# Redesigning the agents' decision machinery

Luis Antunes and Helder Coelho

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

**Abstract.** In a multi-agent system, agents must decide what to do and by what order. Autonomy is a key notion in such a system, since it is mainly the autonomy of the agents that makes the environment unpredictable and complex. From a user standpoint, autonomy is equally important as an ingredient that has to be used with parsimony: too much leads to agents fulfilling their own goals instead of those of the user, too little renders agents that are too dependent upon user commands and choices for their execution. 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. We have proposed the BVG (Beliefs, Values, Goals) architecture with the idea of making decisions using multiple evaluations of a situation, taking the notion of value as central in the motivational mechanisms in the agent's mind. The agent will consider the several evaluations and decide in accordance with its goals in a rational fashion. In this paper we extend this architecture in three different directions: we consider the source of agent's goals, we enhance the decisional mechanisms to consider a wider range of situations, and we introduce emotion as a meta-level control mechanism of the decision processes.

## 1. Introduction

We consider a setting in which multiple agents interact in a shared environment. Usually, this environment is computer-simulated. Sometimes it is self-contained and agents are used in experiments to draw conclusions about socially relevant phenomena; in other cases, there is a user to whom the agent responds to, and a certain amount of subservience is expected from the agent.

Whichever the complexity of agents, they must possess a decision component. Even a compile-time pre-specified agent will be of little use if it is not ready for a certain extent of non-forecast possibilities. As the environment gets more demanding in terms of unpredictability (at least a priori unpredictability) more complex should our agent be in what respects to decision flexibility. The designer must have the means to specify what is expected from the agent even in a new environment s/he has never considered. With the advent of mobile computation and huge, varied artificial environments (such as the internet), we have to enhance our agents with autonomous decision skills.

There is a strong and intertwined relation between the decision (especially, choice) and emotional mechanisms (cf. [14], and the architecture proposed in [5]). In this paper we will present an enhanced overall decision mechanism that is able to incorporate this nested relation. Our account of emotions as meta-level control influences over the decision machinery is yet preliminary: the decisions produced by our previous model are too clean, emotion-free. This paper is also the tentative answer to a challenge about how emotions influence our whole life. Our answer follows the ideas of [5]: values are in charge of filtering candidates for later decision taking; emotions control the overall decision machinery.

### 1.1 Decisions and rational agents

When confronted with a decision situation, an agent is defined as rational if he decides in such a way that pursues his self-interest. A classical way of defining self-interest is by adopting utility theory [27], that requires the agent to know in advance all possible situations and be prepared to express his preference between any two states of the world. Not only do these conditions seem difficult to be fulfilled, but also this theory leads to interesting decision paradoxes that show its limitations [17].

An attempt to escape from this kind of bounded rationality was the BDI (Belief, Desire, Intention) agent model [25]. Here, commitment to past decisions is used as a way to decrease the complexity of decisions, since committed intentions constrain the possibilities for the future, and are only abandoned when fulfilled or believed impossible to fulfil. The preferences of the agents are represented by their desires, and these will be transformed in intentions through a deliberation process.

Simon [29] proposed the idea of aspiration levels along multiple, non comparable dimensions that characterise a decision problem. Aspirations are the minimum standards some solution must meet in order to be adopted. The agent adopts and selects for execution the first solution that meets all of the aspiration levels.

### 1.2 Agents with values

In a similar line of reasoning, we have addressed the issue of choice, as one of the central components in the agent's decision machinery [1, 2, 3]. We have proposed the use of multiple values to assess a decision situation. A value is a dimension against which a situation can be evaluated. By dimension we mean a non empty set endowed with an order relation. Most interesting situations from the decision standpoint will have several such dimensions, and so most decisions are based on multiple evaluations of the situation and alternative courses of action. The agent's choice machinery becomes more clear, as agents express their preferences through the use of this multiple value framework. Choice is performed by collapsing the various assessments into a choice function, that cannot be considered equivalent to a utility function, since it is computed in execution time. The multiple values framework we defend can encompass Simon's aspiration levels, but it is more general, allowing for further flexibility, as is shown in [3]. In [2], we present an example of the use of values in a decision problem: imagine someone wants to solve some trouble with his computer. There are several ways of achieving this goal. We show how a value-endowed agent would successively try to do this, given several failure scenarios, and using as values the probability of success of a given action, the level of patience of our agent, the predicted time delay of the action to perform, and the degree of dependence from other agents each action implies.

The coexistence of these values in mind further allows the enhancement of the adaptability decision capabilities by feeding back assessments of the quality of the previous decision into the agent's decision process. Our agents' decisions no longer depend solely on the past events as known at design time. Instead, events are incorporated into the decision machinery as time passes, and the components of those processes evolve continuously to be aggregated just when a decision is needed. This is done by feeding back evaluative information about the results of the decision taken by the agent. In [2] this assessment of (the results of) the decision was done by using some measure of goodness (an abstract higher value). We also suggested other alternatives, such as the agent's own values, or the designer's values (which amounts to looking for emergent features in the agent's behaviour, that is, agents decide by using some system of values, but the designer is interested in what happens to *another* set of values, to which the agents do not have access). Even in the simplest version, the power of adaptability shown by this schema surpasses by far that of choice based on the maximisation of expected utility. It is enough to remember the demands made on the agents by utility theory: they must know *in advance* all available alternatives and preferences between any two of them [15, 17].

