Active Mental Entities: A step towards achieving Agent Autonomy

Pietro Baroni, Daniela Fogli, Giovanni Guida

Università di Brescia, Dipartimento di Elettronica per l'Automazione,
Via Branze 38, 25123 Brescia, Italy, e-mail: {baroni, guida, fogli}@bsing.ing.unibs.it

Abstract

The property of autonomy is a crucial requirement for intelligent agents. Several definitions of autonomy, with significantly different meanings, are available in the literature. In particular a detailed study about this concept, featuring the distinction between cognitive and social autonomy, has been carried out in [Castelfranchi 95]. With reference to these works, we discuss in this paper how the satisfaction of autonomy requirements can be achieved by a novel approach to agent modeling, based on the concept of active mental entity. After introducing the main features of the approach, we show how it is able to guarantee both social and cognitive autonomy. An application example concerning action planning of an autonomous mobile robot provides a concrete illustration of the proposed approach.

1. Introduction

The design of intelligent autonomous agents is one of the most important challenge in Artificial Intelligence. In particular, *autonomy* is a crucial requirement for intelligent agents. Maes [Maes 95] asserts that "autonomous agents are computational systems that inhabit some complex, dynamic environment, sense and act *autonomously* in this environment, and by doing so realize a set of goals or tasks for which they are designed". The notion of autonomy considered here is related to the concept of situatedness: agents are "situated" in a dynamic environment and must react appropriately to changes occurring in it. The research trend following this view aims at realizing intelligent autonomous agents by simple task-oriented modules implemented through stimulus-response rules [Brooks 91] [Maes 95].

In [Wooldridge & Jennings 95] the authors assert that "autonomy generally means that an agent operates without direct human (or other) intervention or guidance"; this means that an agent must have some kind of control on its actions and its internal state. The authors stress that the explicit representation of mental states, described by attitudes such as beliefs, commitments, obligations, desires, and so on, constitutes a natural way of modeling such kind of control. This assertion is in line with a consolidated research trend [Cohen & Levesque 90][Rao & Georgeff 91] which focuses on the study of a proper representation of agent internal mental activity, conceived as a key factor for achieving high level features such as autonomy, capability to adapt behavior to environmental changes, capability to pursue multiple goals and so on.

The existence of a tight relationship between autonomy and mental activity is also suggested by the definition proposed by [McFarland 94]: "autonomous agents are *self controlling* as opposed to being under the control of an outside agent. To be self-controlling, the agent must have relevant self-knowledge and motivation […]. In other words, an autonomous agent must *know* what to do to exercise control, and must *want* to exercise control in one way and not in another way". The use of terms such as "self-knowledge" and "motivation" suggests once again that an autonomous agent needs an explicit internal mental activity.

A deep investigation about the concept of autonomy is carried out in [Castelfranchi 95]. By starting from the assumption that "autonomy is a relational concept", i.e. that an agent must be autonomous with respect to other systems (mainly the environment and other agents), Castelfranchi deals with two different forms of autonomy, namely cognitive autonomy and social autonomy.

Cognitive autonomy solves the so called "Descartes problem" which can be summarized by the following question: "what agent architecture guarantees that the agent is neither completely determined by stimuli [...], nor completely non reactive to environmental changes?". In order to achieve cognitive autonomy, *beliefs* and *goals* have to be introduced in place of *stimuli* and *reactions*, which are instead at the basis of the reactive approach to agent modeling [Brooks 91] [Maes 95]. In fact, beliefs constitute "interpretations" of stimuli and thus imply some form of autonomous elaboration of direct perception, while goals are "internal manipulated representations of the states to be reached" and, therefore, guide planning and execution of actions.

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 others' goals are taken on the basis of these goals" [Castelfranchi 95].

The aim of this paper is to propose an advancement in the study of the relations between the property of autonomy and the representation of the mental activity of an agent. In particular, we introduce an approach to modeling agent mental activity, based on the concept of active mental entity, and we show how its properties may guarantee, in a natural way, both cognitive and social autonomy.

2. A new architecture for autonomous intelligent agents

In this section we summarize the main aspects of our architecture for autonomous intelligent agents. For the sake of brevity and clarity the description is quite simplified. A more complete and detailed illustration can be found in [Baroni et al. 98b].

