From Reactive to Emotion-based Agents: a prescriptive model

Rodrigo Ventura and **Carlos Pinto-Ferreira**
Institute for Systems and Robotics (ISR)
Av. Rovisco Pais, 1
1049-001 Lisboa
PORTUGAL
{yoda,cpf}@isr.ist.utl.pt

Abstract

The competence level of reactive agents is usually compared with the one of insects. Taking the model of a simple reactive agent as a starting point, this paper proposes two additional levels of competence. These new levels are inspired on neuro-physiological models of emotions, namely on the somatic-marker hypothesis proposed by Antonio Damásio. The first competence level corresponds to a simple reactive agent, while the second level is based on the idea of processing simultaneously two distinct representations of the same stimulus: a simple one, oriented towards feature extraction (shared with the first competence level), and a complex one, oriented towards pattern recognition. The third level makes use of the "movie-in-the-brain" idea, which allows the agent to establish cause-effect relationships about its interaction with the environment.

1 Background

The topic of emotions in Artificial Intelligence is very controversial. On the one hand there is a growing community of researchers proposing either experimental or theoretical arguments in favor of the study of emotions, and on the other hand there are those who dismiss emotions as something superfluous to rationality. To give two salient examples, John McCarthy[1] has stated that it is "artificial" to include emotions in machines, while Marvin Minsky[2] questions whether it is possible to build intelligent machines without emotions.

The origins of the interest for studying emotions in AI can be traced back to two main sources: human-computer interaction (HCI) on an affective basis [Picard, 1995], and the study of the contribution of emotions to intelligent behavior (for instance, see [Sloman, 1999]). The former has been extensively advocated by Rosalind Picard and her research group. Her goals include building devices that recognize emotion expression in people and developing systems which express believable emotions. The idea is to research the interaction of humans with computers in order to include an emotional dimension, rather than just a strictly rational one. The latter source is mostly inspired by neuro-physiological research that suggests a decisive contribution of emotions to human behavior. Some of these studies even defend that emotions are a *sine qua non* condition to rational behavior. This thesis is advocated by Antonio Damásio [Damásio, 1994].

These two origins led to distinct paths of research. More recently it has been suggested that these two sub-fields should not be pursued separately, but rather that one has to learn from the other[3]. The argument is that to attain believability in emotions expression, it is not sufficient to simulate emotions, for instance, by the means of a set of rules that mimic certain behaviors. On the contrary, if believability of emotion expressions is desired, it is essential to base the agent construction on a sound theory of emotions. Moreover, it is interesting to consider emotions expression as a "dashboard" of an agent internal state[4]. In other words, consider an agent internal states of which correspond (to a degree) to emotional states, and that those states are visible from the outside (through the means of an expression mechanism): an external observer (e.g., other agent) equipped with the ability of recognizing those emotional states, could make use of those observations to reason about the agent internal states and possibly about its future behaviors. Emotional expression and recognition therefore can play an important role in social contexts.

2 Related work

Antonio Damásio has been an influential reference for several researchers [Damásio, 1994; 1999] in the area of emotion-based agents, namely because the essential role he attributes to emotion as far as rationality is concerned. Related work having Damásio as a fundamental reference was developed, for instance, by Juan

[1] Personal communication at AAAI Fall Symposium 2001.

[2] *idem.*

[3] *idem.*

[4] Personal communication of Barry Kort at AAAI Fall Symposium 2001.

Velasquez [Velásquez, 1999], which uses the Damásio's idea of somatic marking along with a society of agents approach [Minsky, 1988]. Another approach based on Damásio's is the one of Gadanho [Gadanho and Hallam, 1998], which complements a reinforcement learning architecture with an hormonal system. Several implementations followed another influential reference — the Appraisal Theory of Frijda [Frijda, 1986] — for instance, the TABASCO architecture [Staller and Petta, 1998]. Aaron Sloman has been one of the precursors of emotions in AI, defending an architecture based on a reactive, deliberative and meta-management layers [Sloman, 1999]. More formal approaches to this field have been explored for instance by Arzi-Gonczarowski [Arzi-Gonczarowski, 2000], using categorical theory, and by Gmytrasiewicz [Gmytrasiewicz and Lisetti, 2000], using decision theory.

3 The Proposed Model

This paper explains a model of a situated agent: its sensors receive stimuli from the environment, to which the agent responds with actions upon the environment.

