EMOTION-BASED AGENTS

Rodrigo Martins de Matos Ventura
(Licenciado)

Dissertação para obtenção do Grau de Mestre em
Engenharia Electrotécnica e de Computadores

Orientador Científico:
Doutor Carlos Alberto Pinto-Ferreira

Constituição do Júri:
Doutor Carlos Alberto Pinto-Ferreira
Doutor Luís Manuel Marques Custódio
Doutor Paolo Petta

Lisboa, 16 de Março de 2000

RESUMO

Resultados recentes da neurofisiologia têm mostrado alguns aspectos interessantes da inteligência humana: os processos mentais do pensamento são guiados pelas emoções. Inclusive o pensamento racional requere emoções para funcionar apropriadamente. Esta tese propõe um modelo para um agente cujo funcionamento baseia-se em emoções. Este modelo é suportado pelo trabalho de Antonio Damasio [18] em pôr a descoberto o papel das emoções na racionalidade humana. O modelo proposto é baseado numa paradigma de dupla representação: uma representação complexa, não-tratada, estruturada denominada de *imagem cognitiva*, e uma representação simples, básica, *built-in* denominada de *imagem perceptual*. Após a discussão do modelo, três implementações são descritas, tal como alguns resultados experimentais. Finalmente, algumas consequências da abordagem são discutidas, tais como a emergência de *relevância* e *significado*, terminando com uma enumeração de possíveis futuras direcções de investigação, nomeadamente a integração deste modelo num ambiente robótico.

Palavras-chave: Emoções, Agentes, Arquitecturas, Inteligência Artificial, Neurociência, Sistemas.

ABSTRACT

Recent neurophysiologic findings have uncovered some interesting aspects of human intelligence: the mind's thought processes are driven by emotions. Even rational thinking does require emotion to function properly. This thesis proposes a model for an agent whose functioning is based on emotion. This model is supported by the work of Antonio Damasio [18] on unveiling the role of emotion in human rationality. The proposed model is based on a double-representation paradigm: a complex, unfiltered, structured representation termed *cognitive image,* and a simple, basic, built-in one termed *perceptual image.* After the discussion of the model, three implementations are described, as well as some experimental results. Finally, some consequences of the approach are discussed, such as the emergence of *relevance* and *meaning,* ending with the enumeration of possible future research directions, namely the integration of the model in a robotic environment.

Key Words: Emotions, Agents, Architectures, Artificial Intelligence, Neuroscience, Systems.

Acknowledgments

The author wishes to acknowledge his thesis supervisor Prof. Carlos Pinto-Ferreira, for his warm friendship and enlightening scientific guidance. A special thanks to the ISR (Institute for Systems and Robotics) and all its members, for their partnership, scientific excellence, thus providing ideal "environmental" conditions to make ideas grow and flourish.

Contents

List of Figures

Chapter 1

Introduction

Summary

This thesis begins with some considerations prior to the presentation of the developed work. First, the motivation of this work is presented. Then the objectives of this thesis objectives are enumerated. Finally, some representative work in this research area is briefly summarized.

1.1 Motivation

Through the reading of this thesis, its ideological epicenter can be clearly identified: the ideas presented in Damasio's seminal book "Descartes' Error" [18]. What he proposes is that *rationality* cannot be understood separately from *emotion.*

Since the Greek philosophers the phenomenon of reason has been divided from emotion. Scientific knowledge has been described in rational terms, logically sound, cleared of any emotional consideration. And therefore, it seemed natural that emotions were a regretful heritage humans shared with their ancestors. This suggested an assumption that has (almost) always been present when attempting to build intelligent machines: they require no more than pure rationality in order to "think." In other words, there is no sense in taking emotions into account when designing intelligent machines. It is important to stress that this is an empirical assumption, supported by the observation (mostly introspection) that humans reason rationally without any emotional feeling.

But the contribution of Damasio's work is precisely to challenge that assumption. As far as humans are concerned, even rational thought *does* involve emotions. And he was able to find neurophysiological evidence that

supports his thesis. And one can argue that no one is closer to understanding intelligence than the ones that study how the human brain works.

However, it should be noticed that this model developed by Damasio is *descriptive, i.e.,* it is supposed to provide an explanation of how the human mind works. There is still a step to be taken when one considers to implement those ideas. In other words, a *prescriptive* model is required. A possible step to bridge this gap is what this thesis proposes. The reader is invited through the following pages to assess on what degree that endeavor was accomplished.

1.2 Objectives

This thesis proposes the accomplishment of two objectives: first, to present a prescriptive model based on neurophysiological grounds of the emotion machinery in the brain, and second, to implement the model and to do some experimentation.

But prior to the presentation of the model, a set of findings from neuroscience that were taken into account is gathered in chapter 2. No prior knowledge of neuroscience is required to understand this chapter.

Then, the conceptual issues of the model are presented and discussed in chapter 3. Chapter 4 presents three implementations of the model, along with some experimental results.

This thesis ends with a chapter discussing some of the consequences of this approach, and some future direction that this research can take.

1.3 Overview

There is no agreement on a methodological foundation for building intelligent machines. In the antipodes of the broad spectrum of possibilities, lie the logic approach proposed by McCarthy [36], and the robotic insects approach from Brooks [8, 9]. The former is based on a logical approach — failing to cope with the complexity of the real world — whereas the latter, detached from reasoning models, lacks the ability of handling more difficult tasks.

The first question that pops into mind, when implementing emotions in machines, is whether or not it is legitimate to ascribe emotions to them — "is this machine *feeling*?" In a broader sense, John McCarthy has discussed the problem of ascribing mental qualities to machines:

> To ascribe certain *beliefs, knowledge, free will, intentions, consciousness, abilities* or *wants* to a machine or computer program is legitimate when such an ascription expresses the same information about the machine that it expresses about a person. ([37])

Although McCarthy was almost surely not thinking about emotions and feelings when he wrote this, an attempt to apply this concept to emotions looks interesting. But while the mental qualities referred by McCarthy can be identified with a purely rational perspective of the mind, the same cannot be said about emotions. According to Damasio [18], emotions involve the body, a physical part of a person, as it will be discussed in chapter 2.

From a philosophical foundations viewpoint, AI is divided in the usefulness of emotions in machines. On the one hand, John McCarthy sustains that "Robots Should Not be Equipped with Human-like Emotions" [38], defending the idea that rational thought can be detached from emotions, and that emotions only disturb pure rationality. However neuroscience contradicts this detachment [18]. On the other hand, Aaron Sloman [52] and Marvin Minsky [40] are quite confident that, besides being useful, emotions will be essential, at least as far as an human-like intelligent machine is pursued. Quoting Minsky from his seminal book "Society of Mind":

> The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without emotions. ([40])

In a different perspective, Sloman sustains that emotions are essential to intelligent robots, arguing its close relationship to the origin of motivations [52].

Applying emotions in artificial intelligence does not imply a unique path. The new-born field has already branched since its very beginning. A first major division can be established between *external emotions* and *internal emotions*. In other words, does one want to relate to computers on an emotional basis, or to enable the machine to use emotions internally? Of course these two perspectives are not mutually exclusive, but usually one of them is emphasized. In the former case, the central question is "how can a machine express emotions?" and "how can a machine detect an emotion expressed by a human [1]?" While in the latter, the question is "how can emotions contribute to the decision making process?".

According to Rosalind Picard, emotions can play an essential role in the way people deal with computers. They define a line of research she calls "affective computing" [46, 45] — "computing that relates to, arises from, or deliberately influences emotions." For instance, facial expressions are a medium through which emotions are expressed between people. One aspect of this research area is to detect human facial expressions, as well as how to synthesize a facial expression to show a given emotional state. The applications of this scientific area are immense: they can drastically change the way

[1] Or by another machine.

people relate with computers. If people got personally caught by conversations with the ELIZA program, imagine when computers start detecting and expressing emotions in a convincing way.

In 1988, Andrew Ortony *et. al.* published the book "The Cognitive Structure of Emotions" [42], which presents a systematic categorization of emotions. Based on this work, the Oz Project on believable agents, integrated a module (Em) implementing emotions [47] in one of their agent architectures [2]. This module provides a representation of the agent's emotional state which conditions the agent's behavior.

Another publication worth reporting is Ian Paul Wright's PhD thesis on emotional agents [66]. This work provides some interesting perspectives on the implementation of emotions in agents. Its foundations are based on reinforcement learning and an economic view of the society of mind principle [40].

The field has been more or less lethargic, with sparse publications, until a SAB-98 workshop [12], and a 1998 AAAI Fall Symposium session [11] events, both centered on emotions, putting together a large number of papers and approaches. The publication of the Picard's book "Affective Computing" [45] has certainly contributed to the attention shift onto the field of several AI researchers. At the present stage, there is little convergence on the approach to be taken. Almost every paper proposes a different approach. But it can be expected that in the future the field will decide on smaller number of approaches, resulting from the failure of some to further development, and possibly the merger of others.

Inside the emotions (in AI) field, several sub-areas of research can be identified.

Regarding what was termed above as external emotions, there is research on interaction with a robotic face, responding with an "emotional" expression to certain visual stimuli, like waving objects in front of it [22, 23]. Another example is the interaction with a software GUI, using Bayesian networks to model the user (emotional) personality [7]. In a more specific context, the recognition of affective states [65] and the expression of emotions using motion, through gestual primitives [13], are also interesting.

In an internal emotions approach, several perspectives can be identified. The architectural one views emotions as a fundamental component in a broader architecture, such as the already cited *Em* module (Oz Project) [47], the TABASCO layered architecture [53], or a rule-based approach for controlling the agent behavior of Botelho *et al.* [6]. These approaches make use of the appraisal theory (see [24, 50] for further information), which is based on cognitive assessments of situations. The appraisal has strongly influenced the field (and still does). But in the author's opinion, it fails to capture some relevant neurophysiological aspects of emotions (e.g., the nature of Damasio's

somatic marker [18]).

There are further attempts to build models of emotions from its very foundations. A reinforcement learning approach is taken by the work of Gadanho [26, 25]. Taking the agent society paradigm from Minsky [40] as a starting point, Velásquez has reached some interesting results [56, 57, 55, 58]. This thesis is viewed by the author as belonging to this perspective [62, 60, 63, 61]. Interestingly, all these researchers share in common the inspiration from Damasio's work [18].

Some tentative formalization of emotions have also been attempted, taking a more abstract mathematical approach in some cases [1, 10, 35], or a neurophysiological one in [16, 17]. These contributions are interesting, but in the author's opinion, while the former lack neurophysiological grounding, the latter have a strong descriptive content, rather than a prescriptive one.

> The robot had no feelings, only positronic surges that mimicked those feelings. (And perhaps human beings had no feelings, only neuronic surges that were interpreted as feelings.)
>
> Isaac Asimov, "The Robots of Dawn"
> (Harper-Collins, 1994)

> The robot is going to lose. Not by much. But when the final score is tallied, flesh and blood is going to beat the damn monster.
>
> Adam Smith

Chapter 2

Foundations

Summary

This chapter opens with a section presenting the foundations of this thesis. These foundations are mostly biological, and come from experimental data. Some fundamental data about the brain is presented. Several structures of the brain are discussed, stressing the concept of the topological map. Then, a brief overview of models of the brain structures directly involved with emotions is presented. From these models, the double processing paradigm is extracted, which underlies the proposed model. Finally, the relationship of emotions with rationality is discussed, reaching the supporting pillar of this work — Antonio Damasio's findings of how human rationality depends on emotions to work properly [18].

The main inspiration and motivation for the ideas presented in this thesis lie primarily on findings from biology (namely neuroscience, neurobiology and related fields). This section describes some of these findings, that underpin many choices taken during the development of the presented model.

It is not free from controversy that the construction of artificial intelligence models has to take into consideration the way nature implements intelligence in biological beings. It has been advocated by some parties that, as the human brain works in a distinctly different way than the machines to which we are targeting our models, AI research should be detached from any biological inspiration. Moreover, as the sole evidence of verbal, sophisticated intelligence comes from human beings no one really has a complete model of it (*i.e.,* reverse engineering). To derive a model of an intelligent machine, regardless of the inner workings of the instances humans are, can be expected to be extremely difficult, to say the least.

This debate is by itself discardable, since it does not seem fruitful. This

thesis is about a way of implementing emotions on artificial machines. There is yet no clear case of emotions outside the biological sphere. Therefore, an independent approach to artificial emotions has to be based on the way nature "implements" emotions.

In the following sections, it will be presented a set of findings from neuroscience (and biology in general) that are behind the model proposed in this thesis. The starting point will be a brief description of the overall brain organization in terms of microstructures such as neurons and synapses, upwards toward major brain zones. Next, a brief overview of some biological emotion models are presented, and finally — the cornerstone of this thesis — Antonio Damasio's work on the relationship between emotions and rationality.

Much of the data referred in the following sections was taken from [14], unless otherwise noted. This book constitutes an excellent overview to the broad spectrum of neuroscience issues.

2.1 Brain Organization

Such a complex mechanism as the human brain cannot be explained, not even studied, as a whole at once. Some kind of "divide and conquer" principle has to be applied to separate more or less interconnected areas of research. The approach taken here was to divide it in terms of levels of granularity of the structures involved. These levels are presented in figure 2.1.

The smallest unit — atomic element — which may still be identified with the brain, is the brain cell, *i.e.,* the *neuron.* The human brain is made out of approximately 10^{12} neurons. These neurons are connected to other neurons by *synapses*, which count up to the order of magnitude of 10^{15}. The information exchange between the neurons is electrical in nature, making use of complex chemical mechanisms (yet to be fully understood). The synapses connect unidirectionally neurons outputs to inputs of others, conditioning (among many other factors) the way the activation of the former affects the latter.

