Artificial Emotions
Good Bye Mr. Spock!

Rodrigo Ventura, Luís Custódio, and Carlos Pinto-Ferreira

Instituto de Sistemas e Robótica
Instituto Superior Técnico
Rua Rovisco Pais, 1
1096 Lisboa Codex, Portugal
Tel. +351-1-8418271
{yoda,lmmc,cpf}@isr.ist.utl.pt

Abstract. The question of implementing emotions in robots is twofold: on the on hand it should be verified whether such an effort is valuable, and on the other it should be determined whether the implementation is feasible. The answer to the first question seems easy: besides and beyond the reasons of pure intellectual curiosity and scientific research, emotions should be studied and implemented if the overall behavior of such robots is better than their unemotional counterparts with respect to behaving efficiently in a real world environment. Two diverse opinions have emerged in the previous discussion. One, due to McCarthy, asserts that emotions will introduce obstacles in the communication among robots and human beings [6]. On the other hand, Minsky sustains the opinion that it is impossible to implement intelligence without emotions [7].
In this paper we analyze these perspectives, discuss a possible way to approach the topic, and provide an architecture to implement emotions, which has shown some very interesting characteristics. We sustain that the research on emotions — from the Artificial Intelligence point of view — is valuable and worth pursuing.

You [humans] are, after all, essentially irrational.
...Spock, "Metamorphosis," stardate 3220.3.

1 The need for emotions

It is an uncontroversial assertion that emotions play an important role in the behavior of human beings. What is not so clear is whether emotions should be implemented in robots.

Emotions can be analyzed under two differing points of view: an *external*, behavioral, in which communication among individuals is considered to be helped by cues provided by emotion-based attitudes, and *internal*, functional, in which, following recent results of research, the mechanisms of emotion are crucial in the understanding of decision making processes. Of course, these different aspects correspond to the two sides of the very same coin. However, from a methodological perspective, researchers

interested in studying and modeling emotions should place themselves with respect to these two points of view.

When John McCarthy asserts that "robots should not be equipped with human-like emotions," [6] he is concerned with the additional complexity in understanding the behavior of a robot (the cues it provides) in a relationship with human beings — the behavioral side of the coin. On the other hand, when Marvin Minsky states that it is not possible to achieve intelligence without emotions [7], he is considering the inner workings of the human decision making mechanism, that is to say, the functional side of the coin. We assert that both perspectives rely on sound arguments; however, we sustain that, notwithstanding the difficulties and costs of the enterprise, it is worth pursuing.

It is not an easy task to convince ourselves (and others, of course), that studying emotions is valuable and useful in the framework of Artificial Intelligence. Some very entrenched prejudices and misconceptions are difficult to overcome.

At first sight, emotional behavior seems to be a regrettable heritage from our animal ancestors, always to avoid and to be ashamed of, something whose only value was to be the first step on the journey to 'rational' thinking. Now we suspect that emotional mechanisms are powerful weapons to allow quick decision making in complex environments.

The deeper the study of AI, and other cognitive sciences, the more we conclude that the objective of constructing a robot performing competently among human beings, conducts to a swamp of increasing complexity.

However, dealing with more and more complexity is the saga of AI: at the very beginning, the emulation of intelligent behavior was approached by the incorporation of sophisticated mechanisms of inference. From a certain moment on, it was found out that knowledge — and not only reasoning — was a crucial element to include in this melting pot. Then, some years later, we were told that intelligence is inseparable from perceiving and acting, which took us back to the blackboard, to the task of studying agents and building robots. Social behavior, uncertainty, and other topics were added to help us in this quest to intelligence.

However, recent research in the field of neuroscience has demonstrated without any shadow of doubt that emotions underlie the mechanism to achieve quick and adequate decisions when the situation demands urgent action [3]. When there is plenty of time to decide (that is to say, it seems that nothing very serious is going to happen in the short run), decisions tend to be based on what is called "rational" processes — involving reasoning and deduction.

Nature did not entrust the responsibility of urgent decision making to sophisticated mechanisms of reasoning. When designing and constructing robots, why should we?

Reasoning — and particularly logic — is too heavy in terms of computation to be useful in the vast majority of daily life decision making. It is true that we achieve epistemologic adequacy following a logic-based approach; however, the heuristic adequacy is lost. And, when real-time decision making is the aim, not achieving heuristic adequacy

implies losing epistemologic adequacy, in the sense that we risk to make the correct decision at the wrong time.

Following the research of Damasio [3], we hypothesize that the more urgent and serious the situation is, the less we reason and the more emotion-based the decision is. And this is not to regret or to be sorrow: it is our way to deal with complexity.

When explaining what underlay a certain course of action it is always embarrassing and uncomfortable to state that it was an emotional reaction to the circumstances: even if it was perfectly adequate to the situation, it always seems arbitrary, untaught, and irrational.