In 1998 our group proposed a prescriptive model of emotion-based agents [Ventura and Pinto-Ferreira, 1998]. This model is mainly inspired on the somatic marker hypothesis of Damásio. Since then, our group has been maturing this model and performing experimentation on it [Ventura and Pinto-Ferreira, 1999; Maçãs *et al.*, 2001; Sadio *et al.*, 2001; Ventura *et al.*, 2001]. These experimentations include the exploration of a labyrinth by a robot and the supervision of the controller of an inverted pendulum, among others. Some interesting results have been obtained from these experiments.

What we call an emotion-based agent architecture can be explained in terms of successive levels of competence:

3.1 First level of competence

The first level of competence is accomplished by a simple sensorimotor map between sensors and actuators. This level is designated **perceptual level.** This name derives mostly from the fact that this level performs an interpretation of stimuli in terms of actuation response. The sensorimotor map of this level may either elicit a certain behavior for some stimuli, or may not respond at all to some other stimuli. The reader can think of this level as a genetically evolved part of an agent which assures its survivability in a given environment niche. Some properties shown by an agent with such level implemented are: simplicity, fast response, and robustness. This level is simple and fast because it can be implemented with a simple sensorimotor map (e.g., lookup table, neural network, etc.). The agent design encodes how shall the agent react to certain situations. This hard-wired encoding includes both the detection of those situations as well as the actions elicited. Robustness, here considered in the sense of coping with varying environmental conditions, without explicit world representation, is a consequence of the reactive nature of this level. These properties has been throughly discussed on the literature about reactive agents, e.g., Brooks [Brooks, 1989; 1991] is accounted for being the precursor of the idea of reactive agents which dismiss any explicit representation and reasoning about the world, and Kaelbling [Rosenschein and Kaelbling, 1995] which has further developed these concepts, including formal approaches to them.

The agent does not know why it does what it does: it just does it. What is the *meaning* of a certain stimulus to such an agent? At this level, meaning can only be ascribed by an external observer. Such an observer realizes that whenever the agent is exposed to certain stimuli, it always responds in the same way. From the agent point of view, this constitutes an immediate, built-in response. But an external observer may associate certain environment states with agent behaviors. For instance, people (mostly from outside the field) often ascribe internal states to agent behaviors. In 1979 John McCarthy discussed the validity of ascribing mental states to machines [McCarthy, 1990]. He argued that this ascription is useful to the extend of facilitating the discussion about the systems. But there are some risks: consider for instance the tendency to describe a simple obstacle avoidance behavior as being "afraid" of walls.

Relevance is addressed at this level in a trivial way. As the perceptual map only responds to a subset of stimuli, all stimuli outside this subset are irrelevant to the agent, since they elicit no response at all.

Such a simple perceptual map shows some limitations: once the perceptual map is built-in, it encodes all the agent responses to stimuli that the agent will ever have. To build more competent agents, capable of performing a better processing of the environment stimuli, in a sense that will be clarified below, it is mandatory to embed new and more sophisticated sensors.

3.2 Second level of competence

Increased environment complexity on one hand, and the need of better competence on the other, demand, at a first glance, increased sensor diversity and richness, as well as more sophisticated processing of data. This means that the raw information provided by sensors is richer. However, when the above methodology is attempted, the dimensionality increment of sensor data makes the design of sensory maps an intractable task. Moreover, such built-in maps may show difficulties in accommodating environment changes. In order to preserve performance, the sensory maps would have to incorporate all that environment variety. For this reason, a second level of competence should be introduced.

Let us consider an alternative approach other than incorporating additional complexity at the perceptual level. This approach consists of adding a second level of processing, that has to process stimuli simultaneously, in parallel, with the perceptual level. The parallel nature of these levels is essential so that the level of competence accomplished by the perceptual level alone is not compromised. We call this second level the **cognitive level,** because it is based on a rich and complex representation of stimuli. We readily acknowledge that the names "cognitive" and "perceptual" fail to capture the full nature of each level described. The model discussed in this paper would be the same if we replaced them by any other pair of names, e.g., "high" and "low" levels, "complex" and "simple" levels, "first" and "second" levels, etc.

While the perceptual level manipulates simple and basic representations, the cognitive one uses complex representations of stimuli. The fundamental idea is to create a double representation of stimuli. The complex representation is called **cognitive image**, while the simple one is called **perceptual image.** The first level provides a direct map from the sensors and the actuators. We consider that this mapping is performed in two steps: a first step that extracts a representation of the stimulus of reduced dimensionality, *i.e.,* the **perceptual image,** and a second step which maps the resulting perceptual image space directly to action space. Simultaneously, a cognitive image is extracted from the stimulus. These two representations are then stored in a memory. By associating these two images, the agent establishes a one-to-one link between a rich representation and a basic one. When shall the agent associate and store these pairs of images? It should depend on a relevance assessment made by the agent, e.g., stimuli that elicit a perceptual response (a threat?), novelty, and so on.