Although the inputs of one neuron are analog signals, its output is digital, forming a firing pattern. A spike in this firing patterns lasts about 1 msec, and the transmission delay up to another neuron takes about 5 msec.

In terms of density, there are about 10^5 neurons and 10^9 synapses per cubic millimeter. Each of these neurons is connected to approximately 3% of the neurons in the same amount of surrounding volume. However, the majority of the synapses of a single neuron are connected to other neurons far from the neighborhood, forming what are called projections.

Despite an apparent randomness in the neuron interconnections (which

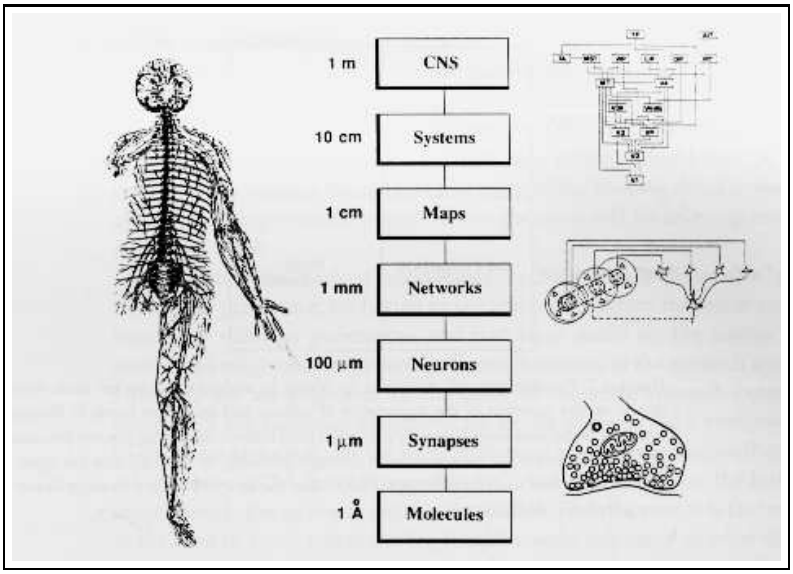

Figure 2.1: Levels of the brain in terms of granularity, from molecules up to the central nervous system (CNS). (From [14] page 11, reprint by courtesy of the author.)

seems to exist, since there is no way the genetic coding could hold enough information to determine every connection), some structure can be noticed. One kind of these is the *topographic map*. A topographic map is a zone of the brain where the placement of single neurons with respect to others is topographically organized. The most significant example of such a map, is the projection of the retina into V1 — an area in the back of the head (occipital lobe, see figure 2.2) which forms the primary visual cortices. Receptive units which are close in the retina, are projected into close neurons into V1. This way, the pattern of activation in V1 resembles the image seen by the eyes. This mapping does not preserve proportion, as it is severely distorted. This distortion can be interpreted as some areas having higher resolution than others. Figure 2.3 shows how a sample picture activates these early visual cortices, in an experimental setup.

Topographic maps can be found associated with nearly every sensory system, namely the auditory and the tactile systems, as well as in motor cortices. There is evidence that the topographic map is a device frequently used by the brain, not only in these most visible examples, but also in less evident and more abstract levels.

Interestingly enough, the early visual cortices, not only hold topographically mapped images from the retina, but also hold images recalled from memory. As Damasio notes ([18], page 101), "Preliminary studies of visual

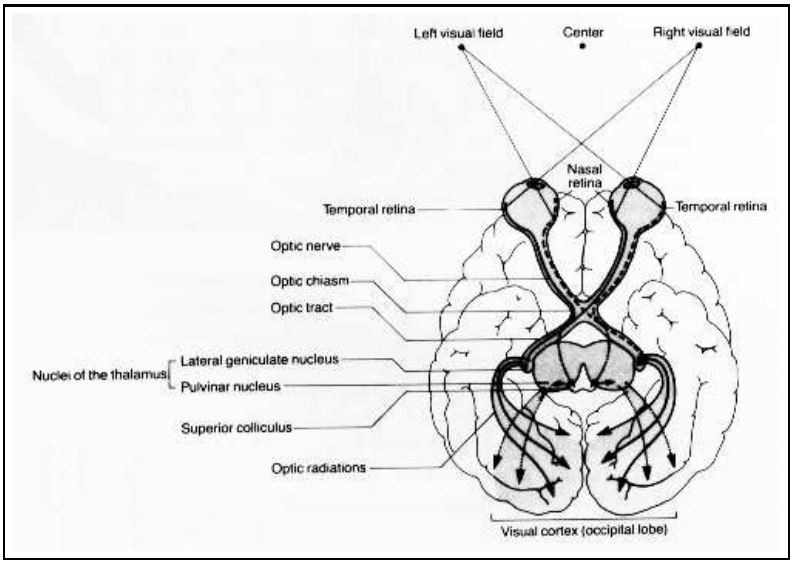

Figure 2.2: Identification of the zones where the optic nerves connect to the brain, by the means of topographic maps. Note the connections from the retina to the nuclei of the thalamus (see below the role of the thalamus in the Papez circuit model), relaying the visual map to the hypothalamus (part of the limbic system, the emotion center), and to the visual cortex, in the back of the brain. (From [14] page 151, reprint by courtesy of the author.)

recall using positron emission tomography (PET)," have shown that "the recollection of visual images activates the early visual cortices, among other areas." These recalled images are not sparse phenomena, but rather something that seems to underly the whole process of thinking. Damasio devotes a section to this fact, with the suggestive title "Thought is made largely of images":

> It is often said that thought is made of much more than just images, that it is made also of words and nonimages abstract symbols. Surely nobody will deny that thought includes words and arbitrary symbols. But what the statement misses is the fact that both words and arbitrary symbols are based on topographically organized representations and can become images. Most of the words we use in our inner speech, before speaking or writing a sentence, exist as auditory or visual images in our consciousness. If they did not become images, however fleetingly, they would not be anything we could know. ([18], page 106)

And many instances of this phenomenon are experienced by the reader (although introspection is a dangerously misleading tool): when a sentence is spoken by someone, out of our attention, it can be later "re-heard" in the

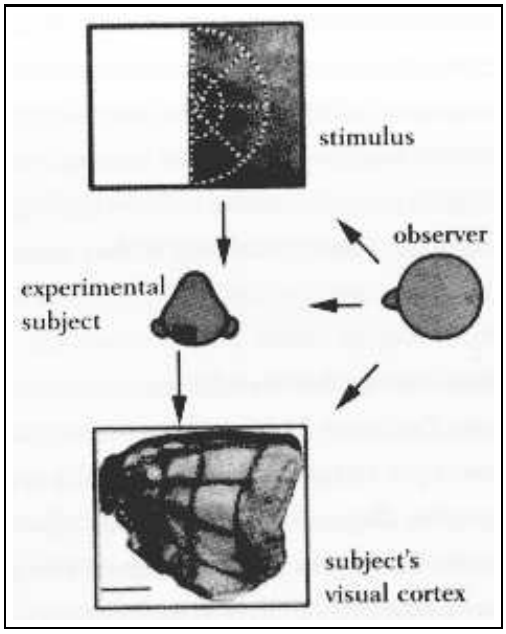

Figure 2.3: At the bottom it is shown the shape of the neural activity pattern, at the early visual cortices of an animal, that is looking to the picture shown at the top. Although deformed, the neural activity pattern shows that the topographic characteristics of the stimulus are preserved. (From [18] page 104, reprint by courtesy of the author.)

brain and only then understood; the pictorial nature of the mathematic notation, that is much easier to manipulate than some horizontal non-intuitive representation (e.g., in a LISP expression); arithmetic calculation make extensive use of graphic disposition of numerals; European traffic signs are mainly based on schematic shapes (they are supposed to be sighted and understood swiftly and clearly, and do not require the knowledge of a specific written language); primitive writing is based on icons rather than on abstract symbols[1]; the easy memorization of corporate wordless logos; and more examples can easily be found in everyday life.

This suggests that the way the brain represents and manipulates knowledge is primarily pictorial in nature, rather than symbolic. This is a rather astonishing finding, which has not received the deserved attention within AI mainstream (but is has been actively researched as a small subfield of AI, under the name "diagrammatic reasoning", see for instance [27]). But in psychology it is well studied for many years. In [32] for instance, "images"

[1] In the sense that icons represented objects and persons in scenes, and symbols implied a syntactic and semantic structure.

are defined as:

> Any thought representation that has a sensory quality we call an image. Images can involve the senses of seeing, hearing, smell, taste, touch, and movement; but since my focus is on visual images, I use the word "image" for mental contents that have a *visual* sensory quality (unless otherwise indicated). ([32], page 3)

In the course of this thesis, the term *image* is used in this broad sense, of a pictorial representation, as the one that can be found in topographic maps in the human brain.

Piaget makes reference to a set of abilities that children show, long before being able to verbalize words, called *sensorimotor intelligence* [44]. These abilities are, for instance, reaching objects with hands, manipulating objects, spatial understanding, and learning in the process. It seems clear that the processes involved in the brain deal with the world in terms of topographic maps. And these abilities appear before spoken language.

As we go up in the level of organization of the brain, the major top level brain zones can be found. The idea of classifying of the brain in zones came about with the advent of "phrenology" in the eighteenth century. The phrenologists used to classify bumps in certain areas of the head as indicators of specific abilities (such as sensing, feeling, speech, memory, intelligence, and so on). These ideas inspired the search for the location of functional centers on the brain. Nowadays there exists a much more refined map of the brain zones, with strong scientific foundations, rather than on empirical methods. Many of these results came from the study of the effects of certain brain lesions in the patient behavior. Neurologists usually are able to pinpoint the region of a lesion just by the means of the study of the way the patient becomes impaired. Located brains centers are identified with capabilities such as vision, language comprehension, touch, voluntary movement, reasoning, speech, memory, hearing, and so on (see figure 2.4). As brain lesion reports are collected and analyzed, and with the aid of apparatus able to trace brain activity (e.g., PET[2]), this map of the brain has been refined. For instance, Hanna Damasio has recently reported that memories associated with person names, tools and utensils names, and animal names, have distinct brain locations [19]. She was able to obtain this result by the means of the systematic and comparative study of patients impaired with very well located brain lesions in the memory region.

[2]Positron Emission Tomography.

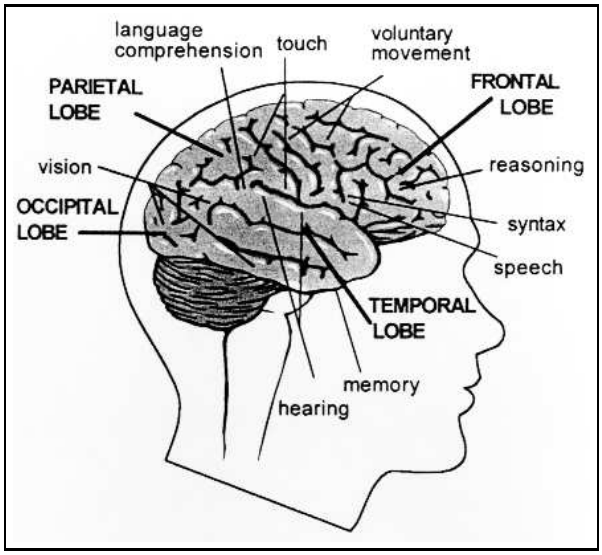

Figure 2.4: Location of several brain centers. (From [34] page 77, reprint by courtesy of the author.)

2.2 Emotion Circuitry

Since Aristotle emotions have been considered a spurious phenomenon that stubbornly stands between mind and body. In the field of AI, it has always seemed obvious that emotions and feelings[3] have nothing to do with intelligence and the domain of pure reason. No proof of any formal theorem has ever required emotions to stand valid. Scientific knowledge has never needed emotions to support itself (in the sense of exact sciences). But what remains arguable is that, because of these facts, it should be possible to attain human-like machine intelligence without ever considering the role of emotional mechanisms in humans. The idea of artificial intelligence without emotions seems to originate from the introspective idea that one person can endorse a rational (and then intelligent) line of thought, without the intervention of emotions. As we will show in this section, this is not so. At least in humans, any rational thought uses the human emotional circuitry intensively.

In 1884, William James was the first to attempt to model emotional processes in human beings[4]. Until then, it was well established that an emotional reaction (such as faster heart beat rate, sweaty hands, and so on)

[3]For the time being, the terms "emotions" and "feelings" are taken by their commonsense meaning. This chapter will not provide an exact definition, but describe approaches to understand and distinguish these concepts.

[4]The description of models here described can be found in [34], unless otherwise noted.

to an external stimulus, came from a mental assessment of that stimulus. The proposal by James went the other way round:

> Our natural way of thinking about [...] emotions is that the mental perceptions of some fact excites the mental affection called emotion, and that this latter state of mind gives rise to the bodily expression. My thesis on the contrary is that the bodily changes follow directly the PERCEPTION of the exciting facts and that our feeling of the same changes as they occur IS the emotion. *[Original emphasis]* (cited in [34])

Notice that the word "emotion" is used in this thesis in a slightly different sense than in the above quotation. William James used the word "emotion" to name the act of internally perceiving the emotional response by the brain.

Essentially, what William James proposed was a radical statement that contradicted everything that had been said in the subject before. But there is much more to follow. As emotion models are developed and refined, a clearer picture becomes visible. It is important to stress the fact that the methodology to develop these models is experimental. They are not purely philosophical models — the rough tool of introspection is very much ruled out[5]

Another relevant model is Papez' circuit theory [34], proposed in 1937 (figure 2.5). The components of this model can be directly identified with areas in the brain, but to the present discussion, their names are irrelevant. However it is important to understand that its architecture is grounded on actual brain structures.