### 1.3. Overview of the paper

In this paper we expand on the functionality of the BVG architecture by considering two important extensions. First, the decision mechanism is enhanced to cope with more complex situations. Second, the agent has to decide in absence of all the relevant evaluations. On the other hand, the source of the agent's goals is addressed in the light shed by the multiple values framework: values guide adoption and generation of goals, by providing the respective mechanisms with reasons and justifications. Finally, we tackle the global issue of control, namely how do emotions influence behaviour. We are less interested in external manifestations of emotions (e.g. facial expressions) than in the ways by which emotional life internally guides the process of decision making (in terms of [14], the so-called 'secondary emotions').

In section 2 we will address the notion of autonomy, and relate it with the agent's self-interest. In section 3 we briefly present the BVG agent architecture. In section 4, we expand on the use of values to enhance autonomy in the decision process. Values illuminate the agent's source and choice of goals, by defining the agent's character. In section 5 we readdress the issue of choice, overcoming some limitations of the previous BVG implementation. In section 6 we show how to exploit the advantages of the multiple values framework in goal adoption. Section 7 expands on emotion-driven control of the decision process. Section 8 concludes by pointing out the most important contributions.

## 2. Principles for autonomy

Social autonomy is meaningful for us because we want to consider agents inserted in societies characterised by complex dynamic environments. Agents will have to take into account their past histories, as well as show initiative and reactivity, in order to make decisions in possibly difficult situations. These can include unknown information, unknown social partners and even unknown environments. Autonomy can be considered from several standpoints. Castelfranchi [11] claims that autonomy is a relational concept (mainly, a social concept: an agent is autonomous just in relation to the influence of other agents) and lists various types of autonomy, as follows:

• executive (means) autonomy: an agent is autonomous relative just to the means (instrumental sub-goals), not to the ends. It could imply some level of decision (choice among means), planning, problem-solving, etc.;

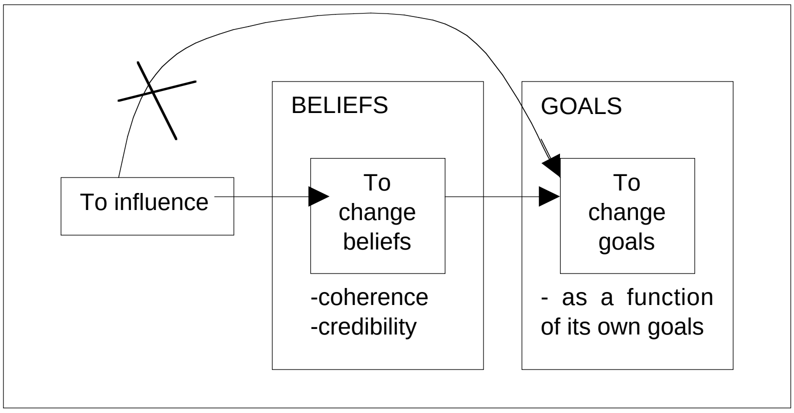

Fig. 1: Castelfranchi’s Double Filter Architecture.

• autonomy from stimuli: the agent’s behaviour should be influenced by external stimuli, but not determined or imposed by them. Behaviour has no causes, but reasons, motives.

In cognitive systems, autonomy is guaranteed by “cognitive mediation:”

• goals autonomy: the system has goals of its own, not received from the outside as commands;

• belief autonomy: an agent controls its acquisition of beliefs.

The following postulates sketch the picture of a socially autonomous agent, and its relations, either with the outside world (including other agents), or with its own mental states, and contribute to the definition of the double filter architecture described in [11] (see fig. 1, and compare it to the one proposed in [5]). Our architecture builds on these postulates, since they solidly draw the basic picture of a self-interested agent, and one that controls its own mental states, thus providing a useful and applicable characterisation of autonomy.

(i) it has its own goals, endogenous, not derived from another agent’s will;

(ii) it is able to make decisions concerning multiple conflicting goals (either its own goals or also goals adopted from outside);

(iii) it adopts goals from outside, from other agents, it is liable to influencing (…);

(iv) it adopts other agents' goals as a consequence of a choice among them and other goals (adoption is thus neither automatic nor rule-governed, and it is not simply required that the goals should not be inconsistent with the current ones);

(v) it adopts other agents' goals only if it sees the adoption as a way of enabling itself to achieve some of its own goals (i.e. the autonomous agent is a self-interested agent);

(vi) it is not possible to directly modify agent's goals from outside: any modification of its goals must be achieved by modifying its beliefs;

(vii) 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. This protects its cognitive autonomy.

We should notice that this version of Castelfranchi's architecture does not take emotions into account (but cf. Castelfranchi's address in [21]).

## 3. The BVG architecture

While following the principles of the double filter architecture, let us draw a new agent cognitive architecture that incorporates the idea of multiple values (fig. 2), and does not include emotions. This new sketch expands on the one presented in [2], where the primary focus was laid on choice. Now, we turn our attention towards control.

We can see the execution cycle of an agent divided in three phases: perception, deliberation and execution of an appropriate action. This description is a crude simplification of a much more complex process, involving several parallel threads, with intricate relations among them. For instance, the perception is clearly guided by the agent's current focus of attention, which cannot be dissociated from its own motivations. If some relevant feature alteration is detected by the agent, it must be immediately taken into consideration, even if the decision process which included the previous value of the feature is almost concluded. We can't imagine that the world will wait until we take our decisions. But generally we have an idea about how much time we can use to decide, and how stable are the beliefs upon which we base our decisions [28].