2.1 Active mental entities

Our main point concerns the fact that attitudes denoting a particular mental state should not be considered as passive entities (information structures on which agent operates according to fixed procedural rules) but as active entities, endowed with the capability of autonomous operation. The basic ideas underlying our approach are the following:

- We can reasonably understand mental processes as the result of the cooperation and conflict between various attitudes, such as intentions, inhibitions, obligations, hopes, desires, etc.
- Since those attitudes have to dynamically interact to produce a globally intelligent behavior, it is essential that they are provided with individual and independent operation capabilities. Therefore, we call them *active mental entities*, stressing that they are autonomous and can operate and cooperate according to their own goals and strategies.
- Agent architecture is thus a distributed structure: all crucial functions are performed through the free interaction between active mental entities, which can dynamically arise or be disposed.
- In this work we consider only two active mental entities, namely *intentions* and *persuasions.* In [Baroni et al. 98a][Baroni et al. 98b] we have shown that they are sufficient to demonstrate the potential of our approach.

Intentions

The concept of intention is related to that of goal. An *intention* is conceived as an active mental entity committed to pursue a persistent (relativized) goal [Cohen & Levesque 90] (called here *subject* of the intention), under the assumption that a given condition (called here *validity condition*) holds. An intention is carried out through a *plan*. A plan is, in the simplest case, a sequence of tasks. A task can be an elementary action, consisting in a computation, a sensorial acquisition or an action on the environment, or a non-elementary action whose accomplishment leads to the creation of a new intention.

We assume that an intention is able to generate plans; this can be performed by using a suitable mechanism in charge of plan generation or by searching in an already available list of precompiled plans. Then, the intention is capable of choosing the most suitable plan and of putting it at work. Finally, the intention is able to revise the selected plan if the current external situation changes.

Intentions are characterized by persistence, that is they remain active until their validity condition no longer holds, otherwise they are dismissed.

Some intentions are valid only until they are achieved (for instance the intention “I want to find Mr. Smith” is dismissed once Mr. Smith has been actually found), whereas others do not depend on specific achievements (such as “I want to preserve my integrity”). For this reason, we distinguish two kinds of intentions:

- *generated* intentions, which are related to transient goals and are created by other mental entities;
- *primitive* intentions, which are related to permanent goals and, therefore, are always active inside the agent, since they are created with it.

Intentions may interact and conflict one another. A conflict between two intentions may be caused by the access to a shared resource; in this case, the involved intentions try to reach an agreement on which one of them must be achieved first. In particular, if the intentions are primitive, we assume that a *priority* attribute allows one to directly establish the prevailing intention. If one conflicting intention is primitive and the opponent one is not, the latter refers to the primitive intention underlying it, to which the conflict resolution is delegated. Finally, if both intentions are not primitive, conflict resolution involves a more articulated interaction protocol which is described in detail in [Baroni et al. 98b].

Persuasions

In our model, the concept of persuasion is related to that of belief. We say that an agent has a *persuasion* when it is interested in knowing the truth value of an interesting proposition (called here *subject* of the persuasion). For this reason, a persuasion is generated when a new interesting proposition is met and is dismissed when the interest in the proposition ceases; in this sense a persuasion is persistent, that is it remains active until the proposition to which it refers is no longer considered interesting. According to this idea, also persuasions can be modeled as autonomous active entities; in fact, we can assume that a persuasion is not just the passive result of some perceptual or reasoning activity, but that it is an autonomous entity definitely committed to find new elements supporting belief or disbelief in the related proposition.

Propositions may concern long-term knowledge and sensory data collected from external environment; moreover, they can be related to other propositions (for example “I can not move forward”, given that “There is a wall in front of me”). Therefore, a persuasion may update or revise the truth value of the associated proposition when long term knowledge changes, new sensory data are acquired or, finally, other persuasions have updated or revised the truth value of related propositions.

The activity of a persuasion is carried out by a process of searching evidences about the truth value of the related proposition. To perform its activity, a persuasion may cooperate or conflict with other persuasions. In particular, it can happen that several persuasions, concerning the same subject, are active at the same time with contradictory truth values. Since they have different origins and rely on different evidences, they may be conflicting. In this case, each persuasion is stimulated to look for supports for its own thesis and for counterexamples to the theses of opponent persuasions, in order to resolve the conflict through a debate. The conflict resolution protocol is described in detail in [Baroni et al. 98b].

Relationships between intentions and persuasions

The following relationships between intentions and persuasions hold.

When an intention is active, its subject and its validity condition become interesting propositions. Therefore, in correspondence to these ones, persuasions are created. These persuasions notify to the intention if the truth value of the related proposition has changed, that is if the subject of the intention has been reached, or if it is no longer achievable. On the basis of this information, the intention may decide to remove itself.

Persuasions are involved also in the phase of plan generation on behalf of an intention. In fact, also plans rely always on validity conditions; therefore, new persuasions are generated in order to determine the truth values of these validity conditions.