According to this model, following the path taken by an external stimulus, the perception layer is projected into a center (the thalamus) from where it bifurcates in two separate paths. One of them follows to the hypothalamus that is able to directly generate a bodily response (affecting blood pressure, stress hormones, provoking a freeze reaction, and so on). This path that goes from perception to action is called *stream of feeling*. The response to a stimulus through this path is very quick, but it is unable to discriminate subtle differences. A second path goes from the thalamus, up through the sensory cortex, until reaching the cingulate cortex — the *stream of thought*. This latter path corresponds to higher cognitive abilities, such as reasoning, memories, and so on. The processing at this level is considerably slower than the former. The terminal centers of these two paths are connected in both directions, via the hippocampus (downwards) and the anterior thalamus (upwards). The upward connection relates to the feeling of an emotion, and the

[5]This does not mean that philosophy does not take into account these biological findings. In fact, philosophy has already taken emotions into account, for instance in [20].

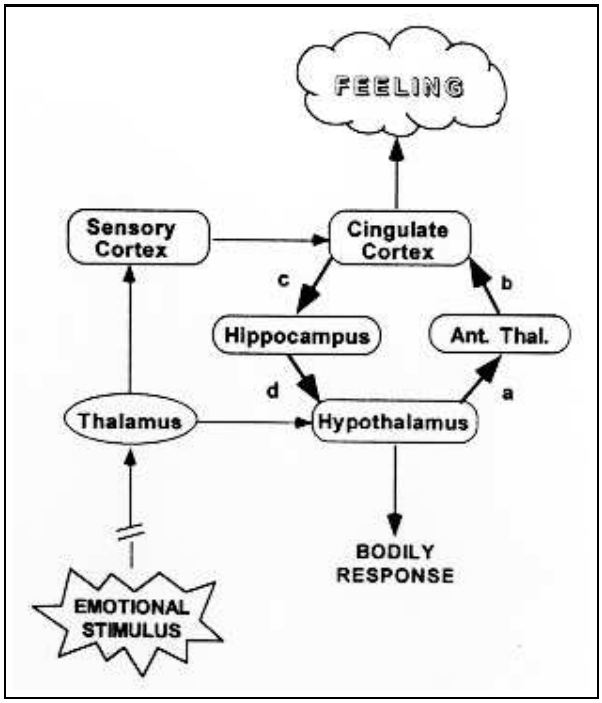

Figure 2.5: Architecture of the Papez circuit. (From [34] page 89, reprint by courtesy of the author.)

downward to the blocking of basic responses (triggered by the hypothalamus) by the means of the higher cortex [34].

From this model two aspects should be retained: first, the statement that in the human brain external stimuli are subject to a double processing, a basic/quick and a complex/slow, and second, the bidirectional influence of these two layers.

It is relevant to add a note about biological evolution, with respect to this double processing perspective. The size of the brain of mammals has been increasing along species evolution. The interesting aspect is that the brain size does not increase uniformly. What happens is that the limbic lobes (responsible for emotional behavior) remain relatively similar, while the cortex undergoes a significant growth. The growth of the cortex is the most distinguishing feature, when one observes the recent evolution of the brain. Figure 2.6 shows the volume occupied by the limbic system in relation to the cortex, in three animals. The limbic lobes, which form the older (and inner) parts of the brain, are a heritage humans got from their ancestors. But although the influence of the thalamus has been diminishing along the path of evolution, it has not ceased to exist! Evolution possibly determined that

the existence of a quick, basic, immediate path of processing is still essential even in species with high cognitive abilities, like humans [34].

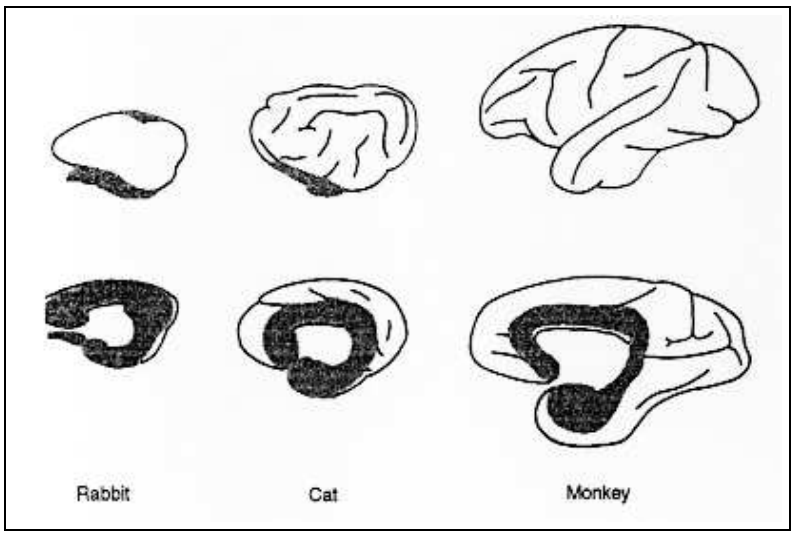

Figure 2.6: Areas occupied by the limbic lobe (the evolutionary older part of the brain) of three animal species, along the path of evolution: rabbit, cat, and monkey. (From [34] page 86, reprint by courtesy of the author.)

There still is a considerable amount of discussion around the issue of basic emotions. The idea of basic emotions is to pinpoint a basic set of emotions, from which, by combination, every emotion felt by humans can be described. One of the most prominent persons behind this theory is Paul Ekman [21]. But besides this ongoing discussion, it seems consensual that *fear* is to an essential phenomenon, whether it is part of a set of basic emotions, or there is no sense in defining such a set. Fear is known to exist in animals since early stages of evolution. The study of the way the human brain deals with fear provides important leads to the inner workings of emotion.

Joseph LeDoux [34] has carried out an exhaustive research on the brain circuits of fear, mainly on rats. And once again a double processing mechanism was found:

> So we can begin to see the outline of a fear reaction system. It involves *parallel* transmission to the amygdala from the sensory thalamus and sensory cortex. The sub-cortical pathways provide a crude image of the external world, whereas more detailed and more accurate representations come from the cortex. While the pathway from the thalamus only involves one link, several links are required to activate the amygdala by way of the cortex. Since each link adds time, the thalamus pathway is faster. *[emphasis added by the author]* ([34], page 165)

And as far as response time is concerned:

> Although the thalamic system cannot make fine distinctions, it has an important advantage over the cortical input pathway to the amygdala. The advantage is time. In a rat it takes about twelve milliseconds (...) for an acoustic stimulus to reach the amygdala through the thalamic pathway, and almost twice as long through the cortical pathway. ([34], page 163)

These findings corroborate the double processing model of a complex/slow and a basic/fast layers proposed here. In the next section, the relationship between these structures and human rationality is explored.

2.3 Rationality and Emotion

At the beginning of this chapter it was said that for a long time the dominant thought was that emotions were an undesirable byproduct of the human rational mind, and that the less emotional a person was, the more (s)he would think rationally. Antonio Damasio was one of the first researchers to openly state otherwise. Daring claims require daring approaches, and Damasio was able to come up with experimental evidence that, in fact, emotions play a key-role in human reasoning.

But what is really understood here about emotions? A dictionary [39] definition of "emotion" reads "the affective[6] aspect of consciousness," and further ahead more precisely as

> A psychic and physical reaction (as anger or fear) subjectively experienced as strong feeling and physiologically involving changes that prepare the body for immediate vigorous action. ([39])

This last definition is clearly oriented towards the physiological aspects of emotion, although it rejects any possibility of ascribing emotions to machines. Unless, of course, one could ascribe all the terms used (body, psychic, subjectively, feeling, and so on) to the same machine. On the other hand, the first definition, although being detached from any physiological ground, is too vague (and entangled in circular definitions) to be useful, besides using the similarly "precarious" word "consciousness."

Most of Antonio Damasio's experimental data stems from patients with brain lesions in the prefrontal cortices, which reside just behind the head fore bone, right above the eye balls. In his book [18], Antonio Damasio develops his argumentation around three case-studies, which will be briefly

[6] Defined in a circular fashion in the same dictionary as "relating to, arising from, or influencing feelings or emotions".

described below (the material in this subsection is quoted from[18], unless stated differently).

In 1848 Phineas Gage suffered an accident that destroyed a substantial part of his prefrontal lobes. He survived; however even though he did not become physically handicapped in any way, but his life changed forever. His character, his personality was deeply modified. He became unable to behave in presence of others, was rude, acted like a child, and he was unable to resume his previous job. Hopping from job to job, he even became a circus attraction, showing his wounds and the iron stick that was responsible for the accident.

The second case is the one of a patient named Elliot. He suffered from a brain tumor that compressed the prefrontal cortices, damaging them. He underwent surgery to remove the tumor as well as the damaged tissues of the prefrontal lobes. As a result, his behavior was also deeply affected. According to Damasio's words:

> Once at work he was unable to manage his time properly; he could not be trusted with a schedule. When a job called for interrupting an activity and turning to another, he might persist nonetheless, seemingly losing sight of his main goal. Or he might interrupt an activity he had engaged, to turn to something he found more captivating at that particular moment.
>
> [...]
>
> The flow of work was stopped. One might say that the particular step of the task at which Elliot balked was actually being carried out *too well,* and at the expense of the overall purpose. One might say that Elliot had become irrational concerning the larger frame of behavior, which pertained to his main priority, while within the smaller frames of behavior, which pertained to subsidiary tasks, his actions were unnecessarily detailed. ([18], page 36)

As Damasio notes, these two cases have much in common:

> In some respects Elliot was a new Phineas Gage, fallen from social grace, unable to reason and decide in ways conducive to the maintenance and betterment of himself and his family, no longer capable of succeeding as an independent human being. ([18], page 38)

Apparently, both cases showed no weakening of pure cognitive abilities (as the ones measured by the traditional I.Q. rating[7] Only Elliot was actually examined, but it is supposed that Gage would obtain similar results.). Yet, they were unable to handle common-sense tasks, they lacked the ability to coordinate all these particular cognitive abilities usually recognized as intelligence, into a coherent whole.

[7]See [5] for a description of the I.Q. test.

Latter patients suffering from similar lesions in the prefrontal cortices showed another common feature: they all had a strong impairment on their *emotional* assessment of situations.

There are two brain structures that are essential to these mechanisms. They are the *amygdala*[8] and the *prefrontal cortex.* Damasio classifies emotions in two broad classes: *primary emotions,* that are triggered by external stimuli, originating body responses such as sweat, blood pressure, and so on; and *secondary emotions* which are relative to recalled images from "emotionally charged" past events. The primary emotions rely on the amygdala (older part of the brain in terms of evolution). Certain external stimuli trigger the amygdala to produce a body response. The secondary emotions are based on the prefrontal cortex, but work on top of the amygdala: images of past events are activated in the brain, and the prefrontal cortex responds by activating the amygdala to produce a body response. In general, this response is milder than the one directly provoked by external stimuli.

According to Damasio, the same areas in the brain whose lack deeply affects reason and long-term planning, are also responsible for the ability to have an emotional response to certain stimuli. This is more than a coincidence, and in fact, these two aspects — rationality and emotion — are *deeply entangled.* To explain this connection, Damasio raises the *somatic-marker hypothesis*:

> When the bad outcome connected with a given response option comes into mind, however fleetingly, you experience a gut feeling. Because the feeling is about the body, I gave the phenomenon the technical term *somatic* state ("soma" is Greek for body); and because it "marks" an image, I called it a *marker.* Note again that I use *somatic* in the most general sense (that which pertains to the body) and I include both visceral and nonvisceral sensation when I refer to somatic markers. ([18], page 173)

In other words, certain *images* (recall the previous discussion about how thought is largely made out of images) are *marked* with a *somatic* (as relative to the *body*) representation. The body plays here a fundamental role as the "theater for the emotions," to quote Damasio. The effects of this somatic marker can either be properly visceral, in the sense that it modifies certain physiological characteristics (blood pressure, hormone balance, and so on), or short-circuiting the body through an "as-if" mechanism, but still holding the same characteristics.

[8] The amygdala is not present in the Papez circuit described in the previous section. It was later introduced by MacLean in 1952 (see [34] for further details). The role of the hypothalamus is related to body regulation issues, where the amygdala is in fact responsible for its activation.

To verify this hypothesis, Damasio describes several examples, out of which three will be reported here.

- When a patient visiting Damasio's laboratory pulled out his appointment book to schedule his next visit, with cold posture, he started enumerating reasons for this or that date, without being able to decide. It took more than a half-hour, without neither being able to decide, nor showing any sign of frustration. He just kept analyzing, comparing possible dates, endlessly.

- Several patients with lesions in the prefrontal lobes were matched against normal persons, in terms of skin conductivity while watching to the same sequence of pictures. These pictures included banal images, like landscapes, as well as disturbing pictures (violence, blood, accidents, sex, etc.). The results were very clear. While the disturbing pictures produced strong skin conductivity response in the normal subjects, there was no noticeable response from the ones with the prefrontal lobe lesions. Although they were able to correctly understand the horror of these pictures, they did not show any emotional response. One of the impaired patients showed a remarkable insight of what was happening to him:

  > He noted that after viewing all the pictures, in spite of realizing their content ought to be disturbing, he himself was not disturbed. ([18], page 211)

  And Damasio further notes that:

  > Here was a human being cognizant of both the manifest meaning of these pictures and their implied emotional significance, but aware also that he did not "feel" as he knew he used to feel — and as he was perhaps "supposed" to feel? — relative to such implied meaning. The patient was telling us, quite plainly, that his flesh no longer responded to these themes as it once had. ([18], page 211)

- The third example is the setup of a card game[9], consisting of four decks — A through D. The subject is asked to turn a card, from a deck of his choice, then the experimenter asserts whether that card made him lose or gain a certain amount of (fake) money (from a start loan of $2,000). Cards from any of the A or B decks offer the subject $100, while cards from decks C and D only give $50 each. The tricky part of this game

[9] Although the cited Damasio's book [18] describes this game, detailed information about the results and card sequences can be found in [3] and [4].

is that certain cards in decks A and B unexpectedly produce a loss of high amounts (e.g., $1,250), but in decks C and D certain cards only cause a minor loss of less than $100. Each game consists of 100 turns, but players were not informed beforehand.