To start with, we propose as a reference a schema of decision which includes goals, candidate actions to be chosen from, beliefs about states of the world, and values about several things, including desirability of those states. This is because we don't want to overload the architecture with too many ingredients, and related mechanisms. It is preferable to keep things manageable, and see how far we can go.

Decision is a complex, multi-staged process in an agent's mind. One stage deals with the origin of goals. Agents can either adopt goals from other agents or generate goals internally, as a result of internal processes. In another stage, goals are considered against other objects in the agent's mind, such as beliefs (which include plans about their feasibility) and classified accordingly. For instance, we can have suspended goals, active goals, etc. Finally, among the active goals, the agent has to serialise them into execution. This is the choice phase, which we have addressed in previous papers, and will expand further herein. One way of achieving this without over-determining the behaviour of the agent, and even so reducing the deliberation time would be to have several partial ordered sets of actions, and computing the final order only when the moment arrives.

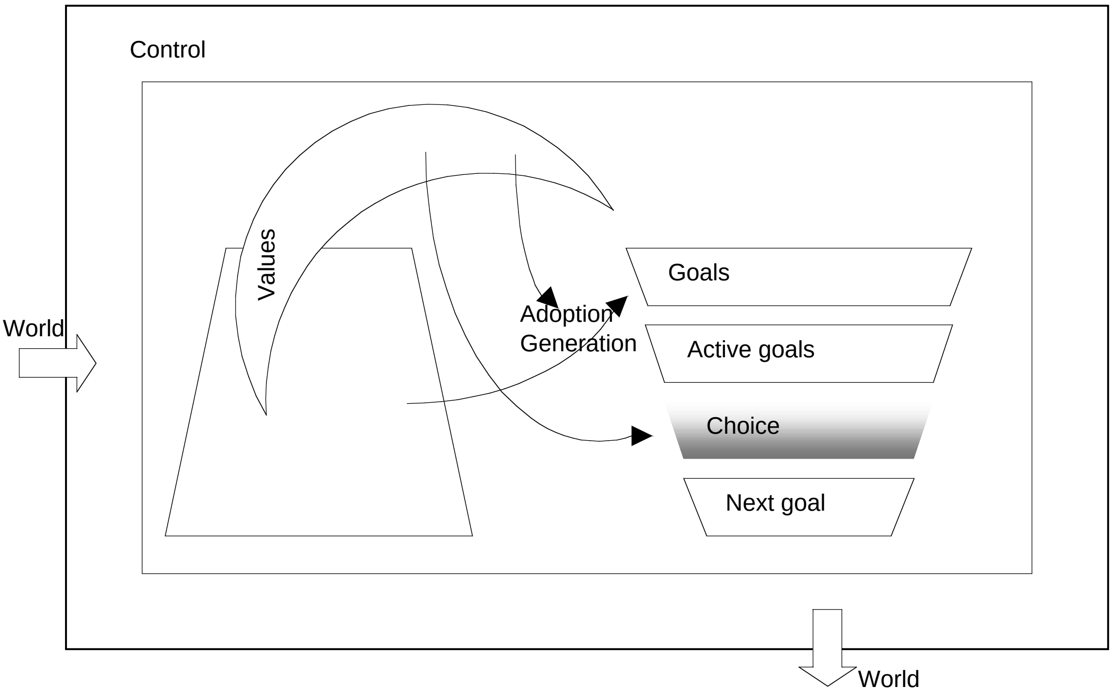


Fig. 2: The BVG architecture.

The BVG architecture introduced in [2] paid special attention to choice. Now it is time to get the whole of the agent's decision machinery in context, and consider a broader picture. The role of beliefs and goals in an architecture such as the above has been thoroughly examined in the literature [13, 10]. The role of values was addressed in [1, 2], and further cleared here. The biggest challenge in BVG nowadays remains to be control. In fig. 2, control mechanisms wrap the whole of the architecture; this is meant to represent a layer responsible for a lot of important mechanisms in the agent's mind. Many of these are cognitive, along the BDI track. In [12] a set of facets were included as control mechanisms of the whole agent architecture: importance, urgency, intensity and persistence. Other control mechanisms are arguably emotional [22]. However, the association between the rational and emotional issues is still unclear [5], and our conjecture is: emotions have a dominant role in controlling the decision-making process.

As explicitly depicted in fig. 2, control mechanisms act over perception and also action. This is meant to represent the myriad of control issues involved in these two activities. Take an action that was scheduled to be executed. There are many things that can go wrong, some perhaps forecast, some completely new to the agent. Even if everything goes well, control cannot rest, for there can be opportunities to be taken, records to be made, evaluations to be performed about how well were expectations met. What kind of machinery should this control layer possess in order to decide *only* if the agent's reasoning should or not be interrupted, and the issue of the current action be reconsidered? How many things are there to control, how many features are there to be examined to solve such a (meta-reasoning) problem [26, 15]? A lot of other roles of control are left implicit in fig. 2. Just as over external actions, control must be exerted over internal actions, such as deductions, constrainings, updates, accesses, etc. One possible solution was suggested in [31]: redundancy. We defend that agents should perform redundant calculations and have triggers that allow normal reasoning to be interrupted. Look below (section 4.2) for a possible such mechanism. [9] presents similar ideas concerning control, when the authors defend the multidimensional mind (the unidimensional mind covers only the use of utility functions) and postulate the need for a balance between action interests and high level normative motives. Imagine another situation, put forward by LeDoux [18], where a bomb is involved and people around get into panic with terrible fear. What influence did this fear have in the people's decision making? Before we further address the answer to this question let us tackle some more architectural issues.

