Enhancing autonomy with value-based goal adoption

Luis Antunes, João Faria and Helder Coelho

Faculdade de Ciências, Universidade de Lisboa
Campo Grande, 1749-016 Lisboa, Portugal
Ph: +351-21-7500087
Fax: +351-21-7500084
{xarax@di., jfaria@, hcoelho@di.}fc.ul.pt

Abstract

Agents must decide what to do and by what order. Autonomy is a key notion in a multi-agent system, since it is the main justification for the unpredictability and complexity of the environment. Autonomy has a role in deciding which are the new goals of the agent, and it has another in choosing which of the agent's goals are going to be addressed next. Our (BVG) agent architecture relies on the use of values (multiple dimensions against which to evaluate a situation) to perform choice among a set of candidate goals. In this paper we propose that values can also be used to guide the adoption of new goals from other agents. In this setting, we introduce rules for adoption and evaluate their impact by conducting experiments to trace the adoption of goals through a simulated society of simple agents.

Introduction

Some entity takes action in order to fulfil its goals and satisfy its preferences. If some witness fails to make sense of those actions, it could be due to a number of reasons, such as not knowing the entity's goals or preferences, or not grasping the relation between these aims and the chosen actions. The witness shouldn't deem the entity irrational just because it cannot figure out what is happening in its mind. It might assume that justifications for the observed behaviour could exist, and they are just not available to the witness.

In recent papers (cf. [Antunes et al 2000] and its references) we have proposed a new model of rationality that goes beyond utility and encompasses the notion of value as a central component for decision-making for computing importance. With its cognitive machinery roughly included in a Belief Desire Intention (BDI) framework, our agent produces its choices from available alternatives by considering, for the given situation, how each relevant value is affected. A calculus performs choices by collapsing all these partial assessments into a sorted list of actions to be performed. This must be made in execution time, since it is not possible to foretell all relevant possibilities beforehand. Evaluative assessments can evolve, as values are dynamic objects in the mind, and so can be added, removed, or changed.

All these operations are favoured by the contact of the agent with the environment, including other agents, or some user. If we want to achieve enhanced adaptability to a complex, dynamic environment, we should provide the agent with motivations, rather than plain commands. A degree of autonomy is requested, as well as responsibility and animation. Autonomy is a social notion, and concerns the influence from other agents (including the user). An autonomous agent should be allowed to refuse some order or suggestion from another agent, but it should be allowed to adopt it as well.

The BVG Architecture

The notion of preference and its consequences on choice are seldom addressed in the BDI architecture [Wooldridge and Parsons 1998], and most of times by recurring to a restrictive utilitarian approach. In the past few years, we have been working at filling this gap by providing a clear framework for decision making that includes preference-based motivations for choice.

The BVG architecture roughly follows Castelfranchi's principles for autonomy contained in his "Double Filter Architecture" [Castelfranchi 1995]. BVG centres around the notion of value, which is a dimension along which situations are evaluated, and actions selected.

A key issue in the BVG architecture is the update of the choice machinery based on assessments of the consequences of the previous decisions. These assessments can be made in terms of (1) some measure of goodness (the quality of the decision, measured through its consequences on the world); (2) the same dimensions that the agent used for decision; and (3) a different set of dimensions, usually the dimensions that the designer is interested in observing, what amounts to look for emergent behaviour.

(1) and (2) were addressed in [Antunes et al 2000]. We proposed a cumulative choice function $F=\sum c_k F_k$ that, given some goal and alternatives characterised by values V_k, would sort those alternatives out, selecting the best of them for execution. Afterwards, the result of this decision gets evaluated and is fed back into the choice mechanism. An

appropriate function performs updates on the features that are deemed relevant for the decision. For case (1), function G takes an assessment in dimension V_n, and distributes credit to every v_{ki}, feature of winning alternative i. In case (2), a more complex function H takes a multidimensional assessment in $V_1 \times \ldots \times V_n$.

Case (3) relates the evaluation of the experiments to the behaviours of the agents. We have a keen concern for autonomous behaviours. But autonomy can only be located in the eyes of the observer. To be autonomous, some behaviour need not be expected, but also not extravagant. This hard-to-define seasoning will be transported into our agents by an update function such as described above, but one that doesn't depend exclusively on agent-available values.

Autonomy in BVG

We propose several steps in this move towards enhanced autonomy. First, let us observe that adaptation and learning should be consequences of the interaction between the agent and the world and its components. If the agent would adapt as a result of any form of orders from its designer, it wouldn't be adaptation or learning, but design. Of course, if the agent is to respond to a user, this user should be present in its world.