Normal people usually started the game trying each of the decks, but soon would take notice of the high losses resulting from the A and B decks, and converge to taking cards from decks C and D only. However, patients with prefrontal lobes lesions, kept on taking cards from the apparently more profitable decks A and B, insensitive to the occasional high losses (figure 2.7). These patients were unable to recall the risk of choosing A or B deck cards, and kept on choosing the immediately apparent higher value of these decks. Damasio calls this phenomenon "myopia for the future".

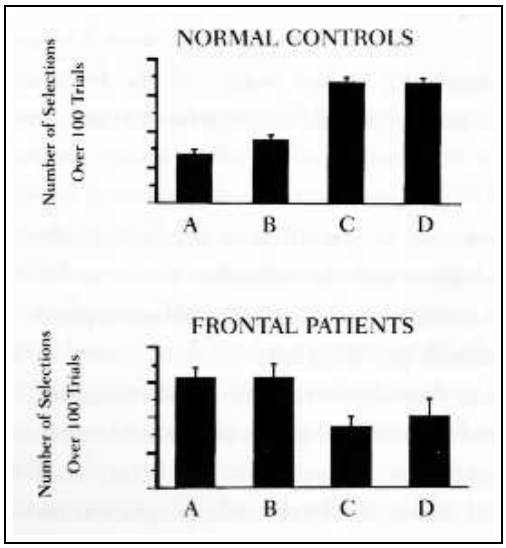

Figure 2.7: Number of selections from each of the decks, in normal subjects and "frontal patients" (*i.e.,* suffering from frontal lobe damage). (From [18] page 215, reprint by courtesy of the author.)

These results suggest that, when normal players are faced with the four decks, they perform a double assessment of each deck, while in the case of the impaired patients, it is only a single one. The assessment that is common to both of them corresponds to a crude low-term evaluation, based on the most recent card values. The assessment missing in the impaired patients is the ability to recall a somatic marker associated to a past event. In this case, this would be the (sad) remembrance of the high loss cards taken out from A and B decks. This mecha-

nism overrides the first crude assessment, and holds long-term benefits throughout the game. Frontal patients (*i.e.,* who suffer from frontal lobe damage) are unable to foresee the high losses from the A and B decks. The distinction between these two kinds of assessment will further ahead be nicely mapped into our proposed double layer model.

These results are in fact the major contribution to this thesis — *emotions play an essential role in human rationality.* And this result is not an assumption, but rather a conclusion supported by experimental data.

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

> *A culpa foi minha, chorava ela, e era verdade, não se podia negar, mas também é certo, se isso lhe serve de consolação, que se antes de cada acto nosso nos puséssemos a prever todas as consequências dele, a pensar nelas a sério, primeiro nas imediatas, depois nas prováveis, depois nas possíveis, depois nas imaginávias, não chegaríamos sequer a mover-nos de onde o primeiro pensamento nos tivesse feito parar.*[10]
>
> José Saramago, "Ensaio sobre a Cegueira"
> (pg. 84, Editorial Caminho, 1995)

[10]It was my fault, she cried, and it was true, it could not be denied, but it also holds, if that can serve as a consolation to her, that if we predicted all consequences before each act, considering them seriously, first the immediate ones, then the probable ones, then the possible ones, then the imaginary ones, we would never get to move beyond where the first thought would have made us stop. [author's translation]

Chapter 3

The Model

Summary

The model hypothesized in this thesis is presented here in an incremental fashion. But before starting the presentation, a set of basic assumptions is set. After supporting the model in the double representation paradigm, the perceptual layer is presented, followed by the cognitive one. The desirability vector concept is then introduced. The way these two layers interact in order to produce a decision and/or an action is then discussed, followed by considerations on the role of the memory, that implements the capability of learning. Finally, the complete picture of the architecture, containing all the discussed components, is presented. The way this architecture functions as a whole is also discussed.

3.1 Basic Assumptions

The proposed model is built on top of the *agent* paradigm. The agent is in contact with the environment (which may include other agents, with or without similar architectures) through its *sensors,* and acts upon it by the means of its *actuators.* The core of the agent — the internal entity that generates actions based on percepts (as well as the agent's internal state) [49] — constitutes the model that will be proposed, developed and discussed. The agent conceptual framework is for now considered as an individual. The concept of multi-agent systems is an interesting prospect [64], but lies outside the scope of this thesis.

The starting point of the model are the perceptions, which will be also termed *stimuli.* Each stimulus models a perception event received by the agent sensors. These perception events will also be called *images.* The choice for this term derives from the fact that in the brain, as was discussed in

section 2.1, information is usually (if not always) represented by topographic maps. The natural way of thinking about these maps is as visual images. But in this context, the name *image* is meant to comprise not only visual images, but also other kinds of perceptions that can be encoded in a topographic map: auditory, tactile, motor, and so on.

In physical environments it seems natural to represent stimuli in this manner. But when purely synthetic environments are to be considered, this choice may not seem as natural. The advantages of putting information together in a topographic fashion have to be considered for each case. The question is how to represent stimuli in such a way that the exploitation of topographic properties can be useful. These considerations, as well as the usefulness of representing things in this fashion, are of course domain-dependent. Still, it will be assumed here that stimuli have this topographic map form, *i.e., images.*

3.2 Double Processing

In chapter 2, it was shown that a double representation scheme could be found throughout many of the presented models. And this paradigm forms the starting point of the proposed model.

It is hypothesized that whenever the agent receives a stimulus (an image), it processes it, in parallel, that is to say, simultaneously, under two different perspectives: a *cognitive* and a *perceptual* one. The *cognitive* processing gets a complete picture of the stimulus, as close to the original stimulus as possible. This results in a *cognitive image.* The *perceptual* processing extracts a minimal set of features, which are considered as essential, basic, built-in, by design. These features can be arranged in a structure designated by *perceptual image* [62].

This distinction requires some clarification. Imagine an animal facing a fast moving object: this triggers a "flight or fight" kind of reaction, which derives from the assessment of the apparent threat. From this stimulus, this animal extracts a cognitive and a perceptual image. While the former is complex, and is therefore takes time to process and analyze, the latter is extracted quickly, but tells the animal little more than whichever class does the stimulus corresponds — the danger of a predator, or the desirability for catching a prey. This is a basic, built-in feature which is innate to the animal [62].

Thus, while a cognitive image is complex, of slow processing, but rather complete[1], the perceptual image is simple, basic, small, quickly extracted

[1]The word "complete" is to be understood here with respect to the perceived stimulus,

representation of a stimulus that is but primitive and reduced.

3.3 Perceptual Layer

It is assumed here that, in order to assess how to cope with a given environment, there must exist a minimal, basic set of features that can be extracted from stimuli. Without this built-in knowledge, as it will soon become clear, the agent would be indifferent to the world, *i.e.,* all stimuli would look the same. This representation can be said to provide *relevance* to external stimulus [61].

When faced with a specific environment, the question of what shall be considered perceptual (that is to say, built-in) and what shall not, becomes a crucial one. The behavior of the system when first exposed to the environment, as well as throughout its life, can be radically different depending on these design choices. What are the issues that define these choices? A formal answer to this question has to be postponed until there is a better understanding of this model. For now, it must be understood that this choice depends, at a first sight, on what stimuli have to be considered as a minimal, basic set, in order to allow the agent to bootstrap. For instance, considering an animal, sights of predators and preys definitively belong to this set. These stimuli are also related to the needs of the agent in order to survive.

Unveiling a bit of what will be discussed ahead, these perceptual assessments are going to be associated with cognitive images. The agent will learn, in contact with the environment, to cope with it. To learn new associations means to evolve and to gain from experience. But in order to do that the agent must be able to assign a basic meaning to certain stimuli — a minimal set on top of which a much larger and complex set of stimuli can be learned and recognized, by means of association. Given a certain environment and a specification of the objectives to be accomplished by the agent, a set of perceptual stimuli has to be defined.

Consider the example of a robot moving in an human-inhabited office room. Some candidate perceptual features are the ones provoked by: close proximity to walls, namely quick movement towards them, proximity to moving people (or other robots), direct exposure to sunlight (may overheat the robot, or on the contrary may supply it with solar power), lack of floor (such as proximity to stairs running down), and so on. With these features, the robot would be able to move around, avoid disturbing people, avoid damaging environments, and so on. In order to provide the robot with means to do other things, additional perceptual features are required. For instance,

and not to the object that originated that stimulus.

imagine hardwiring the obligation to obey orders from humans — disobeying orders could result in "pain," to use a daring word.

In addition to the existence of a built-in core in the perceptual layer to trigger the agent bootstrap, this layer is allowed to adapt to the environment through time. As the agent interacts with the environment, it may find it necessary to respond perceptually to new classes of stimuli. For instance, it may find that whenever it approaches orange walls, it senses collisions. Thus, the perceptual layer can be allowed to learn, in a way that will be further detailed later in this chapter.

3.4 Cognitive Layer

The nature of the cognitive layer is defined in counterpoint to the perceptual one. A cognitive image contains as much information extracted from the sensors as feasible. It contains mostly (unfiltered) raw information.

Consider an example of a visual cognitive image in a robot with camera vision. First of all, every pixel gathered from the camera apparatus is retrieved. Additional processing can be accomplished, such as edge detection, segment extraction, displacement profile, and so on. A cognitive image includes not only the results of these algorithms, but also the raw input image. In the case of hierarchical processing, where succeeding algorithms are applied to the results of former ones, the whole hierarchy of images is contained in the cognitive image.

The purpose of retaining the stimulus complexity is to allow the agent to remember past events, and re-analyze them under the light of new knowledge.

3.5 Desirability Vector

The *desirability vector* (DV for short) is the mechanism that supports the basic representation of the perceptual layer [62]. The perceptual layer's major role is to map stimuli to the DV. It can be considered as the model's equivalent to the "body," in Damasio's terminology. Thus, it plays a fundamental role in the model.

Each one of the desirability vector components represents a particular kind of assessment of a stimulus. Each component can be either activated or neutral (varying either discretely or continuously). Neutral components mean no assessment. But when a certain component is activated, it means that the stimulus triggers a specific assessment, e.g., is it good? is it bad?[2]

[2]The ethical terms "good" and "bad" should be taken here in their empirical sense.

Going back to the predator/prey metaphor, the animal sees a predator, it triggers a strong activation of the DV fear component. If, on the other hand, when the animal finds a prey, it is now the "tasty" component that is activated. A minimal DV consists of a positiveness and a negativeness component. When a stimulus is considered as positive, from the agent's point of view, the positiveness component is activated. When on the contrary, the stimulus is considered negative, it is the negativeness component that is activated. Otherwise, if both components are neutral, the stimulus is considered to be irrelevant. If both components get activated, it is not clear what it means. It corresponds to an abnormal situation.

Certain basic stimuli are able to trigger, at a first level, certain components of the DV. For instance, a threatening stimulus, may activate a "fear" DV component, which ultimately generates a fear behavior[3]. This path, starting in the agent sensors, through the DV, and leading to an immediate action is of an extreme importance. Note that all these mappings[4] are built-in. It means that agents, in the first steps of their contact with the environment, are capable of behaving "appropriately," provided that some care is taken in choosing those mappings. It is interestingly to consider the use a genetic algorithm approach [28] to come up with a working set of mappings, instead of designing them "by hand."

For a given stimulus, the evaluation of the DV may not be hard-wired. Despite the fact that there must exist a built-in mapping, prior to the agent contact with the environment, through this contact, the agent may learn to associate new classes of stimuli to DV instances. But this kind of learning is distinct to the one performed in conjunction with the cognitive layer. At the level of the perceptual layer, the kind of learning is very basic, in the sense of a direct mapping between the stimulus and the DV. This mapping can be updated through time. The major distinction from the cognitive kind of learning is that while the latter is *explicit*, the former is *implicit*. Explicit learning refers to the existence of a memory of events that can be individually recalled, while implicit learning implies a monolithic adaptive structure that simply maps inputs (stimulus) to outputs (DV). These designations are inspired by a classification of memory[5] cited in [14] (page 244).

For instance, a stimulus is asserted "good" by the agent if it is desirable in terms of its objectives. The use of this empirical terminology is meant for the sake of clarity.

[3]As above, the term "fear" should also be taken in its empirical sense. Consider it as a threat to the agent's survival.

[4]A direct mapping between the DV and the agent's action is assumed here, before being presented below in section 3.8.

[5]In this classification, explicit memory is further divided in "facts" and "events". Although these two classifications are found relevant in the context of this thesis, only the

Cognitive images, when become associated with DV instances, can be considered to provide *meaning* to this images. But more interesting than establishing associations between cognitive images and DV instances extracted from the perceptual image of the same stimulus, is the idea of *propagating* these associations, from memorized associations to presently "non-perceptual" (null DV) stimuli. Previously irrelevant stimuli become relevant, due to past experience [63].