## 4. Autonomy and character

As we have seen in Castelfranchi's analysis, autonomy is involved in all of the stages of decision. An agent that generates its own goals is certainly more autonomous than another that doesn't. A further flavour of autonomy is the ability to adopt goals from other agents.

The selection of the next action to be executed results from successively constraining the set of candidates for goals. Following the path in fig. 2, we see that from the set of possible goals we extract the actual goals, either by adoption or by generation. Among these, some are considered active. For instance a goal known to be impossible to fulfil could not be considered active. Some of the active goals are then selected to be executed, and so are sorted according to some criteria.

So we have three processes by which our agent constrains its set of goals. The (1) *creation* of goals is the first phase, including both adoption and generation of goals. After that, the (2) *selection* of active goals occurs. From the active goals the agent (3) *chooses* the goals to be executed. Most of the literature on intelligent agents only focus on the selection phase, because it is the 'cleanest' one, the most technical one. It is possible to accomplish an interesting performance level relying only on technical reasons to limit the set of goals, such as a deductive machinery with modus ponens. See, for instance, the BDI (Belief, Desire and Intention) architecture [24, 25]. Usually the choice phase is either not addressed or implemented with very simple heuristic solutions, such as using simple utilities. Also, the creation phase, many times, is oversimplified by just postulating that the only source of goals is the agent's user.

However, when we want to define autonomous agents, we must address all of these phases, which raises the issue of the agent's character. The character can be defined as the set of collective qualities, especially mental and moral, that distinguish a person or entity. Autonomy implies difference, i. e., the agents should be free to decide differently from one another even in the same situation. Therefore, agents can have different characters. Either in the creation (1) or in the choice (3) phases, the reasons can vary from agent to agent. In the BVG architecture, it is in this two phases that we can locate the personality traits that define the character. In the creation phase, some agents will adopt goals whereas others don't, some agents will generate goals in one way, while others may not generate goals at all, and only obey goals from their creator. In the choice phase, we look at different rationalities in different agents. Some choice is not necessarily irrational just because we cannot identify the reasons that led to it [15]. Even considering agents endowed with the same set of goals and beliefs, their different sets of values should be enough to produce differentiated behaviours. For example, a soccer robot can decide to shoot weakly at goal because he is too close to the goalkeeper and thinks there's some risk of endangering him. Finally, the influence emotions exert in decision making, can also be tuned up differently in different characters. Some agents will be bolder, or more timid, or more risk averse, etc. Subsequently (cf. section 7), we will see how this can be achieved in BVG.

### 4.1 Generation and adoption of goals

The source of goals is usually considered to be the designer of the agent, and acquisition of goals is done once and for all in an a priori fashion. This is clearly very limitative in what concerns autonomy. Acquisition mechanisms are necessary so that the agents can profit from all possibilities the environment can offer.

Castelfranchi proposes to only adopt goals when these are instrumental to other previously existing goals. This is an oversimplistic solution to the adoption problem, that is founded in architectural limitations, the fact that only beliefs and goals are available. This mechanism of adoption would lead to too rigid a structure of goals: if all goals are instrumental to previously existing goals, all sub-goals must be justified by these other higher ranked goals. Who provided these ones? Since all adoption is guided by them, they must have been adopted in design time. But then these higher (or end-) goals are the only ones important, all others could be abandoned as easily as they were adopted, if the agent discovers their uselessness for fulfilling the end-goals. So the agent's autonomy is severely restricted to executive autonomy. How can we then represent the agent's self-interest?

We propose to use values to help define the source of goals, guiding either adoption or generation of goals. When an agent identifies an object of the world as a candidate goal, it can decide to adopt it for a number of reasons. Of course if this new goal is instrumental to a previous existing goal, it should be a stronger candidate, but don't the characteristics of the existing goal matter? We think that the candidate goal should possess some added value itself. On top of the technical reasons (e.g. the new sub-goal allows the super-goal to be fulfilled), other reasons can lead to adoption. If some behaviour is perceived as leading to a situation assessed as desirable, the agent can adopt as their own the goals that triggered that behaviour. And this desirability can and should be expressed in terms of the agent's system of values.

Candidate goals come from everywhere in the agent's life. Virtually any interaction with the environment and its inhabitants can provide candidates for goals. Mechanisms for adoption include imitation of other agents, instincts, training and learning, curiosity, intuition, etc. We can use the values framework to fine-tune these mechanisms, and even apply the same adjustment methods we used in the choice mechanism to enhance them. In section 6 we will see how these mechanisms can be defined in the BVG architecture.

Goal generation mechanisms can also be based on the agent's values. When a particular state of the world is evaluated as desirable, the agent can transform this description of the world into a goal to be achieved. Goal generation is in a lot of ways a more complex problem than goal adoption, since it has to do with creativity, and will be left out in the rest of this study.