Agents can share information, in particular evaluative information. Mechanisms to deal with this information can be fit into BVG in two different ways. (1) An agent receives information from another and treats it as if it was an assessment made by himself (possibly filtered through a credibility assignment function). Choice can be performed as always, since the new information was incorporated into the old one and a coherent picture is available. (2) The receiving agent registers the new information together with the old one, and all is considered together by the choice function F when the time comes. A new component of F, say F_{n+1}, deals with the 'imported' information separately.

Agents can also share goals. They can pick goals from each other for several reasons. We try to characterise those reasons by recurring to values. We have addressed the adoption of goals by imitation (see [Antunes et al 2000]). Other reasons could be curiosity: the appeal of something new in the absence of foreseen drawbacks; affect: the adoption of some goal just because it pleases (or serves) another agent; etc.

Finally, agents can share values. In humans, the acquisition of these takes years, and relies heavily on some ingredients: some 'higher' value, or notion of what's good and what's bad, that, by the way, we would postulate as a common element to any ontology of values; intricate dedicated mechanisms, that include some of the ones we have been addressing, but also more difficult ones like kid's playing and repeating (as opposed to usual computer-like 'one-shot comprehension' [Monk 1998]). Our mechanisms for value sharing are necessarily simple, and consist basically of conveying values and respective components of the choice and update functions.

Information Exchange

We want our agents to communicate with each other within some agency. We start by describing the way they exchange information. To simplify, agents only tell other agents about their opinion on some feature of the alternative. The agent that receives the information can either (i) incorporate it into its own knowledge, or (ii) just register this information and consider it when making the next choice.

Option (i) leads to an operation similar to an update cause by the agent's own evaluation. If the credibility is 1, the receptor agent considers the information as good as if it were its own conclusion. We give an example for dimension quality. Suppose that agent a (whose credibility for b is c>0.5) tells agent b that quality of w is q, after n evaluations. If b knows nothing about the quality of w, b will accept this information unconditionally. If b believes quality(w, q', n'), then it will change this impression into

$$\text{quality}(w, \frac{(n' \times q') + (n \times q \times c)}{n' + n \times c}, n' + n \times c).$$

In option (ii), agent b will only register the information that was conveyed to it. When later it will have to make a decision, this information will be taken into account as follows. The agent will consider (say) quality(w, $\bar{q}$, $\bar{n}$), the mean values of what it was told concerning the quality of w, weighted by the credibility of the emitter agents. When nothing was told about some dimension of w, a default value is considered. Then b's choice function is changed to incorporate a further component, F_7, which uses F to compute the 'global value of all suggestions.' The new choice function, F', is again $\Sigma c_k F_k$, with k now ranging from 1 to n+1. The suggestions b receives are hence filtered by b's own choice machinery, and they are cumulatively incorporated in F' with the desired weight.

Goal Adoption

Castelfranchi [1995] defends an agent should adopt a goal only if this new goal serves an already existent goal of his. We have embodied this idea into a goal adoption rule and expanded it into value-based goal adoption. The general mechanism we introduced was that of imitation. A possible goal is perceived and considered for adoption according to the agent's values. In the case we wanted to maintain Castelfranchi's rule, our rule was to adopt the candidate goal if there would be already a goal that shared some value with the new goal to be adopted. In the absence of a goal to be served by the goal candidate for adoption, we proposed another rule that would base adoption upon the values concerning the imitated agent:

$$\begin{aligned}
&\text{adopt}(agA, \text{goal}(agB, G_0, V_1=\omega_1, \ldots, V_k=\omega_k)) \text{ if} \\
&\quad \exists \text{bel}(agA, \text{val}(agB, V_i=\xi_i, \ldots, V_j=\xi_j)): \\
&\quad\quad \exists \text{val}(agA, V_i=\xi_i', \ldots, V_j=\xi_j'): \\
&\quad\quad\quad \forall m: i \le m \le j, \text{sameValue}(\xi_m, \xi_m')
\end{aligned}$$

A slightly different rule of adoption by imitation, comes from adapting Castelfranchi's requirement, by weakening the final condition linking the two goals, i.e., the values of

agB are known through the description of its goal G_0, and it suffices for agA to have one value coincident with the ones of agB (we could have used agA's goals to infer its values, but conflicts might occur):

adopt(agA, goal(agB, G_0,V_1=ω_1, ..., V_k=ω_k)) if
$\exists$val(agA, V_a=ξ_a): $\exists i \in \{1, \ldots, k\}$:
sameValue(ω_i, ξ_a)

A new mechanism for value-based adoption would be curiosity. An agent will adopt a goal from another agent if all of the values involved in this goal are completely new to him:

adopt(agA, goal(agB, G_0,V_1=ω_1, ..., V_k=ω_k)) if
$\neg\exists$a: $\{1 \leq a \leq k \wedge$ val(agA, V_a=ξ_a)$\}$

Sharing Values

An agent will share its values by making available their associated parameters and mechanisms. The decision of whether to adopt or not could be taken by considering higher values. At our present stage, instead of postulating one or more higher values, we consider the value affinities between exchanging agents. We propose several rules, such as the existence of (a) one, (b) the majority or (c) all values in common.