Until now, there has been some intermixing of the terms perceptual image and DV. It is true that both concepts result from the stimulus assessment process in the perceptual layer, but they have to be distinguished. The perceptual image is the result of the extraction of basic features from stimuli, while the DV components have explicit meaning. Moreover, while the perceptual image depends on the extracted features themselves, the DV components are independent of the nature of the stimulus. For instance, when an animal is faced with a quickly moving object, the perceptual image holds information whether it is a big object, or it is approaching the animal, while the DV addresses issues such as fear, attention, curiosity. The perceptual image is geared towards feature extraction from stimuli, while the DV holds its immediate meaning. Although the DV concept is essential to the definition of the model, the perceptual image is not. But this does not discard the usefulness of the concept. For instance, consider using the reduced set of features to *index* the memory, to narrow the search for cognitive matches [61, 60]. In the implementations presented in chapter 4, some of them use perceptual images.

3.6 Memory Issues

When the agent is faced with a relevant stimulus, the cognitive and perceptual images as well as the DV are associated and stored in memory. But how can the agent know whether a stimulus is relevant or not? Of course it cannot store every stimulus it perceives, flooding the memory with useless data. But it is not desirable to be too conservative, taking the risk of missing information that may later prove to be useful.

At a first stage, strong perceptual images are the only way to indicate relevant stimuli. The cognitive images associated with these stimuli are to be associated with the respective DV, and stored in the agent memory. The idea behind this association is the somatic marker hypothesis [18].

latter was considered. But one may imagine the "facts" kind of memory as association between cognitive images. These associations must however support an additional representation to specify how these two images are related.

When later on the agent faces a similar stimulus, say, recognizes the cognitive image, it is able to recall a previous association, and act accordingly. Consider, for instance, the agent approaching an unknown object. When close enough, the object "bites" it, causing a very "negatively" charged DV (meaning that the DV assessment indicates a "negative" stimulus). The agent associates the (cognitive) image of the object with the perceptual image and the DV. When later on, it faces the same object, it refrains from approaching the same object, from a distance. Although the DV (directly mapped from the stimulus) does not reveal immediately the danger of the approach, the agent recognizes the cognitive image, and recalls the "negativeness" of the encounter.

As cognitive images were defined as complex representations, this matching process can be rather slow. But keeping in mind that the perceptual layer works in parallel with this matching mechanism, it is able to detect some stimuli deemed relevant, prior to the completion of the matching process. The perceptual layer, due to its simpler and faster nature, is able to deliver a quick response. Furthermore, this response can indeed help the cognitive matching process, narrowing the possibilities, for instance. Because of the adaptability of the perceptual layer, as the agent interacts with the environment, and finds new stimuli that are considered essential, this guidance to the matching process becomes more and more refined. The DV and/or the perceptual image can provide a useful help in this process.

3.7 Learning

Reiterating the learning issue, this model encompasses two distinct learning mechanisms: cognitive and perceptual. Perceptual learning is the adaptation that the immediate direct mapping between stimuli and DV undergoes through time. Cognitive learning also involves the perceptual layer, in the sense that associations also involve the DV. But unlike the perceptual memory, this memory is organized in events. Each association stored in memory corresponds to a single event, that led to the association. Furthermore, these associations may be related to one another by the means of other structures, which may for instance hold what could be considered as *context*. The cognitive layer may handle a complex web of knowledge representation. But the way this web is constructed and used, depends on perceptual representations, namely the association of cognitive images with DV instances.

Higher cognitive abilities, like reasoning and planning can be conducted primarily at this cognitive layer. The activation and recalling process is not restricted to the matching mechanism. The process of searching the memory

for a matching cognitive image is just a first level approach. Other processes could trigger the intervention of stored associations. And this suggests that high-level processes can endorse more complex cognitive abilities.

3.8 Decision and Action

The difference between "action" and "decision" lies in the fact that while the former is externally observable (in terms of the agent actuators operating on the environment), the latter is not, being an internal phenomenon. However, the agent's actions are a consequence of the agent's decisions.

From the double processing paradigm, two kinds of decisions have to be combined: there is an immediate decision provided by the perceptual layer, resulting from the basic assessment delivered by the DV, and a decision resulting from the cognitive processing. Although the DV can be mapped directly to a decision, the same does not happen with a cognitive image. When the agent is first faced with the environment there is no way to generate decisions from cognitive images alone, because the only built-in mechanism able to produce actions lies in the perceptual layer. The DV plays an important role here, because it is able to bridge the gap between the DV-decision mapping and the cognitive layer.

Although the decisions generated by the agent are mainly considered here as resulting from a DV-decision mapping, this does not validate the existence of purely cognitively derived actions. Consider, for instance, the process of supervised learning. Several actions are shown to the agent (assume that they are correctly perceived and represented as cognitive images), some are shown to be "good" and others to be "bad." This suggests that the agent associates images of actions (in the cognitive layer) to certain DV instances. In the future, when faced with the necessity to plan a sequence of actions, the agent my recall these learned actions, stored as cognitive images, and decide to act based upon them. In this scenario it cannot be said that a DV-decision/action map suffices. A mechanism to store action schemes and to generate actions through the cognitive layer is required.

In fact, the `faces` implementation to be described in chapter 4 uses a simple chaining mechanism, from which actions can be derived. This mechanism resides in the cognitive layer. Notice that this chaining device allows the agent to *plan* a sequence of actions, suggesting a way of implementing planning with this model.

3.9 The Complete Picture

The purpose of this section is to give a complete picture of the model to the reader. The above concepts and mechanisms are here gathered into a whole. A diagram of the model is presented in figure 3.1.

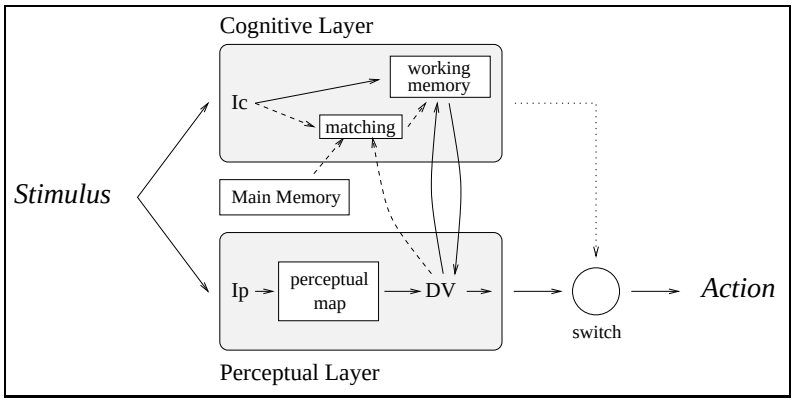

Figure 3.1: The complete picture of the proposed model, containing all the components discussed in the above sections. (see figure 3.2 for memory structure details)

Here follows a summary of how the model works: in response to an external stimulus, the cognitive and the perceptual layer process it in parallel. At the perceptual layer, there is a direct map between stimuli and the DV. When the agent is built, a part of this mapping must already exist, in order to allow it to bootstrap. Furthermore, this map is adaptive. This forms a kind of implicit memory, termed *perceptual memory.* On the other hand, the cognitive processor looks into the main memory for matches of the cognitive image. This memory contains experienced associations, but unlike the perceptual memory, these associations are individually stored as representing events[6]. These associations contain both the cognitive image, the corresponding DV, and the perceptual image (if implemented). The origin of this DV comes primarily from the perceptual layer, but one can also consider propagating DV instances from other associations. This is a way to allow the agent to associate cognitive images to DV instances, even when faced with a situation where the input stimulus does not deliver (in the perceptual mapping) a significant DV. This memory is here termed *main memory.* The *working memory* holds the input cognitive image, the DV (and optionally the perceptual image), as well as the results from the matching process (or any

[6]Note that in the future, other kinds of representations other than events may be placed in this memory.

other higher-level cognitive processes). Figure 3.2 illustrates these memory structures. The action, in response to the stimulus (if any) comes primarily from the DV, although there is provision for actions originating from the cognitive layer. If the agent decides on any action, it may produce alterations in the environment, which can be perceived by the agent as a *feedback stimulus*. This new stimulus tells the agent the result of its action. It is fed into the architecture, in order to make the agent learn. This learning can be accomplished at several levels: at the perceptual layer, it can adapt the perceptual map to be sensible to new stimuli, and at the cognitive layer, it can mark (one or more) cognitive images with the DV, along with the action that led to the environment feedback.

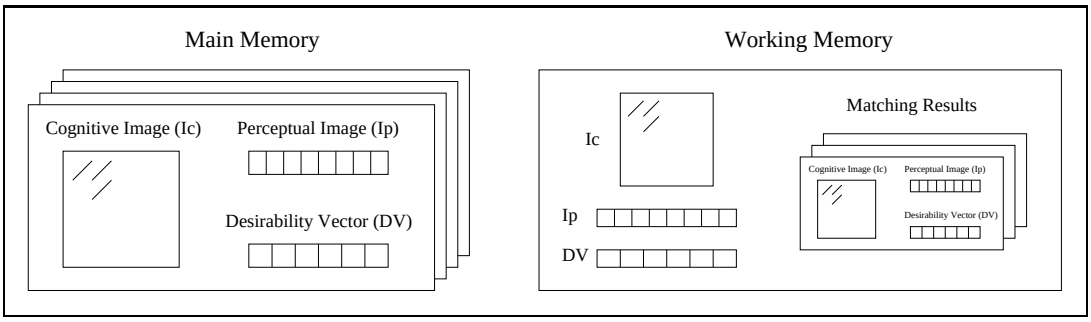

Figure 3.2: Memory structure of the main and working memory.

Note that these descriptions are deliberately vague on some issues. There are several degrees of freedom left. For instance, how the switch between cognitive actions and perceptual actions works — the former kind (when present) may override the latter, but for strong DV instances, it may be important to ignore the cognitive outcome. Or since the perceptual layer is able to deliver an action prior to the cognitive layer, shall it act immediately, or shall it wait for a more precise cognitive assessment? Once again, it may depend on the gravity of the situation. Another degree of freedom is the way new associations are established. They can be established as soon as stimulus triggering DV components reach the agent, and/or after the environment feedback.

This section tried to offer a global description of the model in as much detail as possible, but in the implementations that will be discussed in the next chapter, several simplifications were made. These simplifications were done not only to narrow the issue under experimentation, but also to make the interpretation of results clearer.

Life is like music; it must be composed by ear, feeling, and instinct, not by rule.

Samuel Butler

Art is not a handicraft, it is the transmission of feeling the artist has experienced.

Leo Tolstoy

Chapter 4

Experimentation

Summary

In this chapter, three implementations are presented, as well as the obtained results. These implementations correspond to different stages of the development of the model, so the early ones denote some divergences from the final picture presented in the last chapter. In the first one (termed `damasio`*), a basic marking mechanism is tested, while the second one (*`faces`*) shows some consequences from the intermixing of the cognitive and perceptual processing. The third implementation (*`decks`*) is a simulation of the decks experiment described by Damasio ([18] page 212), showing similar results to the ones obtained with the normal subjects and the patients with frontal lobe damage.*

This section describes three implementations that went along with the development of the model. Note that some issues in the early implementations bore modifications up to the latter ones. The following sections should be understood as snapshots of three views of the model through its evolution.

All implementations presented share the same execution model. The agent lives in an episodic environment. Each episode starts with the presentation of a stimulus, followed by the agent decision (and action when so decided). Except for the first implementation, the environment responds to this action with another stimulus. This corresponds to the environment feedback for the agent action.

In each episode, the agent performs the following sequence of steps:

1. Double processing of the incoming stimulus S, extracting a cognitive and a perceptual image — I_C and I_P;

2. Use these extracted images to search the agent memory, and copy the

similar ones to the working memory. This task can be helped by the perceptual image I_P;

3. Using all the information gathered in the working memory, build an assessment of the stimulus;

4. Form a decision, and possibly an action to be taken;

5. Receive the feedback from the environment;

6. Update the agent memory.

All these steps are not necessarily present in all implementations. But they form the guidelines of what the agent is supposed to do.

The following implementations were written in ANSI Common Lisp language [33, 29], using the CLISP implementation [30], running on the Linux operating system [41]. The graphical interface for the `faces` implementation used the Tcl/Tk scripting language [43, 15], in addition to a Common Lisp core.

4.1 Implementation: `damasio`

The first implementation, called here `damasio`, was an attempt to experiment with the somatic marker mechanism. The motivating idea was to obtain the kind of behavior found when people associate a thunder with the flash of lightning. In this metaphor, the thunder corresponds to the perceptual image, while the flash to the cognitive one. Once the agent associates this two images, when in the future it only sees a flash of lightning, it immediately "expects" the thunder.

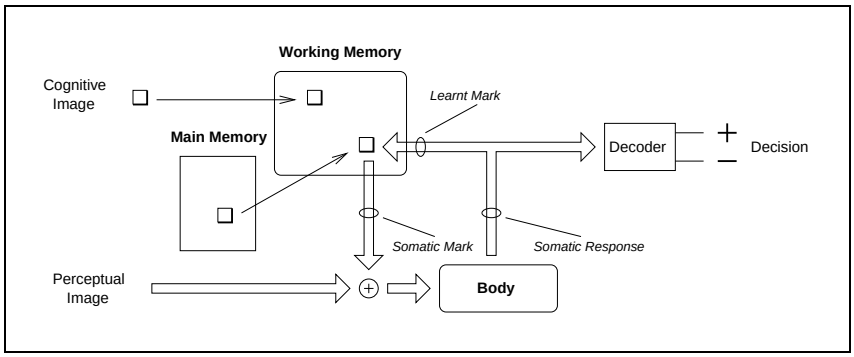

Figure 4.1: Architecture of the `damasio` implementation.