### 4.2 Enhancing the choice mechanisms

In [2] we have seen how choice can be guided by values, and how this choice mechanism can be adapted in accordance with evaluation of successive results. But in that paper, we worked out our models by considering severe limitations that must be overridden. We had a fixed set of values, against which all candidate actions could be compared. We also had a fixed decision function (linear combination of the multiple evaluation of the candidates) and a fixed value adaptation function. Finally, we had only two types of values: aspiration-type values (e.g. probability of success), and capital-type values (e.g. time delay).

Even when considering only the single-agent case, every one of these limitations is unacceptable since they radically restrain the expressive power of the model. When the multiple agent case is considered, they are even more damaging. Consider the ability to communicate values, and consequently goals and actions. Several steps must be performed in planning the experiments, making this path follow the agent's classification according to the use of values. First, increase the number of values present in the choice setting. Then we can consider options that are characterised by values which were not necessarily foretold. That is, the agents must perform choice even in the presence of incomplete information. Afterwards, we will consider a variable set of values. Given the previous setting, this should not raise any problems, but in this case it is important to observe the agent's behaviour in the long run. Remember that value adaptation is performed over the candidate actions, even if this wasn't the only option we could have made. So we must increase the number of candidate actions, to see how the model copes, and check whether an alternative formulation should be used.

An interesting alternative could be the use of value accumulators. By adapting the idea of [6] for memory access, we can have an indicator of how much a value was referred to by other mental objects. A competition among values is cast, allowing the agent to have a relative idea of the importance of the various values. This schema does not necessarily substitute the usual schema of evaluation/choice/adaptation, it just allows the agent to subvert this in special occasions. For instance, if an option is being considered that refers to an important quantity of the highest rated value, execution stops and all calculations are restarted. This could be a simple way of implementing a pre-emptive mechanism, which we deem extremely important in an open environment.

### 4.3 Ontology of values

The BVG architecture is sufficiently general to cope with a series of diversified situations. By selectively choosing the set of relevant values and associated mechanisms for choice and for value adaptation, the designer can create instances of BVG agents which are adequate to his own setting. However, we can conceive specialisations of the most general BVG agent that provide a default basis for the development of agents. One such step has already been taken [2] when we considered the notion of goodness of a decision to help recalibrate the value system that led to it. This goodness is a kind of higher value that would be present in most agent designs (see section 7).

Another strong candidate to be present in a lot of agent designs would be survival (or vital energy) as a fundamental value that would govern most decisions. If we launch a group of robots to explore Mars, we would like them to keep functioning as long as they possibly can, even if they would sometimes undermine other valued possibilities. Survival is particularly important since it is the fact that the agent keeps working that allows the result of its efforts to be exploited, passed along to its user, etc. Several years ago Sloman [30] proposed to characterise goals with three features: urgency, intensity and persistence. These are also candidates to be usually present as values when tackling a choice situation.

As with the credibility of a belief, also the probability of success of a given action is likely to be considered when choosing that action. With time, a designer can develop an ontology of values that characterise most of the choice situations his agents face, or at least to have several ontologies adapted to classes of problems.

## 5. New mechanisms for choice

In [2], our agent had a goal to fulfil that was characterised by some values. The candidate actions were all comparable according to the same values, so choice was straightforward. The agent just computed a real function of those values for all the alternatives, and the highest scorer was chosen. Afterwards, the agent looked at the results of his action by assessing the resulting state of the world against a dimension of goodness (cf. section 1 and [3]), and updated the values appropriately.

We now propose to tackle choice situations where the agent doesn't have all the relevant information. Imagine some goal G is characterised by targets for three (out of five) values: $V_1=\omega_1$, $V_2=\omega_2$, $V_3=\omega_3$. Let the candidate actions be represented by sub-goals $G_1$ and $G_2$, with associated values, respectively: $V_1=\omega_{11}$ , $V_4=\omega_{14}$ , and $V_1=\omega_{21}$ , $V_2=\omega_{22}$ , $V_5=\omega_{51}$.

As long as choice is performed by using a linear combination of some function of the paired (e.g. $\omega_1,\omega_{11}$) values, like in [2], one can just omit the values outside of the intersection of the goal and the candidate characterisation, thus rendering the other values not redundant. If other types of choice functions are used, one must proceed with more care. Anyway, even in the simple case above, there are open problems to be dealt with. First of all, it is necessary to characterise the new adopted goal, say $G_2$. Should $G_2$ include values for $V_3$? Should it keep values for $V_5$? We think that the answer to both questions is positive: $V_3$ should be kept (with target $\omega_3$) because we are adopting $G_2$ just because of G. So it is only fair that we keep whatever guides the attempts at achieving G, and those are the values $V_1$, $V_2$, and $V_3$. For an analogous reason we should include $V_5$. It could be the case that $V_5$ represents important evaluative notions to be considered during the attempts at $G_2$, and so we mustn't give up $V_5$ for our future execution. In both cases, these values will help control the execution of the agent towards his goals, possibly allowing for revisions and recalculations if the chosen goals no longer serve the relevant values at stake.

## 6. Goal adoption

In this section, we illustrate our ideas about goal adoption by proposing a concrete mechanism for adoption: imitation. To simplify, assume an agent perceives as a possible goal [23] some action he observed another agent carrying out (or possibly as a result of some interaction). In the multiple values framework, any object carries with it some characterisation in terms of values. Let us further assume this set is non-empty.

