:

- *Primitive intentions* are created at the same time as the agent and are always active. A primitive intention is kept forever by the agent and does not depend on a specific achievement (its validity condition is always true). Primitive intentions represent therefore very general and fundamental objectives which are intrinsic to the existence of an agent and, in a sense, represent its basic *raison d'être*: "Preserving integrity" is an example of primitive intention.
- *Generated intentions* correspond to transient goals and are created by other mental entities. A generated intention is produced when the achievement of its subject is necessary for the achievement of the subject of another intention (either primitive or generated) or for determining the truth value of a persuasion. Intentions of this kind remain active only until their subject is achieved or their validity condition no more holds.

2.3 Relations between intentions and persuasions

Intentions and persuasions are strictly related to each other: in fact, in order to execute its plans an active entity (either an intention or a persuasion) may need to generate some new active entities, devoted to solve sub-problems on behalf of their generator. According to this perspective, we define here the relationships concerning the dynamic generations of mental entities in our distributed mental activity model.

Generation of persuasions by intentions

Each intention is related to two propositions, namely its subject and its validity condition, whose truth values affect its operation. Therefore, when an intention is generated, such propositions become in turn interesting propositions and, as a consequence, two persuasions are generated, having such propositions as subjects.

Persuasions are also involved in the phases of plan selection and execution by an intention. In fact, they are affected by the applicability conditions of the generated plans. As a consequence, a new persuasion is generated for each applicability condition, in order to allow both the initial plan selection and, possibly, subsequent plan revisions, if applicability conditions change.

Generation of intentions by intentions

Once an intention has selected a plan, it has to put it at work, by executing the various tasks composing the plan. Each task can represent either an elementary action (a computation, a sensorial acquisition, an action on the environment) which can be executed by a suitable operative component, or a non elementary action, whose accomplishment corresponds to generating a new intention whose subject represents the accomplishment of the task.

Generation of persuasions by persuasions

The justification for the truth value of the subject of a persuasion can be obtained through an inference step which uses as premise another proposition. In this case, the related proposition becomes the subject of a new persuasion.

Generation of intentions by persuasions

Each persuasion is in charge of executing plans whose result is the association of a truth value to its subject. Such plans may involve the generation of new intentions, if they encompass the execution of non elementary tasks (for instance concerning acquisition and interpretation of sense data).

The following operational constraints hold between a mental entity (intention or persuasion) and the mental entity that has generated it:

- a persuasion is in charge of notifying the mental entity that has generated it when there is a change in its believed truth value;
- an intention is in charge of notifying the mental entity that has generated it when its subject is achieved or when a definitive failure in its accomplishment is detected.

Finally, every mental entity that has generated another mental entity may decide to suppress it at any moment if the activity of the generated entity is no more considered useful.

3 Overall agent structure

The overall structure of an agent A includes two parts:

- a *static part*, which is created at the moment an agent is defined and which remains unchanged during all its operational life;
- a *dynamic part*, which includes components that are

generated and disposed during agent operation.

The static part *S* is a three-tuple S = <Ip, O, K> where:

- Ip is the set of *primitive intentions* of A;
- O is the set of *operative components* of A, in charge of performing elementary actions, either mechanical, concerning the interaction with external world through sensors and actuators, or symbolic such as inferential and computational activity;
- K is the set of *knowledge bases* available to A for problem-solving purposes; it represents the basic agent competence endowment and can be exploited by agent components during their operation.

The dynamic part *D(t)* is a time-variant pair D(t) = <P(t), Ig(t)> where:

- P(t) is the set of persuasions which are active in A at time t;
- Ig(t) is the set of generated intentions which are active in A at time t.

4 Solving conflicts between mental entities

Possibly the most significant part of the cognitive activity of an agent is constituted by conflict resolution between mental entities. In fact, among the most important capabilities an autonomous agent should be endowed with there are:

- the capability to deal with different contrasting goals, which can be obtained by exploiting suitable methods for conflict resolution between intentions;
- the capability to cope with the uncertainty and ambiguity that affects the perception and representation of the external world, which can be obtained by exploiting suitable methods for conflict resolution between persuasions.

Due to space limitation we can give here only a brief account about this important topic (see [Baroni et al. 97] for a more detailed description).