Moreover, during the process of searching evidences about the validity of the related proposition, a persuasion may require the acquisition of data from the environment; this implies the creation of new intentions carrying out data acquisition tasks.

2.2 Agent architecture

An agent has to perform two kinds of activities:

- *reasoning activity*, understood as the result of the interaction, cooperation and conflict between mental entities;
- *operating activity*, regarded as the result of the action of operative entities able to perform mechanical or computational actions.

Such activities are strictly interconnected inside the agent; in fact, decisions arising from the first one influence the actions performed by the operating activity; whereas results produced by low-level actions provide a feedback to the reasoning activity.

To realize the above mechanisms, an agent features a structured internal micro-organization which includes *knowledge* and *components.*

The agent knowledge represents the basic agent competence endowment; it is available to all agent components and can be exploited by them during their operations.

Components may be classified as follows:

- *mental components*, that is intentions and persuasions;
- *operative components*, in charge of performing actions, either mechanical, concerning the interaction with external world through sensors and actuators, or symbolic such as computational activity.

All agent components are understood as autonomous active entities which interact one another according to a message passing paradigm.

2.3 Agent operation

The overall operation of an agent results from the autonomous operation of its components. We assume that reasoning and operating activities are carried out in parallel by the operation of mental and operative components.

A dedicated component is in charge of interacting with the other agents and of receiving their requests. According to the request a new intention is generated, whose subject coincides with the solution of the problem.

The intention is in charge of identifying a set of alternative plans that might lead to the problem solution, of selecting one of them and of realizing it by resorting to the cooperation of other (mental or operative) components. While the firstly selected plan is executed, each intention is capable, through the cooperation with other mental components (namely, related persuasions) of continuously monitoring the environment and revising its plans, if this is required by changes in the external situation.

3. Active mental entities guarantee autonomy

As mentioned in section 1, many different definitions of *autonomous agent* are available in the literature, reflecting different views about the concept of autonomy. Therefore, we divide our analysis in three subsections: the first one concerns the concept of self-regulation, which can be regarded as a first (low) level autonomy, the second and the third ones focus on the higher level properties of cognitive and social autonomy respectively.

3.1 Autonomy vs. self-regulation

A first step towards real autonomy coincides with the definition of *self-regulation* proposed by [Castelfranchi 95]: “self-regulated agents are goal-governed agents, who given a certain goal, are able to achieve it by themselves: planning, executing actions, adapting and correcting actions”. Thus, self-regulation is a necessary, but not sufficient condition in order to achieve autonomy.

A significantly more flexible level of self-regulation is achieved by the so-called “reactive” or “behavior-based” agents [Maes 95][Brooks 91]. However, in this approach, many of the features relevant to autonomy are in some sense “hardwired”: agent goals are fixed and implicit in the definition of behaviors. Moreover, the selection of the current behavior is determined by a fixed set of stimulus-reaction rules, so that the agent is completely driven in its operation by the external stimuli rather than by its (implicit) goals. In a sense, the agent is at the mercy of its stimuli: its behavior depends entirely on them.

Agents endowed with some form of mental activity seem to be able to achieve a more satisfactory level of self-regulation: they feature an explicit representation of their goals and produce and execute plans in order to fulfill such goals. External stimuli give rise to beliefs and the agent operation mechanism includes some method explicitly devoted to update its plans according to its beliefs [Rao & Georgeff 91]. In this approach the agent is no more at the mercy of the external environment: its operation is fully regulated by its internal capabilities. In particular three types of capabilities are crucial:

- the capability to associate action plans to agent goals;
- the capability to modify agent beliefs according to external stimuli;
- the capability to modify previously adopted plans

and/or goals in accordance with modified beliefs.

These requirements for the property of full self-regulation, that we will call *self control*, are in line with the definitions of autonomy proposed by [McFarland 94] and by [Wooldridge & Jennings 95]. The property of self control, in the meaning defined above, can be regarded as the highest level of self regulation. Whereas self regulation involves some requirements concerning mainly the external operation of an agent, self control puts some constraints also on its internal operation. In fact, it is necessary that an agent has the internal capabilities specified above in order to achieve two types of separations:

- separation between actions and goals;
- separations between beliefs and stimuli.

The quick analysis presented above shows that the property of self control can be achieved only by agents featuring an explicit representation of mental activity and that such property represents a more restrictive (and advanced) form of autonomy with respect to the generic definition of self regulation. However, it has to be remarked that self control, as defined above, is a particular case of self regulation. Therefore it can not guarantee, *per se*, the properties of cognitive and social autonomy defined by [Castelfranchi 95]. In the following sections, we review the definition and the features of such properties and show how they can be satisfied by an architectural paradigm based on active mental entities.