When an agent (agA) decides to accept another agent's (agB) particular value (V_k), agB's standards for V_k are accepted by agA, both in terms of scale (S_k), and agB's default position towards that scale. At this stage, agA will also accept agB's choice function component F_k, i.e., agB's relation to V_k, a mapping from S_k to the interval [0,1], where F takes values. We show only rule (c). In what follows, V_k is a value, and s_k, $\xi_k \in S_k$.

Accept(agA, val(agB, V_k=ξ_k)) if
$\forall$val(agA, V_i=ζ_i), $\exists$bel(agA, val(agB, V_j=ψ_j)):
sameValue(ζ_i, ψ_j)

When this condition is met, agA will take on all agB's definitions related with V_k: its scale S_k and respective default mean value s_k; agB's position towards V_k, ξ_k; F_k(agA,·) becomes equal to F_k(agB,·); and F(agA,·) is modified to incorporate F_k with coefficient c_k inherited from agB and adapted to cope with the other agA's c_is.

Finally, all agB's update functions mentioning V_k and relevant to agA's new set of values are transferred to agA. It could be interesting to transfer all update functions mentioning V_k *tout court*, to let agA have contact with new, not completely mastered values (unlike this process of transference, which is rather complete).

Results

A series of simulations addressed the issue of goal exchange. The most interesting result comes from crossing these simulations with the choice-perform-update ones. We concluded that what were the reasons for the selected actions may have been lost, with a high probability. Agents are committed to selections but they no longer have access to the rationality that caused those commitments, and so will have trouble deciding whether or not to keep them. Even worse, this hazards the update process and undermines the whole choice model, since this model by and large depends on a feedback circle that is now broken.

It is apparent that the overlooking of grounded goal management was the cause for this problem. The focus on preference-based choice was too heavy, and a return to a more solidly built mental basis may be in need, before readdressing the choice issues. BDI models could provide this basis.

Conclusions

The main force behind the BVG architecture has always been motivation: what makes us do what we do; what allows us to balance what we ought to do, with what we can do, with what we want to do. And even how do we rationalise a bad (for our own judgement) decision in order not to pain over it again and again. The BVG model provides a framework where multiple values are used to compute practical decisions (reasoning from beliefs to values, to goals and actions), and then these values are updated according to the outcome of those decisions.

Our model departs from BDI by stating that agents need reasons for behaviour that surpass the simple technical details of what can be done, how it can be done and when it can be done. These reasons (which we have called 'dirty,' as opposed to 'clean,' logical ones) are then the basis of the agent's character, and are founded on values as representatives of the agent's preferences. The dropping of those 'clean' reasons was perhaps hasty, and even a preference-conscious agent must possess a sound basis of logical mental concepts on which to ground. It is then only natural to return again to BDI as a provider of such a machinery, and to build the multiple-value model on top of it. Also, the necessity of goals on which to base behaviour. Values are not enough, even when they can cause goals to be adopted. A careful weaving is still in order.

Acknowledgements

This research has been carried out within the research unit LabMAC, and partially supported by project Praxis 2/2.1/TIT/1662/95 (SARA). We would like to thank one of the anonymous referees.

References

Antunes, L., Faria, J. and Coelho, H. 2000. Improving choice mechanisms within the BVG architecture, in *Proceedings of 7th ATAL (Agent Theories, Architectures and Languages),* July, Boston MA (to appear in Springer-Verlag).

Castelfranchi, C. 1995. Guarantees for autonomy in cognitive agent architecture, in *Intelligent Agents, Agent Theories, Architectures, and Languages*, ECAI'94 Workshop (ATAL), Springer-Verlag, Lecture Notes in AI no. 890.

Monk, A. 1998. Cyclic interaction: a unitary approach to intention, action and the environment, *Cognition* 68.

Wooldridge, M. and Parsons, S. 1998. Intention Reconsideration Reconsidered, in *Intelligent Agents V: agent theories, architectures, and languages*, Proceedings of ICMAS'98 workshop (ATAL), Springer-Verlag LNAI 1635.