However, rationality cannot and should not be confused with optimality: as we have painfully learned in the past few decades, the goal of achieving optimal solutions is not compatible with finding answers to real-world problems. On the other hand, understanding rationality as solution adequacy — considering what the agent knows about the situation — suggests a different approach to cope with difficult, real problems. Based on the Damasio's work [3], it seems that intuition — possibly a result of the machinery of emotions exhibits these characteristics of adequacy we are searching for.

Not surprisingly, when trying to convince someone about an argument (in the technical sense of the word, a set of premises and a conclusion), somebody saying just 'I feel that such and such fact imply the conclusion,' is not being very persuasive...

In fact, one of the drawbacks of implementing emotions is that the resulting behaviors are not explainable. And explanations are crucial in teaching and convincing others about our own decisions. On the other hand, it is essential to understand how emotions work in human beings and animals to develop a framework underlying future implementations.

Understanding emotions is difficult because, as they are not derived from verbal thinking, they are very difficult to translate verbally. Following the Western Civilization traditional approach, what cannot be stated verbally, as cannot be communicated, does not deserve consideration. However, the fact of being unable to understand a topic, an idea or a concept does not necessarily means that it is not relevant...

The question is how to teach (and learn!) adequate emotional behavior. As it is based on experience, it is not possible to transmit this kind of knowledge. So, how to teach such kind of behavior to robots? Of course, the only way is to expose them to situations demanding urgent and proper action, provided that they include a mechanism to deal with emotions.

2 The proposed approach

There have been several approaches to incorporate emotions in agents. These approaches can be divided in two groups: the first one is based on a non-emotion layer, and adds emotion-like capabilities on top of that — a *behavioral* approach. For instance, in the context of the OZ project at CMU [10], the *Em* module, adds emotional behavior to

agents architectures. At the lower levels of the architecture, reactive and planning modules [1] bridges the gap between the perceive-think-react loop and these higher level components. These kind of architectures, instead of being supported by emotions, are rather enhanced by a higher level module implementing emotions. The emotion model of the *Em* module was based on a cognitive approach to human emotions due to Ortony *et al* [8].

We shall call the second a *functional* approach, and is constructed in an emotion-oriented paradigm from its foundations. One such system can be found in [13], which pursuits emotional behavior by building a society of agents (in the sense of [7]). Each agent (called "emotion proto-specialist") contributes to the outcoming emotional behavior in a particular way. These agents can be identified with basic emotions [4]. This particular system is strongly oriented towards the simulation of human emotions, all the way down to the level of hormone chemistry.

Albeit the ideas presented in this paper are based on human emotions, we detach from taking into account excessive detail (with respect to fisiological issues), preferring a more abstract level. We are rather interested in understanding how the mechanisms underlying human emotions can contribute to a more general context of *machine intelligence* [15].

Each one of the above approaches is motivated by different ways of understanding the role of emotions in human behavior. The first one derives from viewing them as an extra mechanism humans make use of. They are not viewed as an *essential* part of the workings of the human mind. On the opposite, we believe that emotions are the foundations of a rational mind. Recent trends in neuroscience [3, 5] motivate and support this belief. In fact emotions are much more than just "emotions" or what is currently described as emotional behavior. They result from a mechanism on top of which, more complex and elaborative functioning is built.

The agent architecture we hypothesize is based on a double perspective from which external stimuli are processed: a *cognitive*, elaborative — which allows the agent to understand what is happening and what it knows about the world, and a *perceptual*, immediate — which permits them to react quickly, and therefore has a simpler and more basic representation than the former. For instance, the image of a zebra can be viewed as an animal with four legs, with a striped coloring, etc. A myriad of considerations can be drawn by a careful observer from this image. But to a lion, these considerations have little importance, since the zebra's image triggers on it a predator behavior.

The architecture we propose in this paper is then based on this double perspective. External stimuli are simultaneously processed by two systems: a *cognitive processor* which extracts the cognitive features of the stimulus, and a *perceptual processor* which provides a more basic assessment of the same stimulus.

In a neuroscience context, this double layering has been discussed in several biological models of human emotions: namely the Cannon-Bard theory ([5], pg. 82–85) and the Papez circuit theory ([5], pg. 87–90).

The objects that result from the cognitive processor are complex, rich, divisible (in parts, maybe hierarchically), structured, therefore presenting difficulties in handling them. Examples of such objects are visual images, auditory time-frequency represen-

tations, etc. There is strong evidence that humans reason directly at the level of visual images ([3], pg 106). We thus shall call such images as *generalized images* (GIM).