If we take Castelfranchi's rule for adoption, our agent will adopt this possible goal as a goal of his own, only if he perceives this new goal as serving one of his previously existing goals:

Adopt(agentA, goal(agentB, $G_0$)) iff

$\exists$goal(agentA, $G_1$): $\exists$plan(agentA, $G_1$, $P_1$, …, $P_n$): $G_0 \supset G_1$

These are what we have called technical reasons for adoption. In BVG, the agent has reasons for adoption that are founded in his values, that represent his preferences. If we want to maintain Castelfranchi's rule, our rule of adoption by imitation could be to adopt the candidate goal if there is already a goal that shares some value with the new goal to adopt:

Adopt(agentA, goal(agentB, $G_0$,$V_1$=$\omega_1$, …, $V_k$=$\omega_k$)) iff

$\exists i\in\{1, \ldots, k\}$: $\exists$goal(agentA, $G_1$,…, $V_i$=$\omega_i$', …): $\omega_i * \omega_i' > 0$

In the absence of a goal to be served by the goal candidate for adoption, we could propose another rule of adoption by imitation that would base adoption upon the values concerning the imitated agent:

Adopt(agentA, goal(agentB, $G_0$,$V_1$=$\omega_1$, …, $V_k$=$\omega_k$)) iff

$\exists$bel(agentA, val(agentB, $V_i$=$\xi_i$,…, $V_j$=$\xi_j$)):

$\exists$val(agentA, $V_i$=$\xi_i$',…, $V_j$=$\xi_j$'): $\forall l\in\{i, \ldots, j\}$ $\xi_l * \xi_l' > 0$

This mechanism (imitation) is also interesting for an emotional machinery (cf. [18]). Young children imitate emotional reactions much before they internalise the real emotions. Damasio's secondary emotions are cognitively generated, and arise later, as a result of identified "systematic connections (…) between primary emotions and categories of objects and situations" [22]. Children build these mental connections by going through imitation and game-playing: they emulate the emotional behaviour. In the panic situation mentioned above, people use imitation as an emotional-driven goal adoption strategy. There is time to consider only one or two alternative courses of action (run, dive), and then a decision must be reached.

## 7. Affective reasoning

In [2], the adaptation of the agent system of values has three possible sources of evaluative information, that assess the quality of the decision taken: (i) some measure of goodness; (ii) the values that led to the decision themselves; and (iii) the designer's values. We now propose an alternative meaning for these assessment measures.

We suggest that the goodness of a decision (i) amounts to the agent's affective appraisal of the outcome produced by that decision. That is, goodness is not an abstract notion, independent of whoever performs that assessment, it is subjective, and it involves the agent that performs the assessment, in the exact conditions the assessment is made, and considering the conditions in which the decision was taken.

Alternative (ii) is the 'clean' option, for is performed in the same terms as the decision. That is, if one is already looking to optimise some things, just go and measure them. Observation subjectivity apart, the agent should have no difficulty with this. So, these values do not appear very interesting as a candidates for emotion arousal. Since goodness (i) was not considered in deciding (unless decision was taken using simple utility), the natural extension is to consider a different set of evaluative dimensions to assess the outcome of the decision (iii).

The interesting possibility about the designer's values is to consider this set to be the set of emotional features raised by the issue at stake (see figure 3d/e). These can be either inspired by the agent's own affective perception, or driven by some interaction language that interfaces to the user, or other agents. We humans frequently evaluate our decisions by examining the emotional reactions they raised in the surrounding people [18]. In fact it is often like this we notice we goofed up. Of course, we raise here a regress problem: if we want to enhance our decision capabilities by using others' views about the result of our actions, we should use our decision skills to judge whether to trust our impressions about those views (since we don't have direct access to them). To solve this problem we must accept that subjectivity has its limits: we only know what we know, and never what others know [4].

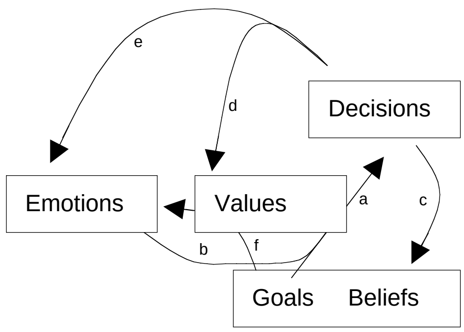


Fig. 3: The dynamics of emotions. (a) value-informed decision making; (b) emotions exert control over decision; (c) watching consequences of decisions; (d) feeding back information to enhance decision machinery; (e) getting emotional information about the results of decisions; (f) cognitive evaluations arising emotions.

The view just described concerns primarily the observation of emotions raised as a consequence of the agent's behaviour (figure 3e). In a sense, it is an external view of emotions, although considering the consequences of the emotional assessment in the calibration of the decision machinery. What is lacking here is the direct influence of emotions in decision making (figure 3b), as suggested by recent literature (especially [5]). At this point of our research, we can only point out a broad picture, which we deem coherent, and also consistent with the most recent neuroscience views on emotion and its strong influence on decision making.