3.2 Cognitive autonomy

Cognitive autonomy represents the autonomy of an agent from its physical context: “behavior is influenced by external stimuli but is not determined or imposed by them” [Castelfranchi 95]. The author suggests that cognitive autonomy is achieved when “stimuli are replaced by interpretations” and “reactions are replaced by goals”, and summarizes “the agent reacts to a belief and/or with a goal”. However a key factor for cognitive autonomy is also that “it is impossible to change automatically the beliefs of an agent”. The last crucial property is not guaranteed for a self controlled agent: it depends on the mechanism of belief updating rather than just on the fact that the agent has some internal beliefs. On the other hand, such property is guaranteed by the active persuasions described in section 2: in fact a new persuasion, possibly originated from outside or from another agent, does not automatically overwrite the preexisting persuasions of the agent. However, if there exists a contradiction, the persuasion conflict resolution mechanism, which is an internal feature of the agent, will be in charge of solving it, possibly through the acquisition of further information.

Moreover, consider the case of an agent who regards all its beliefs as equally important and therefore is constantly engaged in verifying whether any minimal aspect of the external world has changed. This very meticulous agent will probably spend much of its time in updating many irrelevant beliefs and will be probably unable to actually carry out any useful action.

This means that, in order to realize cognitive autonomy, it is not sufficient to have a separation between stimuli and beliefs: it is also necessary to be able to select which stimuli have to be taken into account and to focus attention only on the aspects of the environment which are considered important with respect to the current goals. Moreover agent beliefs have not to be bound to the stimuli that are spontaneously provided by the environment, but rather an agent should be able to search actively for some interesting information in the external world.

Therefore, we claim that the definition of cognitive autonomy should be somewhat extended, requiring also that a cognitively autonomous agent has to be able to:

- select the information it is interested in, rather than collecting any stimulus provided by the external environment;
- search for interesting information when it is not immediately available.

These capabilities are not encompassed by the generic definition of self controlled agents endowed with mental states. On the other hand, these capabilities are a built-in feature of the approach based on active mental entities described in section 2. In fact, active persuasions are generated by other mental entities (either intentions or persuasions) 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.

3.3 Social autonomy

Social autonomy, as defined by [Castelfranchi 95], concerns the interaction between different agents and consists of some postulates that can be roughly summarized as follows: an agent has to be available to adopt other agents' goals but has not to do it blindly, since the adoption of exogenous goals has necessarily to be mediated with agent's own endogenous goals.

Also this property is not guaranteed for a self controlled agent, since the requirement that an agent has an explicit representation of its goals does not imply that some of them are endogenous nor that the agent is able to mediate external requests with its preexisting goals.

On the other hand, this property is enforced in the approach described in section 2.

In fact, 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 (for instance self integrity, in the case of a physical agent).

Similarly to the case of persuasions, externally generated intentions (such as those ones deriving from the requests of other agents) can not overwrite preexisting intentions, but rather have to cohabit with them. If an exogenous intention is not compatible with another preexisting intention, a conflict arises and the internal conflict resolution mechanism solves it.

Our proposal is therefore in perfect agreement with the postulates of social autonomy, stating that an agent “adopts other agents' goals as a consequence of a choice among them and other goals” and “only if it sees the adoption as a way of enabling itself to achieve some of its own goals”.

4. An application example

In this section we present an example in order to demonstrate the practical applicability of our approach. In particular, two emblematic situations are analyzed with the purpose of showing how the property of autonomy is realized straightforward by modeling agent cognitive activity through active mental entities and conflict resolution mechanisms.

The example concerns a department mail delivery robot, to which the user consigns an envelope to be delivered to Mr. X. We suppose that the primitive intentions “obey-the-user” and “preserve-energy-level” are innate, and therefore always active, inside the robot. After receiving the request of delivering the mail to Mr. X, a new intention having subject “deliver-mail-to-Mr.X” is generated by “obey-the-user”. This intention may for instance adopt the following plan:

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

Task 1 still concerns a quite generic and high-level task and must therefore be associated to a new intention. While the intention “go-to-the-office-of-Mr.X” is trying to pursue its subject, it may happen that the robot energy reaches the minimum threshold. Then, the persuasion with subject “battery-is-drying-up”, which represents the validity condition of the plan for battery recharge associated to the intention “preserve-energy-level”, becomes true. In particular, the following plan must be put at work by “preserve-energy-level”:

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

Now, a new intention whose subject is “go-to-the-recharging-point” is generated. Thus, both “go-to-the-office-of-Mr.X” and “go-to-the-recharging-point” attempt to control the robot movement system and, since this is a shared resource, intentions enter in conflict one another. In practice, several solutions exist for this conflict depending on the external situation and on plans adopted by the conflicting intentions. Let us simply suppose that the recharging point and the office of Mr. X are in opposite directions; then, since the conflicting intentions are not able to reach an agreement about which one of them must be accomplished first, they delegate the conflict resolution to their generating intentions, that is to “preserve-energy-level” and “obey-the-user” respectively. Each of these intentions has a priority attribute which allows one to solve the conflict immediately; in particular, we can reasonably hypothesize that the first one has higher priority than the latter, so it wins the conflict and propagates this information to the other intentions. This represents a typical case in which a request coming from outside must be postponed to an agent internal need (a so-called “endogenous goal” of Castelfranchi).

Consider now the case that Mr. X is not found in his office. The intention “deliver-mail-to-Mr.X” selects a different actuation plan, namely:

task 1: find Mr. X around in the department
task 2: go near Mr. X
task 3: deliver the envelope to Mr. X

Let us suppose now that, while pursuing task 1, the robot arrives near a glass wall behind which there is Mr. X. Thus, task 2 must be carried out and the intention “go-near-Mr.X” is generated for this purpose. Since Mr. X is standing just in front of the robot, the intention of navigating towards such a fixed target is reduced to the elementary action "go-forward".

While the robot is moving forward, the components associated to the video camera and the sonar acquire and process data about the external world. Doing this, they continuously generate or update persuasions about the environment, whose subject is communicated to a primitive intention “avoid-collision”. In this case, two contradicting persuasions are communicated to the intention “avoid-collision”; in fact, the sonar has the persuasion “there-is-an-obstacle”, whilst the video camera has the persuasion “no-obstacle”. Therefore, in order to solve the conflict, persuasions are put face to face and a debate between them starts. First of all, an analysis of the motivations supporting them is carried out: “there-is-an-obstacle” is justified by the fact that sonar received reflected echoes, whilst the other one is justified by the fact that in the image acquired by video camera nothing but Mr. X is seen. Persuasions are then in charge of looking for evidence or other persuasions corroborating their supports or undermining the opponent's ones. For instance, “no-obstacle” can resort to general knowledge that sonar readings are often erroneous and notify it to the other persuasion. In turn, “there-is-an-obstacle” may reply that sonar readings are erroneous in specific conditions (near wall corners, in presence of noise sources, etc.) that are not met in the present case. Moreover, “there-is-an-obstacle” may attack directly “no-obstacle” by resorting to domain knowledge, that, in the building, there are invisible obstacles (such as transparent glass walls). Since “no-obstacle” is not able to reply to these arguments, “there-is-an-obstacle” prevails: the presence of an obstacle is accepted and the intention “avoid-collision” intervenes to modify the motion plan, going around the glass wall and eventually reaching Mr. X.

By this second example we have shown that, thanks to active persuasions and to the conflict mechanism, some form of sophisticated reasoning about world perception can be realized. This matches perfectly with the seventh postulate of Castelfranchi: “it is impossible to change automatically the beliefs of an agent. The adoption of a belief is a special ‘decision’ that the agent takes on the basis of many criteria and checks” [Castelfranchi 95].

References

[Baroni et al. 98a] P. Baroni, D. Fogli, G. Guida, S. Mussi, Modeling the mental activity of an autonomous agent: an implementation based on intentions and persuasions. *Proceedings of the 14th European Meeting on Cybernetics and Systems Research (EMCSR'98),* Vienna, Aprile 1998, 737-743.

[Baroni et al. 98b] P. Baroni, D. Fogli, G. Guida, S. Mussi, Achieving autonomous behavior through active mental entities: an approach to the implementation of the strong agency concept. TR-199802-16, Università di Brescia, February 1998.

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

[Maes 95] P. Maes, Artificial Life Meets Entertainment: Lifelike Autonomous Agents, *Communications of the ACM*, , vol. 38, no. 11, November 1995, 108-114.

[McFarland 94] D. McFarland, Towards robot cooperation, *Proc. Simulation of Adaptive Behavior Conference*, Brighton, UK, MIT Press, Cambridge, MA, 1994.

[Rao & 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.

[Wooldridge & Jennings 95] M. Wooldridge, N. R. Jennings, Agent Theories, Architectures, and Languages: A Survey, M. J. Wooldridge, N. R. Jennings (eds.), *Intelligent Agents*, LNAI-890, Springer-Verlag, 1995,1-39