4.1 Conflict resolution between persuasions

The findings of different persuasions, which are based on different justifications and may have been acquired at different times, may lead to contradictory conclusions, so generating a conflict, which needs to be solved if such conclusions affect agent decisions.

Here we consider only a simple conflict resolution mechanism for persuasions, based on the comparison between the justification types of plans adopted by persuasions. In practice, we assume that a pre-defined priority order exists between the justification types "knowledge" and "sense data". In particular, the latter, being based on recent and up-to-date data acquisitions, is assumed to be stronger than the first one (of course this is a working assumption which might not be adequate in some specific situations, e. g. where sensory devices are particularly unreliable). Regarding the "related proposition" type, it has to be considered that this kind of justification in general brings about a chain of related propositions. At the root of this chain, a terminal node represents a justification which can be in turn "knowledge" or "sense data". Thus, each persuasion which has adopted a plan whose justification type is "related proposition" must search first for the justification type ascribed to the terminal node of the chain of propositions supporting its subject and then determine its justification strength (for the sake of simplicity we do not consider here the case where multiple justification chains are available).

Therefore, conflict resolution is carried out by the involved persuasions by determining which one has the stronger justification. If both persuasions have the same justification type, a more articulated conflict resolution mechanism is necessary: its description is however beyond the limits of this paper.

4.2 Conflict resolution between intentions

A conflict between two intentions may arise when they concurrently try to assign contrasting tasks to a shared operative component. If the involved intentions are primitive, we assume that a *priority* attribute makes it possible to directly establish the prevailing intention. If one conflicting intention is primitive and the opponent one is generated, the latter refers to the priority of the primitive intention underlying it. Finally, if both intentions are generated, conflict resolution involves a more articulated interaction protocol. For the sake of simplicity, we will consider here only a simple example of protocol, where the conflict is solved by postponing the execution of the plan of a conflicting intention to that one of the opponent. In order to decide which plan should be postponed, each intention simulates the execution of a new compound plan, which is obtained from the previous one by hypothesizing to give priority to the opponent intention. If the conflict cannot still be solved, i.e. both intentions are unable to accept to be postponed, the conflict is transferred at the level of primitive intentions, where it can be directly solved according to their priority attribute.

5 Implementation and experimentation

On the basis of the general approach to agent design introduced in the previous sections, a prototype programming environment for the development of agent architecture endowed with active mental entities has been developed. The implementation of such prototype has been developed in C++. In particular, the *Cooroutine Library* [Stroustrup and Shopiro 87] has been exploited, that contains basic facilities for multi-thread programming. In this environment, a thread can be implemented as an instance of a user-defined class derived from the basic class *task* and can be suspended and resumed when necessary. All components of an agent are thus implemented as classes derived from class *task* and an agent is realized as a multi-thread system, where each component is associated to a thread executed concurrently with respect to other ones. In particular, operative components and primitive intentions, which belong to the static part of an agent, correspond to threads starting their operation at the moment of the creation of an agent, while persuasions and generated intentions, which belong to the dynamic part of an agent, correspond to threads which are dynamically created by other threads. Finally, generated threads may terminate spontaneously or may be killed by their generator, according to the circumstances. Communication among different agent components is carried out through a message passing paradigm by exploiting specific primitives available in the library.

The prototype has been experimented in the implementation, at a high level of abstraction, of the simulated control architecture of a department mail delivery robot. An example is presented below in order to give an idea of how the prototype works.

5.1 Solving conflict between persuasions

Let us suppose that the primitive intentions "obey-the-user" and "preserve-energy-level" are primitive, and therefore always active inside the robot. After receiving the request of delivering an envelope to Mr. X, a new intention having subject "deliver-mail-to-Mr.X" is generated by "obey-the-user". This intention may generate different plans for its achievement (actually in our implementation we resort to the commonly adopted solution [Rao and Georgeff 91] of exploiting a set of pre-compiled plans).

For instance a simple plan whose applicability condition is "Mr. X is in his office" is:

task 1: go to Mr. X office
task 2: deliver the envelope to Mr. X.

An alternative plan whose applicability condition is "Mr. X seen in front of robot" is:

task 1: go near Mr. X
task 2: deliver the envelope to Mr. X.

