

Towards a model of goal autonomous agents

M. Bonifacio¹, P. Bouquet¹, R. Ferrario¹, and D. Ponte¹

Dept. of Information and Communication Technologies
University of Trento (Trento, Italy)

bonifacio@itc.it, bouquet@dit.unitn.it, ferrix@cs.unitn.it, ponte@itc.it

Abstract. In this paper we sketch a model in which agents are autonomous not only because they can choose among alternative courses of action (*executive autonomy*), but also because they can select and pursue new goals on the basis of endogenously generated interests (*goal autonomy*). We use economic models to argue that goal autonomy can be defined by introducing a precise notion of an agent's identity.

1 Executive vs. goal autonomy

The notion of autonomy is central to many definition of (human and software) agent. However, in many agent models – including BDI models – autonomy is thought of as goal-directed behavior and as the possibility of choosing among alternative courses of action. This notion of autonomy, which Castelfranchi [1] calls *executive autonomy*, is very weak, and some authors would argue that it is not autonomy at all.

Executive autonomy is strictly related to traditional theories of choice, based on the paradigm that choosing implies deciding the best course of action in order to achieve a goal [5]. As March pointed out [2, 3], this means-ends paradigm presupposes an anticipatory, causative, consequential type of rationality. The strong limitation of this view is that it presupposes that preferences and goals, once set, cannot conflict with the interest of the agent. However, this form of rationality produces rational behavior only if the environment is stable (or changes in a predictable way), and an agent has complete knowledge about it. Otherwise, it may happen that agents, with no control on their goals, can irrationally pursue unrealistic goals or evaluate goals on the basis of unrealistic preferences.

To overcome these limitations, we suggest that an agent must be *goal autonomous*, namely must have the possibility to decide not only how to achieve a goal, but also which goals are to be preferred and pursued on the basis of an endogenously generated interest. The main object of our research is to propose an agent model in which the source of such an interest is found in a very general principle, which we call the *principle of sunk costs*, which in turn is strongly related to a form of rationality, which March [4] calls *ex-post rationalization*.

2 Identity and the principle of sunk costs

March suggests that rational agents are entities that not only can set appropriate courses of action (including sub-goals) to achieve a given goal, but can also change their mind

about their top level goals and preferences when planned achievements become unrealistic. The research question now is: is there any principled way in which we can explain when and how agents should adopt new goals or change their preferences? Our research theses are: that such a principled explanation is possible; that is based on the notion of an agent's identity; that an agent's identity can be defined in terms of economical principles, namely economies of scale and irreversibility of investments.

In short, the idea is the following. First of all, it is clear that agents sustain costs to acquire a capability or the right to use a resource. Not always these costs are (completely) reversible, and this generates *sunk costs*. Therefore, the more such a capability (resource) is used, the more its costs are amortized (economies of scale effect). Under this respect, we believe that acquired capabilities (resources) are an essential part of an agent identity (what an agent is), and play a crucial role in deciding what goals are to be preferred and pursued on the basis of endogenously generated interests (not using an available capability, especially when it is not reversible, implies a loss of value generated by the lost opportunity of a cost saving!!).

The conclusion is that rational agents should consider not only the current costs of achieving a goal, but also the losses generated by the non-use of sunk investments. Now the point is that in a non predictable environment, circumstances can lead an agent to develop and acquire resources that, to some extent, have no use in order to achieve the current goal. Our thesis is that sometimes the cost of changing one's mind about what is desirable is lower than the cost of going on in the pursuit of current intentions. This happens when, in the decision function, the weight of sunk costs overcomes the weight of current opportunities. In such a situation, instead of reasoning about means necessary to achieve ends that happen to be irrational, rational agents may rationalize their current state as an end which is appropriate to his means, and to change their preferences accordingly. In this sense the sunk cost effect is an attempt to demonstrate the rationality of behaviors that are otherwise not explained and thus labelled as "irrational" by traditional theories of rationality.

References

1. C. Castelfranchi. Guarantees for autonomy in cognitive agent architecture. In M. Wooldridge and N. R. Jennings, editors, *Intelligent Agents: Theories, Architectures, and Languages (LNAI Volume 890)*, pages 56–70. Springer-Verlag: Heidelberg, Germany, 1995.
2. M. D. Cohen, J. G. March, and J. P. Olsen. A garbage can model of organizational choice. *Administrative Science Quarterly*, 17:1–25, 1972.
3. J. G. March. How decisions happen in organizations. *Human Computer Interaction*, 6:95–117, 1991.
4. J. G. March. *A primer on decision making : how decisions happen*. The Free Press, 1994.
5. H. A. Simon. *Reason in human affairs*. Stanford University press, 1983.