In [7,8], emotions are considered as meta-level controls that help the agent organism to regulate itself in order to improve its adaptation to the environment. They have an impact on cognitive processes such as the allocation of cognitive resources, attention focus and adaptive learning, either boosting some ones or freezing others. Also here, the key idea is that emotions establish control over the decision making process. They serve as signals, or somatic markers, to be used in case retrieval, like a sort of pre-cognitive assessment of situations, based on similar experiences from the past. Emotions also determine the amount of time the agent will have for the decision process. One strategy could be to lead the agent to a quick reaction, as a way of gaining some time to spend on a more thorough decision process. For instance, in LeDoux's bomb example, one dives into the ground, and freezes any decision while considering other options.

Let us go a little further into the detail of these mechanisms. An agent should be ready for action taking at any point of its life (or it won't survive long in a hostile environment). This means that although deliberation is certainly useful and many times unavoidable, it should possess anytime character. If we take this idea to its limits, we must include in our architecture some immediate feedback mechanism, what is usually called a reactive component. We propose that this spectrum from reactive action (deliberation time equals zero) to fully deliberative action (unlimited deliberation time) can be achieved through a cumulative deliberation framework. Emotions tune up the amount of time available for the choice function to perform (for instance an additive function). In a strongly emotional stress situation (as in a state of fear, or panic), actions are taken from a small repertoire of readily accessible candidates, possibly with no deliberation at all (e.g. we do the first thing that comes to mind [18]). With some more time available, our cumulative assessment can begin: we pick a larger repertoire of agents, at the same time start the calculation of their choice value. This is done by considering the most important value, and performing the choice calculation for each of the candidates with respect to this value. At each moment, we have an assessment (a rating) that allows us to perform choice. With infinite time available, choice will be perfect, i.e. in accordance to our beliefs and evaluations about the relevant issues at stake. This broad picture is compatible with Damasio's evidence that "Normal people choose advantageously before realising which strategy works best" (the idea of importance, of valued and weighted decision), and that "non-conscious biases guide behaviour before conscious knowledge does" (emotional reactions are triggered before any rational deliberation) [5].

When we evaluate some situation, weighting pros and cons, we mix emotions and feelings with reasons, and if we succeed, we may say that the whole process is a mixture of reasons and emotions. Emotions are not simple reflexes, but patterns (collections of answers) supporting fast reactions (in order to maintain us alive), faster than cognitive actions, because from an evolution point of view the emotional machinery is older than the cognitive machinery. Quoting from [5]: “The sensory representation of a situation that requires decision leads to two largely parallel but interacting chains of events”: emotional ones based upon previous individual experience (complex process of non-conscious signalling) and rational ones based upon processes of cognitive evaluation and reasoning.

We now take the theory of emotions in [20], in what respects the variables that affect emotion intensity. [19] lists the variables affecting anger: the degree of judged blameworthiness, the degree of deviation from personal or role-based expectations, and the degree to which the event is undesirable. We have already seen how emotions influence decision making (fig. 3b). Emotions also influence cognitive acquisition (for instance, through determining attention focus [6], and so through a decision-mediated process, see fig. 3b/a/c). Now we see what influences emotions. [19] tells us that “emotions are dependent on beliefs,” and that “emotions are positively or negatively valenced reactions.” Notice that all those prototypical variables above are cognitive assessments, evaluations about the reciprocal influences between the event and the agent’s informational and pro-active states. That is, we have beliefs, goals and (what we call) values influencing emotions. The circle is completed, and we have now cognitive values as the basis of emotion arousal (figure 3f).

## 8. Concluding remarks

In this paper, we have restated the fundamentals of the BVG architecture, namely the use of multiple values to inform choice, and the feedback of assessment information to recalibrate the choice machinery. We took Castelfranchi’s notions about agent autonomy to enhance the BVG architecture in two ways. First, we considered and overcame some limitations of the previous implementation, such as the inability to deal with incomplete evaluative information. Second, we proposed to use the agent’s system of values to inform the mechanisms of creation of goals. As an example, we presented rules for adoption that could implement the mechanism of imitation. The primary concern throughout the paper is agent autonomy, and its relations with the agent’s character, or personality. We defend that the multiple values framework is especially adequate to provide the tools to successfully tackle these issues. This was further elaborated by the suggestion that not only cognitive but also emotive aspects of the agent’s reasoning can be addressed in this framework. Emotions play a determinant role in decision processing. They influence and are influenced by decisions and their consequences. Together with values, emotions provide the fundamental mechanisms for efficient and adaptable decision processing.

We could only hint at applications of these ideas. Some experiments were made, but the results are not yet conclusive. So, future research will aim at experimental demonstration of the ideas presented herein. Other research issues remain to be the increase of the agent’s adaptability by enhancing the use of the fed back information, and the expansion of an agent’s system of values as a result of interaction with other agents.

**Acknowledgements.** We wish to express our thanks to João Balsa, José Castro Caldas, João Faria, José Cascalho, Cristiano Castelfranchi, Rosaria Conte, Maria Miceli, Luis Moniz, Leonel Nóbrega, Pedro Rodrigues, Jorge Louçã, the editor of this volume, and the anonymous referees. This research has been carried out within the research unit LabMAC, and partially supported by project Praxis 2/2.1/TIT/1662/95 (SARA).