In order to enable plan selection, two persuasions, having as subject the two applicability conditions, are created and are then in charge of finding justifications. For instance "Mr. X is in his office" may be justified by default knowledge that an employee is normally in his office, whereas "Mr. X seen in front of robot" may generate the intention "recognize-face-of-Mr.X" for acquiring and processing data coming from the video camera.

Let us suppose that the robot realizes that Mr. X is in front of it. Then, a conflict between the persuasions "Mr. X is his office" and "Mr. X seen in front of robot" must be solved. Since the first one has justification type *knowledge*, whilst the latter is justified by *sense data*, the conflict is solved in favour of the second persuasion. Accordingly, the plan relying on such persuasion is put at work. The first task of the plan, namely "go-near-Mr.X", is not elementary and therefore it becomes the subject of a new generated intention which drives robot towards Mr. X by resorting to the operative components controlling wheels.

5.2 Solving conflict between intentions

Let us suppose now that, while the intention "go-near-Mr.X" is active, the robot energy reaches the minimum threshold. This fact is noticed by a persuasion which is constantly monitoring the energy level and whose subject is "battery-is-drying-up". This subject represents the applicability condition of a plan of the primitive intention "preserve-energy-level": the persuasion notifies the intention of the change occurred in its truth value and, as a consequence, the primitive intention undertakes the execution of the following plan:

task 1: go to the recharging point
task 2: wait for the complete battery recharge.

Task 1 involves the generation of a new intention which tries to direct the robot towards the recharging point. Thus, both intentions "go-near-Mr.X" and "go-to-the-recharging-point" attempt to control the robot movement system and enter in conflict one another. In order to solve the conflict, intentions operate as follows:

- "go-near-Mr.X" estimates the time for the execution of a (simulated) plan where reaching Mr.X is postponed to the completion of the plan to which the opponent intention belongs;
- "go-to-the-recharging-point" estimates the time for the execution of a (simulated) plan which delays the arrival at the recharging point after the completion of mail delivery.

Then, since generated intentions have an associate *deadline* attribute, they compare the obtained time estimations with such deadlines in order to decide an acceptable scheduling.

If both times of simulated plans are incompatible with the intentions deadlines, conflict resolution is delegated to primitive intentions "obey-the-user" and "preserve-energy-level" which solve the conflict according to their priorities.

6 Discussion

In this section, we outline how our approach based on the notion of active mental entities is definitely appropriate for the actual realization of autonomous agents.

A deep investigation about the concept of autonomy is carried out by Castelfranchi in [Castelfranchi 95]. In particular, Castelfranchi deals with two different forms of autonomy, namely *cognitive autonomy* and *social autonomy*. In order to achieve cognitive autonomy, beliefs and goals are introduced in place of stimuli and reactions. On the other side, social autonomy concerns the relationship between the goals of different agents. In order to obtain this kind of autonomy "the system is endowed with goals of its own, which it has not received from outside as contingent commands. And its decisions to adoption of others' goals are taken on the basis of these goals" [Castelfranchi 95]. Moreover, Castelfranchi asserts that a key factor for cognitive autonomy is that "it is impossible to change automatically the beliefs of an agent", since it depends on the mechanism of belief updating rather than just on the fact that the agent has some internal beliefs. A similar consideration can be made for goals, since an agent should not be blindly available to adopt exogenous goals. Therefore, having an explicit representation of beliefs and goals is not sufficient to guarantee that an agent is cognitively and socially autonomous.

In order to achieve cognitive autonomy an agent should be also able to:

a) select the information it is interested in, rather than collecting any stimulus provided by the external environment;
b) search for interesting information when it is not immediately available;
c) recognize when an information is no more interesting;
d) solve conflicts between contradicting information;
e) find appropriate plans to pursue its goals by maintaining a commitment both to plans and to goals.

These capabilities are naturally encompassed by the approach based on active mental entities. In fact, active persuasions are generated by other mental entities and this guarantees that the agent focuses its attention only on the aspects of the world which are of some interest for it. Moreover, since persuasions are active entities, they do not simply acquire the readily available information from the environment, but may start also information acquisition activities through the generation of suitable intentions. Then, since persuasions are dynamic entities, they remains active only until the belief to which they refer is no more considered interesting.

The explicit representation of contradicting points of view is a key feature for acting in realistic contexts, and the solution of conflict which eventually arises from these contradictions may be carried out directly by the mechanism of interaction between persuasions, as explained in subsection 4.1.