Let us examine closely the consequences of associating a complex (cognitive image) with a simple representation (perceptual image) of the same stimulus. The purpose of storing these pairs (associations) is twofold: on one hand, the agent may use the cognitive image extracted from the stimulus to search the memory for a pair containing a similar cognitive image — we call this *matching* —, and on the other hand, the perceptual image extracted from the same stimulus may be used to guide the matching mechanism — we call this *indexing.* On one hand, the perceptual image, in such an association, ascribes a sensorimotor-based representation to a complex representation. For instance, imagine that the agent associated the shape of some object with certain features that triggered a run-away behavior (built-in). The cognitive image containing a rich representation of this shape (e.g., a bitmap) was stored together with the perceptual image representing the threat level that elicited the run-away behavior. Whenever the agent encounters a stimulus with a similar cognitive image (*i.e.,* a similar shape), this cognitive image is matched against the memory, and the previous association is recalled. Depending for instance on the degree of similarity, the agent may exhibit the same run-away behavior, even when the perceptual image of the stimulus does not trigger it by itself. Moreover, when the perceptual image is obtained prior to the cognitive one, the former can be used to guide the search for matching cognitive images. This corresponds to the *indexing* mechanism, and it is essential in order to make the search for a cognitive match computationally feasible, since the number of stored associations can become very large. We argue that this mechanism allows the agent to ascribe relevance to the stored associations, in the sense of constraining the search for cognitive matches to a subset. This subset corresponds to the associations indexed by the perceptual image extracted from the stimulus.

The idea of marking the cognitive image with a perceptual one is based on the somatic marker hypothesis developed by Antonio Damásio [Damásio, 1994; Bechara *et al.*, 1997]. Antonio Damásio argues that the brain is able to associate the image of an observed object with the body state of the observer, in such a way that when it encounters the same (or similar) object again, that body state is recalled and influences decision-making.

The operationalization of the above concepts is performed by three mechanisms: (1) the *marking* mechanism, which establishes and stores in memory associations between the cognitive and the perceptual images,(2) the *matching* mechanism, which searches the memory for a previously stored association, with the same (or similar) cognitive image, and (3) the *indexing* mechanism which guides the matching mechanism, preventing it to exhaustively search for all stored associations.

The marking mechanism creates an association between a cognitive and a perceptual image. This happens according to a built-in criteria, e.g. for certain perceptual image values. A criterion is needed in order to prevent the memory to be flooded with repeated and/or irrelevant associations. The use of these associations, stored in the agent memory, is done by the matching mechanism. This mechanism is activated each time the agent is exposed to a stimulus. The agent uses the extracted cognitive image to search the memory for associations which contain similar cognitive images. Recall that the perceptual image is obtained first, and that it can be used to guide this search. When a match is found (e.g., the one with the most similar cognitive image), the agent ascribes the associated perceptual image to the stimulus. These two perceptual images — the one extracted directly from the stimulus, and the one obtained by the matching mechanism — may give rise to different actions, from which the agent would have to choose. We have no definitive answer to this problem, which appears whenever multi-layered architectures (where some or all of the layers can output an action) are considered. A possible answer: when no sufficiently close match is found, choose the perceptual action; otherwise, let the perceptual level decide whether its action should override a cognitive one (*i.e.,* a built-in priorization, defined over the perceptual image space).

This level opens several degrees of freedom. As for the marking mechanism: when shall the agent create these associations? Possible answer: whenever the stimulus is relevant, in terms of eliciting a non-null perceptual representation. In other words, the relevance is in this case primarily assigned by the built-in perceptual map. How shall the agent prevent from storing repeated associations? How shall the agent update previous associations when faced with contradictory stimuli? (*i.e.,* the two perceptual images, one extracted from the current stimulus, and another obtained from a cognitive match) As with the matching mechanism: what is the criteria for a cognitive match? (e.g., exact match? a metric among images? until what distance two cognitive images do match?) When a match is found, how shall the agent decide between the current perceptual image and the one obtained from a cognitive match? Since the perceptual level is faster, when/how shall the agent wait for the (necessarily slower) cognitive matching mechanism? What is the impact of this wait to the agent performance? Note that the fast response is an important property of the perceptual level, from the survivability point of view.