On the other hand, objects resulting from the perceptual processor are simple, basic, indivisible, therefore easily handled, although they lack richness to represent more complex objects. We shall call such an object a *vector of desirability* — an ordered set of basic built-in characteristics, which provides a direct assessment of the stimulus under a perceptual point of view (is it positive/negative?, desirable/avoidable?, etc.).

After these two kinds of objects are obtained, there are two complementary mechanisms that act upon them. First, cognitive objects are marked by perceptual ones. For instance, the cognitively processed image of a zebra must be associated with a very basic predatorial instinct in the lion's mind. This mechanism is inspired in the "somatic marker" concept that Damasio [3] hypothesizes. This marking can be said to assign *meaning* to the corresponding cognitive object. The second mechanism indexes cognitive objects by the means of perceptual ones. This allows the agent to have quick access to a cognitive context (for reasoning purposes, for instance), given a basic, primitive stimulus. For instance, picture a big fast object moving towards the reader: your first impulse is basic, instinctive, thus based on perceptual information (such as color changes, optical flow, etc.). Only after some time (which is probably spent with diverting from the object's path), the higher, slower parts of the brain can reason about identifying the object. Nevertheless the first basic perceptual images offered a useful cue about the nature of the moving object.

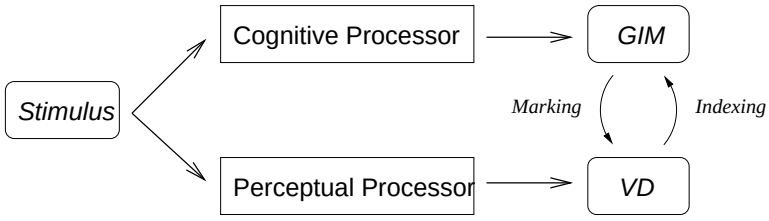

Fig. 1. Proposed architecture: the cognitive processor and the perceptual processor generate generalized images (GIM) and vectors of desirability (VD). The former is marked by the later, while the later indexes the former.

The architecture shown in figure 1 illustrates the discussed ideas. Note that the words *objects* and *images* are used interchangeably here, although they mean the same concept in this context.

3 A path to implementation

A simpler version of the above architecture was implemented. A more complete one is underway. The goal was to prove that a plain implementation of some of the ideas discussed here is capable of exhibiting an interesting behavior. A sketch of the implemented architecture can be found on figure 2.

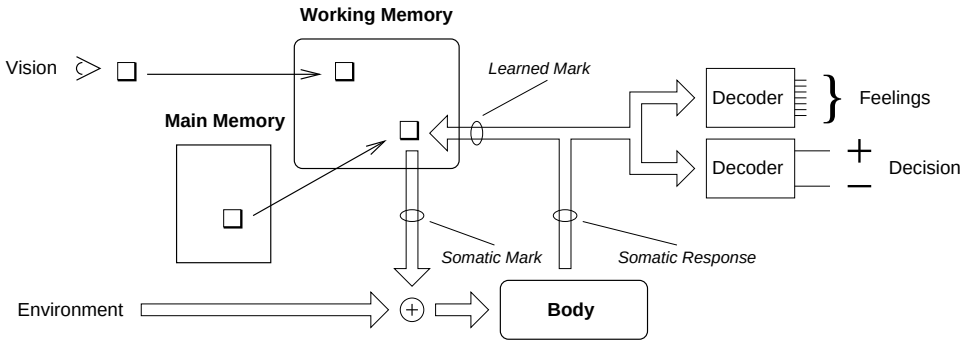

Fig. 2. Synopsis of the implemented architecture. See text for details.

This architecture differs from the one discussed in the previous section on the following points: it is assumed that the agent perceives external stimuli via two different sensors, namely *vision* (that performs the role of pure cognitive images) and the *environment* (which corresponds to the perceptual input); and only the marking mechanism is implemented (there is no indexing) [1].

The system works as follows: first, a external stimulus produces in the agent a visual image (cognitive) as well as environment input (perceptual). The visual image is used to recall from the main memory all images, through some similarity criterion. Each one of these images stored in the main memory, contains a marker[2], which has the same nature as the environment input. All similar images are recalled, along with the input image (unmarked, as it is seen from a cognitive sensor), and put in the working memory area. Each recalled image is also associated with a *relevance value*, which reflects the amount of similarity found. For each image in this area, its mark (if any, otherwise it is null) is composed with the environment input producing what we call the *somatic response* (the "body" box is just a placeholder for this response). This somatic response is used not only to update the marker at the working memory, but also forms the response of the system to the recalled image. The original mark has as much weight in the final marker value as the relevance value. The updated image is afterwards transfered to the main memory, which allows the agent's behavior (and may we say, knowledge) to be updated as time goes by, *i.e.* learning. On the other hand, the system output has two parts: first it forms what the agent is "feeling" about the recalled image, and second it gives a positive or negative (or neutral) response to the same image. Of course the use of such strong terms such as "agent's feelings" is not free from controversy. We are not claiming that humans "feel" in the same way, only that we call this output "feeling" due to the way this architecture was inspired in human feeling mechanisms.