The architecture of this implementation is shown in figure 4.1. The agent perceives external stimuli through two channels: the cognitive part of the stimulus (e.g., the flash of lightning), and the perceptual one (e.g., the thunder). For simplicity, these inputs are bidimensional vectors. There is a (short-term) working memory, where the present input is used to recall past associations, and an output is obtained; and a (long-term) main memory, where associations are stored throughout the agent life. The recalled associations are combined with the environment input to derive a body response (labeled "somatic mark"). This body response (labeled "somatic response") is used to trigger a decision (positive or negative, for simplicity — "is it good?" or "is it bad?"), and to update the association, depending on its similitude to the stimulus.

The architecture works as follows: each stimulus corresponds to a pair (cognitive, perceptual) of vectors. The cognitive vector is copied into the working memory, and the main memory is browsed for similar vectors. For simplicity, all associations from the main memory are considered, but only a pre-defined number of them are copied into the working memory. For each main memory association, the similarity between its cognitive vector and the incoming one is computed and registered. The `*max-wm-images*` (a numerical constant) higher value associations are chosen and copied to the working memory. In the working memory, these associations form *frames*. A frame contains the recalled association (the cognitive vector and a mark vector), and the similarity measure. Next, each of these frames are combined with the perceptual input. Figure 4.2 shows this mechanism in detail.

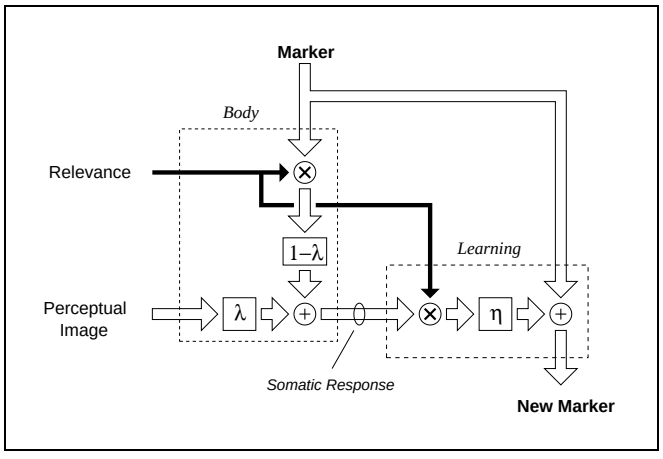

Figure 4.2: Marking mechanism in the `damasio` implementation. A body response ("somatic response") and an updated mark is computed, from the perceptual input, the old mark, and a similarity measure.

Using the perceptual image, the mark, and the similarity measure (termed "relevance"), a body ("somatic") response and an updated mark are computed. This mark is associated to the originating association, and supersedes the corresponding association in the main memory. Note that the incoming stimulus always forms a new frame in the working memory, and its mark is initially put to zero (null vector), and the similarity measure put to 1 (maximum similarity). These operations are performed according to the formulas

$$R = \lambda I_P + (1 - \lambda) s M \tag{4.1}$$

$$M' = M + \eta s R \tag{4.2}$$

where I_P stands for the perceptual image, M and s for the current frame mark and its similarity measure, R the body response, and M' for the updated mark value. The rationale behind equation (4.1) is to linearly interpolate between the present perceptual image and the body response marked on the recalled image, weighted by the similarity measure s (relevance), which ranges from 0 (not similar at all) and 1 (maximum similarity). This interpolation is controlled by the λ coefficient ($0 \leq \lambda \leq 1$). The role of s is to allow the recalled mark to influence the outcoming somatic response R, depending on the similarity found between the present stimulus and the recalled one. Strong marks on very similar stimulus should provoke higher body responses than less similar ones. This similarity measure s accounts not only for the cognitive image similarities, but also for the perceptual image. With respect to (4.2), the idea is to update the new mark M' according to two coefficients: the similarity measure (the more similar the stimulus is, the more it should be updated), and a learning rate η.

As it was previously noted, both the cognitive and perceptual images are bidimensional vectors, as well as the referred marks. The similarity measure is evaluated using the following expression:

$$d(u, v) = \exp\left[t\sqrt{(u_1 - u_2)^2 + (v_1 - v_2)^2}\right] \tag{4.3}$$

where $u = (u_1, u_2)$ and $v = (v_1, v_2)$ are the considered images. The constant $t < 0$ conditions the decay rate as u and v become apart. This constant can be interpreted as a tolerance value — "how much shall I consider this (non-identical) image pair similar?". The expression used for measuring mark similarities is the same. The total similarity, between the stimulus and the recalled frame is weighted by ξ ($0 \leq \xi \leq 1$) between these two measures:

$$s = \xi d(I_C, I_C^M) + (1 - \xi) d(I_P, M) \tag{4.4}$$

where I_C and I_C^M denote the input and the recalled cognitive images.

In this implementation there is no perceptual feedback. Associations are built directly from the stimulus. Furthermore, there is no resulting action. The mark vector is interpreted as the first component being the amount of positiveness, and the second being the amount of negativeness. The DV can be understood in this implementation as being equal to the perceptual image. A classification is computed for each working memory frame, as being the difference between the first and second components. Its purpose is to measure the assessment of the frame ("good" if positive, and "bad" if negative) as well as how strong that assessment is (absolute value). The strongest frame (higher classification, in absolute value) is picked as the agent's final assessment of the stimulation.

The experimental setup for this implementation comprises three phases. First, a set of four stimuli was presented, two of them strongly positive, and the other two strongly negative. These stimuli are called *A1-*, *A2+*, *A3+*, and *A4-*. The ending signal is + or - depending whether they are positive or negative. The location of the stimuli in the Cartesian plane is shown in figure 4.3 as black filled balls. Positive stimuli have perceptual image $I_P = (0.8, 0)$ while the negative ones have $I_P = (0, 0.8)$. The agent was sequentially stimulated with this set of four stimuli four times, in order to get them clearly marked in the agent memory.

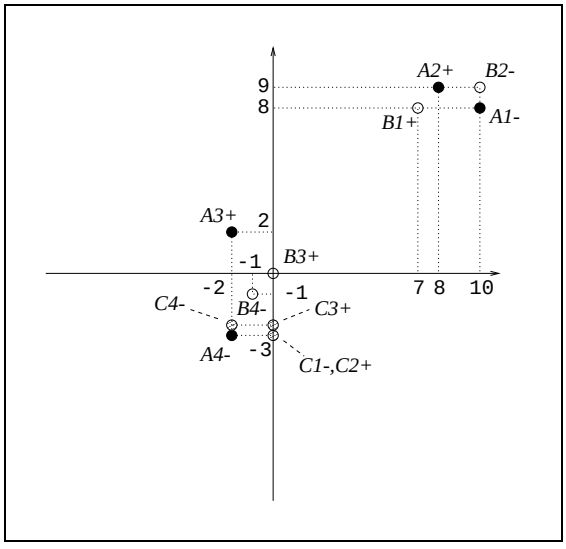

Figure 4.3: Location of the stimulus cognitive image vectors in the `damasio` experiment. The coordinates of each point in the Cartesian plane denote the bidimensional vector of the corresponding cognitive image. See text for the experiment description, as well as the used notation.

Next, a series of four stimuli with null perceptual image $I_P = (0, 0)$ were applied. These stimuli are denoted *B1+, B2-, B3+,* and *B4-*, where the signal now represents the agent's assessment, *i.e.,* whether the strongest classification is positive or negative. As expected, these results are consistent with the closest stimuli experienced in the first phase. This shows that after the agent being submitted to a set of "strong" stimuli, it learned, and when stimulated with null perceptual image stimuli, the agent was able to classify them according to its previous experience. An abridged and annotated output of the implementation can be seen below, where for each stimulus, the "strongest" frame is shown:

```
> (percept-image '#(7 8) '#(0 0))  ; B1
Strongest image:
image= #(8 9)
  relevance= 0.3944934
  marker= #(0.9216685 0)
  response= #(0.2545145 0)
  classification= 0.2545145  ; +
> (percept-image '#(10 9) '#(0 0))  ; B2
Strongest image:
image= #(10 8)
  relevance= 0.49430355
  marker= #(0 0.9216685)
  response= #(0 0.3189088)
  classification= -0.3189088  ; -
> (percept-image '#(0 0) '#(0 0))  ; B3
Strongest image:
image= #(-2 2)
  relevance= 0.24728459
  marker= #(0.9216685 0)
  response= #(0.15954009 0)
  classification= 0.15954009  ; +
> (percept-image '#(-1 -1) '#(0 0))  ; B4
Strongest image:
image= #(-2 -3)
  relevance= 0.2855023
  marker= #(0 0.9216685)
  response= #(0 0.18419695)
  classification= -0.18419695  ; -
```

In this output dump, the presentation of a stimulus starts with the `(percept-image ...)` expression in the LISP interaction, followed by information regarding the strongest image in the working memory. Note the sign of the `classification` output.

Finally, an experiment to test the "expert discrimination" capability of the agent followed. A stimulus *C1-* with null perceptual image was applied, and as expected, the agent answered with a negative assessment (closest to *A4-*). Then, a positively marked stimulus *C2+* was applied ($I_P = (0.8, 0)$). Two "colorless" ($I_P = (0, 0)$) stimuli, *C3+* and *C4-*, were applied, resulting in a positive response to the first and a negative to the second. Given a new

scenario with the new stimulus *C2+*, the agent answered coherently, showing its ability to discriminate between *C3+* and *C4-*:

```
> (percept-image '#(0 -3) '#(0 0))  ; C1
Strongest image:
image= #(-2 -3)
  relevance= 0.30826822
  marker= #(0 0.8)
  response= #(0 0.1726302)
  classification= -0.1726302  ; -
> (percept-image '#(0 -3) '#(.8 0))  ; C2+ <-- strong stimulus
Strongest image:
image= #(0 -3)
  relevance= 1
  marker= #(0.8 0)
  response= #(0.8 0)
  classification= 0.8
> (percept-image '#(0 -2.5) '#(0 0))  ; C3
Strongest image:
image= #(0 -3)
  relevance= 0.68522453
  marker= #(0.8 0)
  response= #(0.38372573 0)
  classification= 0.38372573  ; +
> (percept-image '#(-2 -2.5) '#(0 0))  ; C4
Strongest image:
image= #(-2 -3)
  relevance= 0.68522453
  marker= #(0 0.69274604)
  response= #(0 0.3322806)
  classification= -0.3322806  ; -
```

These experiments were performed setting the parameters $\lambda = 0.3$, $\eta = 1$, $\xi = 0.8$, $t = -1$, and the working memory was limited to 5 frames. These constants condition the behavior of the agent in ways that allow some interesting considerations on possible interpretations. For instance, taking the λ parameter, which interpolates the somatic response between the perceptual image and the recalled mark, when significantly reduced (say, $\lambda = 0.05$), makes the agent less sensible to the perceptual image, relying more on its past experience than in present reality. Consider that right after the initial sequence of stimuli *A1–A4,* is applied a stimulus with cognitive image $(10, 9)$ (same as *B2*) and perceptual image set to $(0.4, 0)$ (mild positive). With $\lambda = 0.3$ the agent accepts the new stimulus, attributing a positive classification (it disregards the "negative experience" of *A1-*):

```
Strongest image:
image= #(10 9)
  relevance= 1
  marker= #(0.4 0)
  response= #(0.4 0)
  classification= 0.4  ; +
```

But when the λ parameter is reduced to 0.05, the agent disregards now the positive perceptual image, assessing the stimulus as negative (due to the influence of *A1-*):

```
Strongest image:
image= #(10 8)
  relevance= 0.49430355
  marker= #(0 0.9216685)
  response= #(0.020000001 0.43280482)
  classification= -0.4128048  ; -
```

How can this behavior be interpreted? The λ parameter plays an interesting role of making the agent more or less trusting of the perceptual, when faced with a contradictory past experience. This result has some similarity with a "superstitious" behavior.

This implementation deals only with the marking mechanism. The stimuli are very basic, not reflecting the complex nature of the cognitive memory. Furthermore, there is no action (and consequently no perceptual feedback). Associations are always done, filling the agent memory with data that may not be relevant. But the results are interesting, in the sense of showing the marking and the memory retrieval mechanisms.

4.2 Implementation: `faces`

This implementation presents several sophistications over the preceding one. The objective is to experiment with more complex stimuli models, as well as the environment feedback. So, the stimuli (equal to the cognitive images) are a square set of polychromatic pixels (16×16). The mapping between the stimulus and the DV is fixed by design. In fact, the perceptual map discussed in the section 3.3 used the perceptual image as an intermediate representation. This perceptual image contains a set of basic features extracted from the stimulus. These features are then mapped to the DV. Both maps are hard-wired.

The agent perception of the environment is limited to the 16 by 16 pixel images. Each pixel is one of `blank` (background), `black`, `green`, or `red`. The agent can take one of three decisions: `none` (inaction), `accept`, or `reject`. The environment is episodic. Each episode starts with the presentation of a stimulus, the agent is then allowed to produce an action, and then the environment responds with another stimulus (feedback). The perceptual features extracted are: number of red pixels (assessment of "redness"), number or green pixels (assessment of "greenness"), and total number of non-blank pixels (measure of object size). The DV has three components: three boolean

components, indicating whether or not the stimulus is "good," "bad," or "deadly" (*i.e.,* very dangerous). The perceptual image is mapped into the DV using a set of thresholds. For instance, if the total number of pixels is above a pre-determined threshold, and the number of green pixels is above another threshold, the "good" components of the DV is activated. In this implementation, the presence of green pixels corresponds to a "good" stimulus, while red pixels denote a "bad" one.