Finally, active intentions allow the realization of the two identified forms of commitment with respect to goals and to plans. They are persistent and are capable of managing external events which could influence such commitments: the intention operation mechanism includes the capability of revising plans, when they are no more applicable, and of dropping goals, when they are no more achievable.

Turning to social autonomy, an intelligent autonomous agent should be able to:

a) distinguish between endogenous and exogenous goals;
b) mediate external requests with its pre-existing goals, by eventually solving conflicts between contrasting needs.

Both these properties are guaranteed by our notion of intention as active mental entity. The concept of primitive intentions guarantees that any agent is endowed with an endogenous and permanent set of intentions, which represent the objectives that the agent is permanently committed to achieve.

Externally generated intentions (i.e. those ones deriving from the requests of other agents) can not overwrite pre-existing intentions, but rather have to cohabit with them. If an exogenous intention is not compatible with another pre-existent intention, a conflict arises and is solved through the internal conflict resolution mechanism.

7 Conclusions

In this paper, we have presented a new approach for designing and implementing autonomous agents. We have remarked, as an important feature of the deliberative approach, the separation occurring between goals and actions and between beliefs and stimuli. Then, we have proposed, as a starting point of our modeling perspective, a new kind of separation: namely the separation between goals and intentions and between beliefs and persuasions. In particular, intentions and persuasions are conceived as active mental entities, i.e. they are able to carry out autonomous activities.

Starting from this standpoint, we have proposed a formalization of the structure and operation of persuasions and intentions and of the overall agent structure and architecture, conceived as a dynamic and distributed system. A prototypical software implementation has been developed and some examples concerning the control architecture of a simulated mobile robot have been presented.

References

[Agre and Chapman 87] P. E. Agre, D. Chapman, Pengi: An Implementation of a Theory of Activity. *Proceeding of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, Seattle, WA, 1987, 268-272.

[Baroni et al. 97] P. Baroni, D. Fogli, G. Guida, S. Mussi, Active mental entities: a new approach to endow multi-agent systems with intelligent behavior. *Proc. of the Workshop on Multi-Agent Systems: Theory and Applications (MASTA'97)*, Coimbra, October 1997, 1-15.

[Bratman et al. 88] M. E. Bratman, D. J. Israel, M. E. Pollack, Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4, 1988, 349-355.

[Brooks 91] R. A. Brooks, Intelligence without representation. *Artificial Intelligence*, 47, 1991, 139-159.

[Castelfranchi 95] C. Castelfranchi, Guarantees for Autonomy in Cognitive Agent Architecture, M. J. Wooldridge, N. R. Jennings (eds.), *Intelligent Agents*, LNAI-890, Springer-Verlag, Berlin, 1995, 56-70.

[Cohen and Levesque 90] P. R. Cohen, H. J. Levesque, Intention is choice with commitment. *Artificial Intelligence*,42(3),1990,213-261.

[Georgeff and Lansky 87] M. P. Georgeff, A. L. Lansky: Reactive Reasoning and Planning. *Proceeding of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, Seattle, WA, 1987, 268-272.

[Jennings 95] N. R. Jennings, Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. *Artificial Intelligence*, 75(2), 1995, 195-240.

[Maes 95] P. Maes, Artificial Life Meets Entertainment: Lifelike Autonomous Agents, *Communications of the ACM*, 38(11), 1995, 108-114.

[Pollack 92] M. E. Pollack, The uses of plans, *Artificial Intelligence*, 57 (1), 1992, 43-68.

[Rao and Georgeff 91] A. S. Rao, M. P. Georgeff, Modeling Rational Agents within a BDI-Architecture, *Proc. of KR&R-91 Int. Conf. on Knowledge Representation and Reasoning*. Cambridge, MA, 1991, 473-484.

[Stroustrup and Shopiro 87] B. Stroustrup, J. Shopiro, A set of C++ classes for Co-routine Style Programming, *Proc. of USENIX C++*, Santa Fe, November 1987, 417-439.

[Wooldridge and Jennings 95] M. Wooldridge, N. R. Jennings, Intelligent Agents: Theory and Practice. *Knowledge Engineering Review*, vol.10(2),1995,115-152.