To assess the behavior of an agent based on the described architecture, several image-environment input pairs were presented to the system. When an image similar to a previously perceived one is presented without any environment input, the similar im-

[1] The terminology is still extensively taken from Damasio's [3] book.

[2] We call it "somatic marker", although is used here in a much simpler sense than in the original bibliography [3].

age is vividly (*i.e.* high relevance value) recalled, along with its marking. It can be said that the pure cognitive image was perceived and "interpreted" accordingly to previous experience, showing clear learning abilities. However, if the pure cognitive image keeps on being fed to the architecture, the previous mark fades out. The same can be said about changing environment input to the same image. This means that the architecture adapts itself to changes in the environment.

4 Conclusions and future work

The purpose of this paper has two aspects: to support the idea that, since the mechanisms of emotion play a fundamental role in human rationality, machine intelligence should also incorporate such mechanisms; and second, to propose that such functionality can be achieved with a double-processing paradigm — a *cognitive* and a *perceptual* flow of information, and a mechanism that binds these two representations together.

It can be argued that machine intelligence must follow a distinct path of development than the human, since it is based on different grounds (computers are mostly serial processing devices with strong efficiency concerns, while the human brain is massively parallel with substantial redundancy). However, it is known that when the human brain performs search (for the purpose of decision making), and certain cortex zones entangled with emotion become damaged (namely the frontal lobes), the subject becomes unable to decide appropriately (e.g. the case of Elliot described in Damasio's book, chap. 3, [3]). This resembles the behavior of traditional search algorithms facing complexity. We wonder if the solutions found by mother nature to cope with this problem can be applied in a machine intelligence context. We believe they can, and support our belief throughout the current research.

Finally, some experiments with an architecture based on these ideas were described. The results indicate that the underlying assumptions originate interesting behavior.

Albeit the relative sophistication achieved in several AI fields, we are still too far from a human level of competence. A more integrated and efficiency-aware framework is needed. We claim that the path leading to the next qualitative step in terms of competence will require emotions — *artificial emotions*, a new born field of Artificial Intelligence [1, 10, 7, 9, 11–14, 2].

However, we are afraid that, to reach intelligent behavior, we will end up implementing some of the (apparently) most stupid things human beings exhibit, including bad temper...

References

1. Joseph Bates, A. Bryan Loyall, and W. Scott Reilly. An architecture for action, emotion, and social behavior. In *Proceedings of the Fourth European Workshop on Modeling Autonomous Agents in a Multi-Agent World*, Decentralized AI Series. Elsevier/North Holland, July 1992.
2. Dolores Cañamero. Modelling motivations and emotions as a basis for intelligent behavior. In *Proceedings of Agents'97*. ACM, 1997.
3. Antonio R. Damasio. *Descartes' Error: Emotion, Reason and the Human Brain*. Picador, 1994.
4. Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3/4):169–200, 1992.
5. Joseph LeDoux. *The Emotional Brain*. Simon & Schuster, 1996.
6. John McCarthy. Making robots conscious of their mental states. URL: http://www-formal.stanford.edu/jmc/consciousness-submit/consciousness-submit.html, 1995.
7. Marvin Minsky. *The Society of Mind*. Touchstone, 1988.
8. A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, UK, 1988.
9. Rosalind W. Picard. Affective computing. Technical Report 321, M.I.T. Media Laboratory; Perceptual Computing Section, November 1995.
10. W. Scott Reilly and Joseph Bates. Building emotional agents. Technical Report CMU-CS-92-143, CMU, School of Computer Science, Carnegie Mellon University, May 1992.
11. Aaron Sloman and Monica Croucher. Why robots will have emotions. In *Proceedings IJCAI 1981*, June 1981.
12. Aaron Sloman and Riccardo Poli. *intelligent Agents*, volume II, chapter SIM_AGENT: A toolkit for exploring agent designs, pages 392–407. Springer-Verlag, 1995. (ATAL-95).
13. Juan D. Velásquez. Modeling emotions and other motivations in synthetic agents. In *Proceedings AAAI-97*, pages 10–15. AAAI, AAAI Press and The MIT Press, 1997.
14. Rodrigo M. M. Ventura and Carlos A. Pinto-Ferreira. Experiments on emotional systems. Technical report, Instituto de Sistemas e Robótica, Instituto Superor Técnico, Lisbon, Portugal, 1997.
15. Wai Kiang Yeap. "Emperor AI, Where Is Your New Mind?". *AI magazine*, 18(4):137–144, Winter 1997.