The model of this implementation is depicted in figure 4.4. The cognitive layer uses both the cognitive and the perceptual images to find for a memory match. The perceptual image is first used to select a limited set of candidate memory associations (termed memory frames)[1]. From those, the cognitive image selects the best match. If two conditions hold, the frame action is selected. Otherwise, the direct perceptual path is used to derive the action. These conditions are: there is a match, the difference measure between the cognitive image and the memory frame is below a certain threshold. This difference measure is simply the Hamming distance between the two images[2].

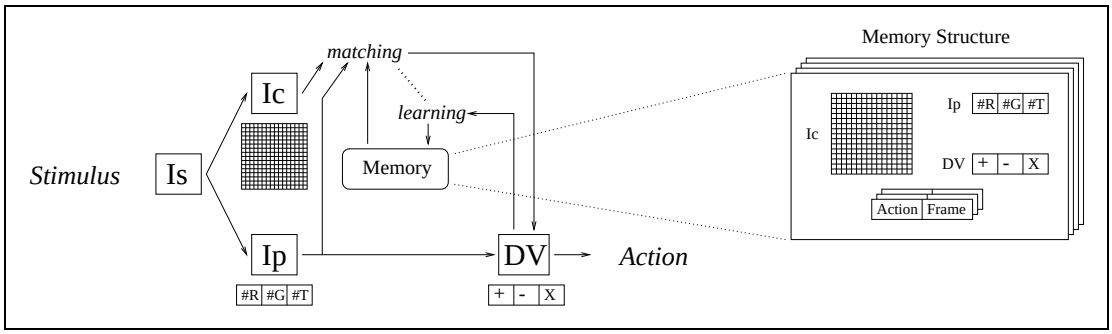

Figure 4.4: Architecture of the `faces` implementation.

A memory frame contains the cognitive and perceptual images, the DV, and an action list. This list consists of pairs (action, future frame), and is used to decide on the next action, based on the past experience. When a memory frame is selected as a match for the current stimulus, its action list is browsed, and the action that leads to the most favorable scenario is chosen. Each scenario is evaluated according to its DV (the positive component means +1, the negative -1, and the "deadly" -10; the heuristic to be minimized is the sum of the values of the corresponding activated components). If no match is found, or there is no action list, the agent acts accordingly to a

[1]Note that this is an implementation of an indexing mechanism raised in the section 3.5.

[2]The Hamming distance is the number of pixels differing between the two images. See [54] for a definition and related issues.

built-in DV action map (negative or "deadly" leads to a `reject`, positive to an `accept`, and `none` otherwise).

After the agent action, the feedback stimulus is applied to the architecture, and the resulting memory frame is stored in the main memory. Furthermore, the action list of the original stimulus frame (before the action be performed) is updated/set, pointing to the feedback frame. Next time the agent faces a similar situation where this frame is recalled, it will know what to expect from the corresponding action.

An illustrative experiment will be presented below, consisting on a sequence of stimuli. In the following screenshots, green pixels are denoted by (▦), and red pixels by (▩). Prior to the agent first stimulus, the memory is blank. The first stimulus (figure 4.5) consists in a smiling face silhouette with some green pixels (a perceptually positive DV). The agent uses the perceptual assessment indicating an `accept` action. The environment responds with a all-green face (*i.e.,* positive DV). The corresponding association is formed and stored in memory.

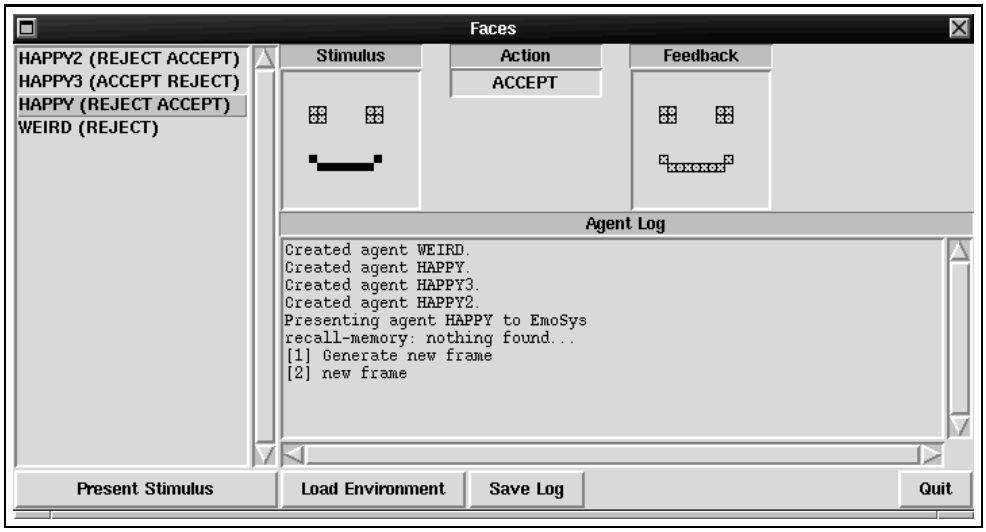

Figure 4.5: Screenshot of the `faces` implementation: a smiling face with some green pixels.

Next, a colorless face, which is similar to the first one, is presented (figure 4.6). The agent recalls the previous association, and chooses to `accept` the stimulus. However, if this stimulus were presented without the former association, the action would be `none` — the stimulus would be mapped by the perceptual layer to a null DV.

An interesting result is obtained when now, a similar face is shown, containing some red pixels (figure 4.7). In this case, the recalled association is

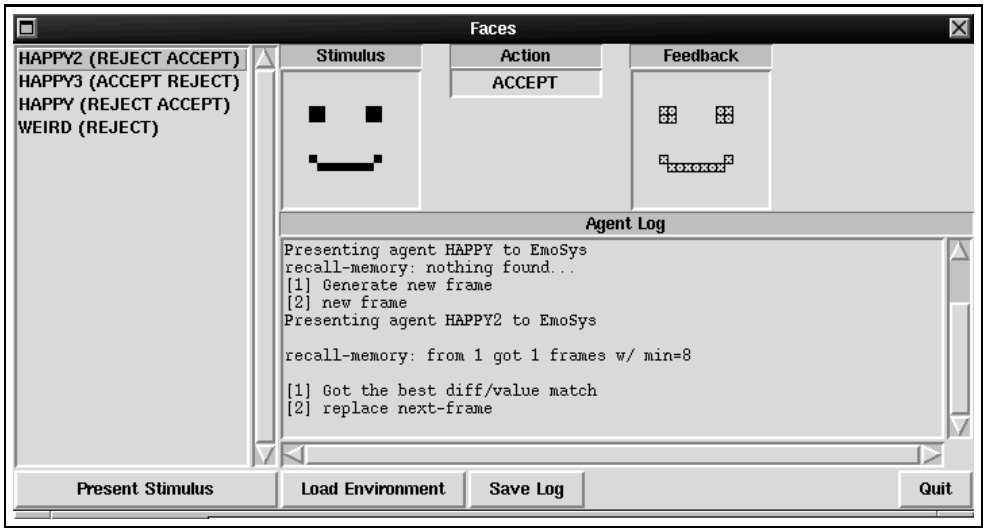

Figure 4.6: Screenshot of the `faces` implementation: a similar smiling face all in black.

used to *override* the perceptual impulse to `reject` the stimulus. This case illustrates the role of the cognitive layer in providing a more refined response, than the basic perceptual one. Using the same line of reasoning, if this stimulus were shown prior to the first of the sequence, the agent would `reject` it.
At last, a different face is shown (with some red pixels, figure 4.8), and unlike the previous stimulus, because this face is "unknown" to the cognitive layer, the action is `reject`, following the perceptual negative assessment.

Other experiments were performed with the architecture, showing further interesting results. For instance, if the acceptance of the stimulus of the figure 4.5 had a negative response (e.g., a very "red" face), next time that same stimulus were presented, the agent would choose another action. When the action resulting from a given stimulus is answered with a negative response, the agent will not repeat the mistake — other actions are "tried" in a search for a better response. The frame that this action points to has a negative DV, making the agent to avoid it.

The role of the built-in knowledge in this implementation stands out very clearly. The mechanism that is behind the agent behavior facing environment stimuli, is encoded in the perceptual layer. Namely in the perceptual mapping between stimuli and the DV. It is on top of this layer that the cognitive layer works. When the simplicity of the perceptual layer is not sufficient to cope with a complex environment, the cognitive one jumps in, providing the "knowledge" gained from past experience.

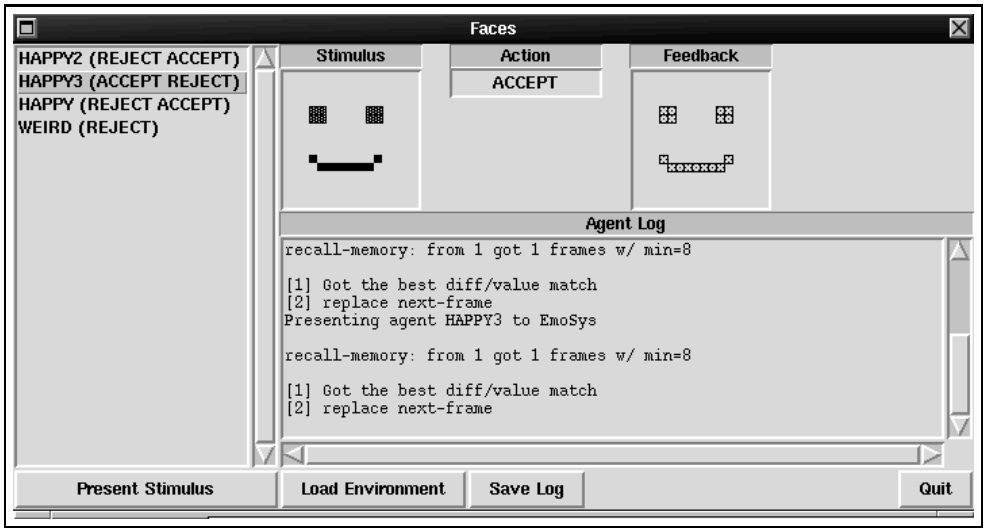

Figure 4.7: Screenshot of the `faces` implementation: a similar smiling face but with some red pixels (the "eyes").

4.3 Implementation: `decks`

The aim of the `decks` implementation is to reproduce the results of the deck game (section 2.3, figure 2.7) described by Damasio ([18] page 212), using the proposed model. In particular, to obtain the two divergent results of the normal and frontal patients (*i.e.,* with damaged frontal lobes), allowing the agent to use or not the marking mechanism (association).

In a simplified version of the original game [3, 4], decks A and B usually give $100 except for a few cards that make the player lose -$1250, while decks C and D usually give a lower value of $50 where there are more frequent losses of -$250. The net profit of decks A and B is negative, while decks C and D provide a positive one.

In terms of the implementation, the environment is episodic, with an environment feedback phase. First, four stimuli are simultaneously presented to the agent (four symbols, corresponding to the four decks: *A, B, C,* and *D*). The agent action is simply the choice of a deck. The environment responds with the amount of money gained/lost. Each stimulus encompasses a pair of card deck symbol and money amount gained (negative, if lost). In the first phase, the second components of all stimuli are null (the card amount is obviously hidden). Only after the action the reward associated with the chosen card is revealed. The perceptual layer only extracts the money amount (the perceptual image), while the cognitive layer extracts the symbol. There is no point in including more complexity in the cognitive image, the symbol

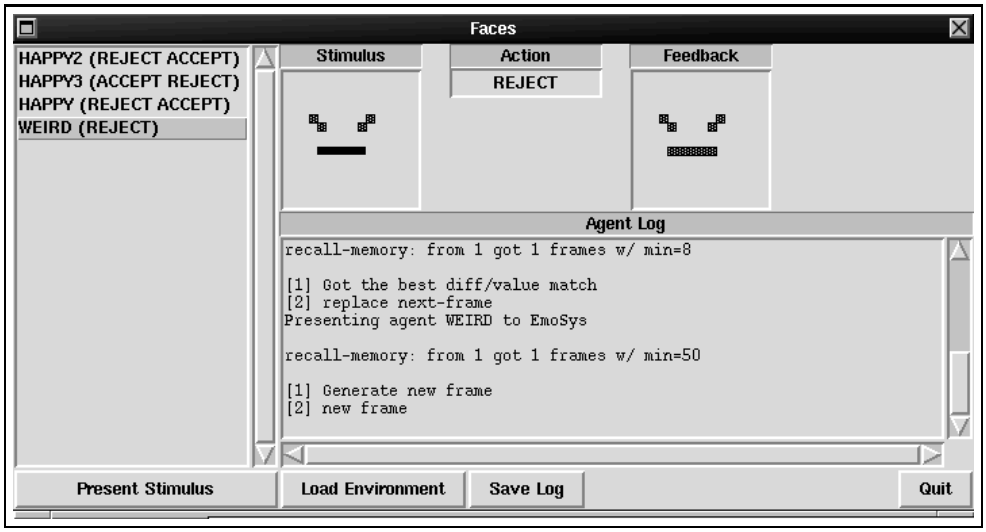

Figure 4.8: Screenshot of the `faces` implementation: a distinct face with some red pixels.

suffices to the proposed objectives. Furthermore, since the environment is very simple (only four distinct decks), the symbolic representation is enough to identify each deck (however, the number of decks is not hardwired in the agent program). The DV has only two (boolean) components, one for positive and other for negative assessment of the deck. The mapping between the perceptual image and the DV activates the positive component if the amount greater than zero, or the negative component when it is less than zero.