The presence of a second level on top the first one may give origin to conflicting situations. These situations occur whenever the output of the perceptual level (action or behavior) differs, according to some metric, more or less dramatically, from the perceptual images from the matching mechanism. Different approaches to resolve this conflict can be used. This degree of freedom can be exploited in order to obtain different kinds of behavior. For instance, and agent taking excessively into account the outcome of the cognitive match (memory), may experience difficulty on discriminating fine distinctions on the environment: since it tends to consider the perceptual images from the associations in memory, the perceptual image extracted from the stimulus tends consequently to be dismissed. On the contrary, an agent taking the cognitive match too little into account, may not be exploiting appropriately the "lessons" of the past: the tendency to only take into account the perceptual image from the stimulus, prevents the agent from anticipating future consequences, that could be exploited/avoided by the use of the stored associations.

3.3 Third level competence

Planning plays an important role in complex environments. An agent may have to envision a sequence of actions in order to obtain some desired goal. Associating cognitive and perceptual representations *per se,* as in the previous competence level, does not contain any representations of time or sequence. This is the motivation for a third level of competence. This level comprises the idea of the "movie-in-the-brain" (MITB). The MITB consists of storing an association between the cognitive and perceptual images, together with the action taken, for a sequence of time instants. This forms a "movie" representing the agent's recent history of interaction with the environment. This mechanism was implemented and experimented with in a simple test-bed (control of an inverted pendulum) [Ventura *et al.*, 2001].

The contents of the MITB can be viewed as a trajectory, parameterized by time, in the space of cognitive images. Each point of this trajectory also contains the associated perceptual image and the action taken. As time passes, the cognitive images obtained from stimuli may come close to other trajectory zones. In the context of the second level of competence, this corresponds to a cognitive match. However, instead of picking up from memory a single association as in the second competence level, the MITB provides a trajectory segment. Consider for now a single point over this trajectory, such that it provides the best cognitive match. This point divides the trajectory in two parts: one corresponding to the forward flow of time, and another corresponding to the backward flow of time. The latter one can be used to confirm or refute the hypothesis that the agent is in the same situation as it was at that time: this hypothesis is more plausible as the preceeding stimuli sub-sequences are more similar. The two sub-sequences that are compared are: the sub-sequence preceeding the matching point, and the one preceeding the present instant of time, *i.e.,* most recently stored in the MITB. The former one, that is, the sub-sequence following the match, gives an indication of future consequences, when the course of action stored along the trajectory was taken. Several cognitive matches (*i.e.,* matching sub-sequences), each one with varied courses of action provide the agent with precious information about future consequences of its actions.

As with the previous level, this one also raises a myriad of open problems (some corresponding to degrees of freedom of the system). To name a few: How shall the agent behave in order to fill the memory with variety in terms of courses of action? (experimentation with the environment, leading to the old issue of exploration *versus* exploitation). How to choose among a set of alternative courses of action? How to discriminate between the consequences of agent actions and others? (the issue of establishing cause-effect relationships).

One severe limitation of the MITB is that the storage requirements of the full history of the agent may not be practical, for two reasons: bounded memory size, and increasingly computational requirements (for the matching mechanism). This suggests the need for a long-term memory mechanism. All the data stored in the MITB has somehow to be compressed. The relevant information has to be extracted.

4 Future Perspectives

There are many theories that address emotions, the majority of which adopt a descriptive point of view (e.g., [Damásio, 1999; Ortony *et al.*, 1988; Frijda, 1986]). These theories present a description of the mechanisms of emotions. Several of these theories originated from neuro-physiological studies of emotions in live beings. However, there are very few theories addressing how emotions can be implemented on an artificial machine. We claim that some of the biggest challenges this field faces are the development of good prescriptive theories. Such theories are expected to generalize the fundamental mechanisms of emotions (or a subset of them), as they are found in living beings, to a broad range of domains (specially the domain of the artificial). The search for such distilled mechanisms is one of our current research goals.

Our current efforts are oriented towards a formal framework that would provide a theoretical grounding to the model proposed in this paper. Such a framework shall state precisely the nature of the mechanisms described in this paper — the marking, matching, indexing, and "movie-in-the-brain" mechanisms — in a way that the discussed properties can be derived from the theory. The advantages of a formal approach are well known. But once a formal approach is attained, the justification of such a model does not need neuro-physiological arguments. It becomes a self-contained theory. Under this perspective, terms like "emotions" are useful just to the extent that they guide us towards the goal of a formal theory: emotions are "crutches," from this point of view.

References

[Arzi-Gonczarowski, 2000] Zippora Arzi-Gonczarowski. A categorization of autonomous action tendencies: The mathematics of emotions. In *Proceedings of the EMCSR2000, Symposium 'Autonomy Control: Lessons from the Emotional'*, volume 2, pages 683–688, university of Vienna, Austria, April 2000.