## References

1. Antunes, L., Towards a model for value-based motivated agents, in proceedings of MASTA'97 (EPIA '97 workshop on Multi-agent Systems: Theory and Applications), Coimbra, October 1997.
2. Antunes, L. and Coelho, H., Decisions based upon multiple values: the BVG agent architecture, in Barahona, P. and Alferes, J. (eds.), Progress in Artificial Intelligence, proceedings of EPIA'99, Springer-Verlag, Lecture Notes in AI no. 1695, September 1999.
3. Antunes, L. and Coelho, H., Rationalising decisions using multiple values, in proceedings of the European Conference on Cognitive Science, Siena, October 1999.
4. Antunes, L., Moniz, L. and Azevedo, C., RB+: The dynamic estimation of the opponent's strength, in proceedings of the AISB Conference, IOS Press, Birmingham 1993.
5. Bechara, A., Damasio, H., Tranel, D. and Damasio, A. R., Deciding advantageously before knowing the advantageous strategy, Science, Vol. 275, 28 February, 1997.
6. Botelho, L. and Coelho, H., Emotion-based attention shift in autonomous agents, in Müller, J.P., Wooldridge, M.J., Jennings, N.R., Intelligent Agents III, Agent Theories, Architectures, and Languages, ECAI'96 Workshop (ATAL), Springer-Verlag, Lecture Notes in AI no. 1193, 1997.
7. Botelho, L. and Coelho, H. Adaptive agents: emotion learning, Proceedings of the Workshop on Grounding Emotions in Adaptive Systems, Fifth International Conference of the Society for Adaptive Behaviour 98 (SAB'98), Zurich, August 21, pp. 19-24, 1998.
8. Botelho, L. and Coelho, H. Artificial autonomous agents with artificial emotions, Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98), Minneapolis/St. Paul, May 10-13, pp. 449-450, 1998.
9. Caldas, J. M. C. and Coelho, H. The origin of institutions, socio-economic processes, choice, norms and conventions, Journal of Artificial Societies and Social Simulation (JASSS), Vol. 2, No. 2, 1999.
10. Castelfranchi, C., Social Power. A point missed in Multi-Agent, DAI and HCI., in Decentralized AI - Proceedings of MAAMAW'90, Demazeau, Y. and Müller, J. P. (Eds.), Elsevier Science Publishers B. V., Amesterdam, 1990.
11. Castelfranchi, C., Guarantees for autonomy in cognitive agent architecture, in Wooldridge, M.J., Jennings, N.R., Intelligent Agents, Agent Theories, Architectures, and Languages, ECAI'94 Workshop (ATAL), Springer-Verlag, Lecture Notes in AI no. 890, 1995.
12. Corrêa, M. and Coelho, H. From mental states and architectures to agents' programming, in proceedings of the 6th Ibero-american Conference in Artificial Intelligence, Lisboa, October 5-9, 1998, Coelho, H. (ed.), "Progress in Artificial Intelligence - Iberamia'98," Lecture Notes in AI no. 1484, Springer-Verlag, pp. 64-75, 1998.

13. Cohen, P. R. and Levesque, H. J., Intention=Choice+Commitment, in proceedings of AAAI'87, 1987.
14. Damasio, A., Descartes' Error: Emotion, Reason, and the Human's Brain, G. P. Putnam's Sons, New York, 1994.
15. Doyle, J., Rationality and its roles in reasoning, Computational Intelligence, vol. 8, no. 2, May 1992.
16. Elliott, C., Research problems in the use of a shallow artificial intelligence model of personality and emotion, in proceedings of AAAI'94, 1994.
17. Hollis, M., The Philosophy of Social Science - An Introduction. Cambridge: Cambridge University Press, 1994.
18. LeDoux, J., The Emotional Brain, Touchstone (Simon and Schuster), New York, 1998.
19. O'Rorke, P. and Ortony, A., Explaining Emotions (Revised), Tech. Rep. ICS-TR-92-22, University of California, Irvine, Department of Information and Computer Science, June 1993.
20. Ortony, A., Clore, G. and Collins, A., The Cognitive Structure of Emotions, Cambridge University Press, Cambridge MA, 1988.
21. Paiva, A. and Martinho, C., Proceedings of the workshop on Affect in Interactions (towards a new generations of interfaces), of the 3rd i3 annual conference, Siena, October 1999.
22. Picard, R. W., Affective Computing, M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 321, November 1995.
23. Pollack, M. E., Overloading Intentions for Efficient Practical Reasoning, Noûs, vol. 25, no. 4, 1991.
24. Pollack, M. E. and Ringuette, M., Introducing the tileworld: experimentally evaluating agent architectures, in proceedings of AAAI'90, 1990.
25. Rao, A. S. and Georgeff, M. P., Modeling Rational Agents within a BDI-Architecture, in proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning, Morgan Kaufmann, 1991.
26. Russell, S., Rationality and Intelligence, Artificial Intelligence, vol. 94 (1-2), Elsevier, July 1997.
27. Russell, S. and Norvig, P., Artificial Intelligence: A Modern Approach, Prentice Hall, 1995.
28. Russell, S. and Wefald, E., Do the right thing - studies in limited rationality, The MIT Press, 1991.
29. Simon, H., The Sciences of the Artificial (3rd edition), the MIT Press, Cambridge, 1996.
30. Sloman, A., Motives, Mechanisms and Emotions, in Emotion and Cognition 1, 3, 1987.
31. Sloman, A., Prolegomena to a Theory of Communication and Affect, in Ortony, A., Slack, J., and Stock, O. (Eds.), AI and Cognitive Science: Perspectives on Communication, Springer-Verlag, 1991.