[Bechara *et al.*, 1997] Antoine Bechara, Hanna Damásio, Daniel Tranel, and Antonio R. Damásio. Decising advantageously before knowing the advantageous strategy. *Science*, 275:1293–1295, February 1997.

[Brooks, 1989] Rodney A. Brooks. A robust layered control system for a mobile robot. *IEEE Journ. of Robotics and Automation*, RA-2(1):14–23, March 1989.

[Brooks, 1991] Rodney A. Brooks. Intelligence without reason. In *Proceedings of IJCAI-91*. IJCAI, 1991.

[Damásio, 1994] Antonio R. Damásio. *Descartes' Error: Emotion, Reason and the Human Brain.* Picador, 1994.

[Damásio, 1999] Antonio R. Damásio. *The Feeling of What Happens: body and emotion in the making of consciousness.* Harcourt Brace, 1999.

[Frijda, 1986] N. H. Frijda. *The Emotions.* Cambridge University Press, Editions de la Maison des Sciences de l'Homme, Paris, 1986.

[Gadanho and Hallam, 1998] Sandra Clara Gadanho and John Hallam. Exploring the role of emotions in autonomous robot learning. In Dolores Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 84–89, 1998.

[Gmytrasiewicz and Lisetti, 2000] Piotr J. Gmytrasiewicz and Christine L. Lisetti. Using decision theory to formalize emotions for multi-agent systems. In *Second ICMAS-2000 Workshop on Game Theoretic and Decision Theoretic Agents*, Boston, 2000.

[Maçãs *et al.*, 2001] Márcia Maçãs, Rodrigo Ventura, Luis Custódio, and Carlos Pinto-Ferreira. Experiments with an emotion-based agent using the DARE architecture. In *Proceedings of the Symposium on Emotion, Cognition, and Affective Computing (AISB'01 Convention)*, UK, March 2001.

[McCarthy, 1990] John McCarthy. Ascribing mental qualities to machines. In Vladimir Lifschitz, editor, *Formalizing Common Sense: Papers by John McCarthy*, pages 93–118. Ablex Publishing Corporation, Norwood, New Jersey, 1990.

[Minsky, 1988] Marvin Minsky. *The Society of Mind.* Touchstone, 1988.

[Ortony *et al.*, 1988] A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions.* Cambridge University Press, Cambridge, UK, 1988.

[Picard, 1995] Rosalind W. Picard. Affective computing. Technical Report 321, M.I.T. Media Laboratory; Perceptual Computing Section, November 1995.

[Rosenschein and Kaelbling, 1995] Stanley J. Rosenschein and Leslie Pack Kaelbling. A situated view of representation and control. *Artificial Intelligence*, 73, 1995.

[Sadio *et al.*, 2001] Rui Sadio, Goncalo Tavares, Rodrigo Ventura, and Luis Custódio. An emotion-based agent architecture application with real robots. In *Emotional and Intelligent II: The Tangled Knot of Social Cognition*, 2001 AAAI Fall Symposium Series, pages 117–122. AAAI, 2001.

[Sloman, 1999] Aaron Sloman. Beyond shallow models of emotion. In *i3 Spring Days Workshop on Behavior planning for life-like characters and avatars*, Sitges, Spain, March 1999.

[Staller and Petta, 1998] Alexander Staller and Paolo Petta. Towards a tractable appraisal-based architecture. In Dolores Cañamero, Chisato Numaoka, and Paolo Petta, editors, *Workshop: Grounding Emotions in Adaptive Systems*, pages 56–61. SAB'98: From Animals to Animats, August 1998.

[Velásquez, 1999] Juan D. Velásquez. From affect programs to higher cognitive emotions: An emotion-based control approach. In Juan Velásquez, editor, *Workshop on Emotion-Based Agent Architectures (EBAA'99)*, pages 114–120, May 1999.

[Ventura and Pinto-Ferreira, 1998] Rodrigo Ventura and Carlos Pinto-Ferreira. Emotion-based agents. In *Proceedings AAAI-98*, page 1204. AAAI, AAAI Press and The MIT Press, 1998.

[Ventura and Pinto-Ferreira, 1999] Rodrigo Ventura and Carlos Pinto-Ferreira. Emotion-based agents: Three approaches to implementation (preliminary report). In Juan Velásquez, editor, *Workshop on Emotion-Based Agent Architectures (EBAA'99)*, pages 121–129, May 1999.

[Ventura *et al.*, 2001] Rodrigo Ventura, Luis Custódio, and Carlos Pinto-Ferreira. Learning courses of action using the "movie-in-the-brain" paradigm. In *Emotional and Intelligent II: The Tangled Knot of Social Cognition*, 2001 AAAI Fall Symposium, pages 147–152. AAAI, 2001.