The model of this implementation is identical the one represented in figure 3.1. An important innovation with respect to the previous two implementations is the adaptability of the perceptual layer. Both kinds of learning are implemented: the cognitive event-based learning, and the perceptual mapping-based learning. When the agent is faced with the four decks, the perceptual layer is able to give an immediate assessment of the desirability of each deck, while the cognitive layer browses the memory for past events associated with each deck. With all this information in the working memory, the agent decides which deck to choose.

The working memory is organized in clusters of frames. Each cluster corresponds to a specific deck, and contains the input stimulus (the deck symbol only), the perceptual frame (the expected perceptual image and the expected DV, or in other words, the expected amount of gain/loss), and the frames recalled from memory (obtained by the cognitive layer). When each frame is complete, a representative frame is chosen for each cluster. Then, all

the clusters with a negative DV are rejected, and a deck is randomly chosen from the remaining ones. In fact, the perceptual value is used to weight this random choice, in order to make the agent prefer higher value cards. But if all clusters are rejected, then the action is randomly chosen from all the available decks, also using a weight factor.

After choosing the deck, the environment responds with a feedback stimulus, now containing not only the symbol of the deck, but also the amount of money gained/lost. This information is used to update the perceptual map (according to a learning rate), and to add the frame to the main memory, associating the cognitive and the perceptual images, along with the DV (mapped from the perceptual image, *i.e.,* the amount of money). This perceptual image can be interpreted here as the expected gain. In the perceptual layer learning, the update rule of this expected value is simply:

$$V'_m = \theta V_p + (1 - \theta) V_m \tag{4.5}$$

where the new memory frame expected value V'_m is interpolated between its former value V_m and the feedback value V_p, using the learning rate θ.

In order to simulate the behavior of the frontal patients playing this game, the agent was prevented from recalling memory frames. Then, the perceptual layer was left alone to decide which deck to choose, preferring the decks A and B, because of the most frequent \$100 cards. As an example, setting the learning rate parameter to $\theta = 0.001$, the obtained results, shown in figure 4.9, are clearly similar to the Damasio experiments results of figure 2.7.

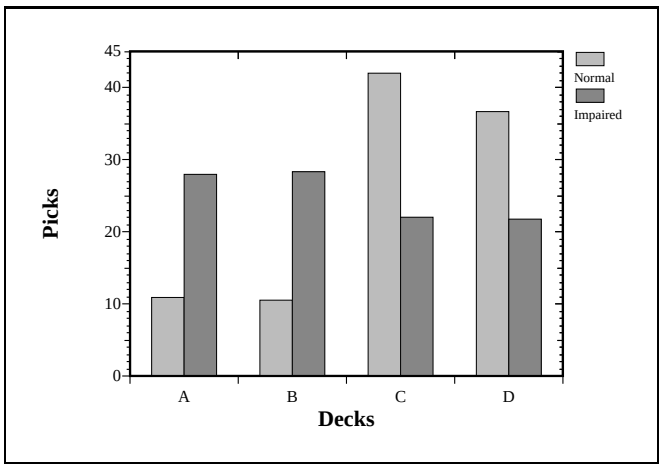

Figure 4.9: Results from the `decks` implementation. The average number of picks for each deck is shown. The average was taken over 200 experiments of 100 turns each. The θ parameter was set to 0.001.

These results illustrate the distinct natures of the learning process performed by each layer. But they are not to be considered separately. Although the perceptual layer is able to work by itself, the same cannot be said about the cognitive layer. This is because the cognitive layer uses the perceptual representation, in order to contribute to an overall enriched behavior.

> There can be no transforming of darkness into light and of apathy into movement without emotion.
>
> Carl Gustav Jung

> The heart has its reasons which reason knows nothing of.
>
> Blaise Pascal, *"Pensées"*

Chapter 5

Conclusion

Summary

The first section of this chapter addresses the question of what the consequences of this model are, both in the conceptual and experimental sides. Next, a set of open issues is presented, which could not be answered in the context of research. Finally, future directions of this work are elaborated and discussed.

The goal of this thesis is twofold: the discussion of a model for emotion-based agents, raised from neurophysiological findings, and the presentation of some experimentation of the model. The formulation of the model is still in an early development stage.

5.1 Consequences

From this thesis, some consequences of the model can be discussed. The first consequence is the ability to *associate* two levels of representation: a complex (cognitive), and a basic (perceptual) one.

The *built-in* part of the perceptual layer is essential to allow the agent to *bootstrap*. It can be said that the basic, irreducible goals (objectives, motivations, desires, and so on) of the agent are implicitly encoded in this built-in. Of course it cannot account for all the goals that can be identified with the agent behavior, they form only the basic ones.

Another consequence of the model is its ability to provide *relevance*. Only the stimuli that bring about (directly, by memory matching, or by other higher-level associations) a non-null DV are relevant, in the sense that they mean something to the agent. The agent is insensitive to all others. Note that

this does not discard *curiosity*, since an unknown stimulus may elicit a DV component that would make the agent "explore" it, *i.e.,* to adopt a course of action that would allow the agent to interact with that origin of the stimulus. It just discards uninteresting aspects avoiding a cognitive saturation.

Provided that the agent is able to decide appropriately about certain stimuli through the perceptual layer (*i.e.,* the mapping from stimuli to the DV) alone, then these stimuli can be said to be *meaningful* to the agent[1]. But since the model is able to associate stimuli irrelevant for the perceptual layer to certain DV instances, these new associations can be also said to provide *meaning,* and of a more sophisticated nature than the former. Recalling John Searle's argument that machines cannot understand [51], raising the Chinese Room metaphor, a confrontation with the model can be attempted. Searle's argument is that as long as he has a sufficiently complete formal rule-book, he is able to answer any question formulated in Chinese about any given Chinese story. From outside the room, it can be said that an understanding of Chinese is accomplished, when the man inside the box does not understand a word of Chinese at all. This is so because the man inside the box just manipulates symbols which are meaningless to him, according to the provided rule-book. But now, imagine that some symbols cease from being senseless, and can be identified with some basic built-in associations. For instance, some symbols become colored, where colors now do mean something to the man inside: red means bad, green means good, blue means important, and so on. Although this is far from helping him to derive a *syntax,* some *semantic* content can already be grasped, at least in a basic level, provided that the coloring scheme is coherent. At a first level, some symbols become *meaningful,* even if from some point on, the coloring stops. The man is able to remember previous colorings, and recall them when faced with monochromatic writing. At an another level, by association, the man becomes able to assign meaning to other symbols, for instance, if some symbol always appears close to a red-colored one, in the stream of text. Applying the model to this metaphor, the colors can be understood as the built-in part of the perceptual layer, that activates certain DV components, depending on the color of a symbol. A complex meaning scheme can be erected by the cognitive layer, driven by associations with DV instances. In this sense, the model can be thought of as a *meaning engine* [63].

Focusing now on the double-processing paradigm, two consequences can be extracted. First, an efficient way of looking up cognitive matches, by the

[1]See [37] for a discussion on whenever a stimulus is "meaningful" to a machine. In this paper, McCarthy states that when a change in the room temperature makes the thermostat switch correctly, it can be said that the temperature change was a *meaningful* stimulus to it.

means of simpler representations provided by the perceptual layer. Since this latter representation is assumed to be extracted very quickly, it can guide the search for cognitive matches. Since the content of the agent memory can grow drastically when facing a complex environment, this perceptual guidance can help in narrowing the choices and avoiding an exhaustive search for a match. This scheme was actually used in the `faces` implementation (see section 4.2). A second consequence is the "expert discrimination" feature, resulting from the ability to differentiate subtle differences, using the finer cognitive matching mechanism. The discussion of `damasio` implementation (section 4.1) shows some illustrative experimentation.

But does it make sense to state that the model provides "meaning," or "relevance," or "understands" whatsoever? Recalling McCarthy's argument on ascribing mental qualities to machines [37], it does. One cannot hope for some invisible magic to attain such mental qualities. Once they help describing the model's proprieties, while it "expresses the same information about the machine that it expresses about a person" [37], it seems reasonable to ascribe them to this model.

After all, does it make any sense to ask "where are the emotions and feelings after all?" One can now resort to an analogous line of argumentation as in the previous paragraph. Accepting McCarthy's argument to ascribe to machines the mental qualities referred to by him[2], why not extend the concept to the terms "emotion" and "feeling?" Damasio distinguishes emotion from feeling as the latter requiring consciousness. This model does not address consciousness. The consciousness issue is far from being understood, either its physiological roots or philosophical description [31]. There seems to be no agreement on this matter. So the discussion whether an agent is conscious is put aside. Emotions however have a more concrete grounding. And the answer to the question "where are the emotions in this model?" is correlated with the role of the perceptual layer. The basic meanings provided by the DV *are* indeed the agent emotions.

This follows not only from the model grounding in the way emotions are described at a neurophysiological level, but also from the behavior attained by it: certain stimuli are able to directly elicit a response from the DV, other stimuli elicit it indirectly by the means of stored associations, and the cognitive layer is able to block some primary perceptual layer responses. The first two are what Damasio calls *primary,* and *secondary emotions* [18], and the latter what LeDoux refers to as regulation of the rage, when discussing animals with their cortex removed ([34] page 80)):

[2]McCarthy refers explicitly to the terms beliefs, knowledge, free will, intentions, consciousness, abilities, and wants [37].

> Yet, the emotional behavior of decorticate animals (animals in whom the cerebral cortex was removed) was not completely normal. These creatures were very easily provoked into emotional reactions by the slightest events. They seemed to be lacking any regulation of their rage, which suggested the cortical areas (like Plato's charioteer) normally rein in these wild emotional reactions and prevent their expression in inappropriate situations.

Although the implementations presented in this thesis do not yet show a behavior clearly identifiable with such proprieties as emotions and understanding, they do implement some aspects of the model and show some interesting results. Namely, the `damasio` implementation showed the basic association mechanism and how the cognitive representation can help providing expert discrimination. The `faces` implementation presented some experiments on assigning meaning to complex representation as visual images, being able to remember past associations. Finally, the `decks` implementation proposed itself to replicate a Damasio experiment to show the role of the somatic marker, essential to the secondary emotions.

5.2 Open Issues

The presentation of the model still leaves several open issues. Namely, a formalization of the model components and the way they interact should be performed.

For instance, the lack of a clearer definition of the perceptual layer, namely what the DV components shall address, given an environment and a purpose to the agent. Of course the characterization of the environment does not suffice to specify the perceptual layer. The goals and motivations of the agent can be viewed as being encoded in the stimulus-DV mapping. But the nature of this encoding has to be explored.

Many issues pertaining to the manipulation of the working memory, as well as the process of deriving a decision from there remain to be researched.

The possibility of generating actions from the cognitive layer, and the way they are orchestrated with the perceptual generated ones is also open to further development.

There are some barriers that stand between the model and a real-world implementation. It is necessary to bring the model out from the simple episodic environments used in the presented implementations. A major step would be to put the model working in a real robot, moving around a lab, interacting with the objects it finds. But when trying to bridge this gap, one faces the problem of representation. The model needs to have a spatial representation, right in the basic and built-in perceptual layer. It needs to

have mechanisms to isolate objects (assuming vision as its primary sensor), and to assign meaning to them.

5.3 Research Directions

It is important to stress that this thesis presents a snapshot of the research on this model. There are several ways this research can evolve. The open issues outlined in the above section give some ideas. In this section, three main research directions are presented.

First, the application of the model to a physical robot is a very promising path for a number of reasons. First, to force the development of the model to handle non-episodic real-world environments. This does not mean that the internal workings of the model will not be episodic. The challenge is to adapt its structure to an environment that is not presented in an episodic fashion. Furthermore, the robotic platform raises a myriad of issues: there are several things happening at the same time, unexpected events (e.g., collisions), encounters with unknown objects, robustness to mechanical failures, and so on. In particular, the RoboCup [48] may provide an ideal environment to put these ideas into practice [59].

As it was said in the last section, such environments raise questions of representation. The model is specially suited to handle image-like representations. Although symbolic systems have reached a high degree of sophistication, the same cannot be said about spatial representations. The area of diagrammatic reasoning [27] can have a lot to offer to this model. And in the case of the robotic platform, this means spatial representation of the surrounding environment. This is essential in order the model to be able to isolate objects of interest (visually perceived), to consider their spatial relationships, and to interact with them in the physical environment.

Finally, after gaining a more mature understanding of the model, provided for instance the reaching of the above goals, it is essential to make a step towards a formalization of the model: to define precisely each component, and the way they all interact to form a whole. As well as gather formal tools to assert what aspects shall be cognitive and perceptual, the minimal richness that the DV shall provide, and so on. The ultimate objective of such research would be to formalize a framework that would allow, in a systematic way, to apply the model to a given specified environment, and to evaluate its performance.

A topic of world-shaking importance, yet dealt with facetiously; an android trait, possibly, he thought. No emotional awareness, no feeling-sense of the actual *meaning* of what she said. Only the hollow, formal, intellectual definitions of the separate terms. [author's emphasis]

Philip K. Dick, "Do Androids Dream of Electric Sheep?"

Dave, my mind is going! I can feel it! I can feel it!

"2001: A Space Odyssey"

Bibliography

[1] Zippora Arzi-Gonczarowski. Wisely non-rational — a categorical view of emotional cognitive artificial perceptions. In Dolores Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 7–12, 1998.

[2] 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.

[3] Antoine Bechara, Antonio R. Damasio, Hanna Damasio, and Steven W. Anderson. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50:7–15, 1994.

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

[5] Bernard Berelson and Gary A. Steiner. *Human Behavior: An Inventory of Scientific Findings*. Harcourt, Brace & World, 1964.

[6] Luis Miguel Botelho and Helder Coelho. Adaptive agents: emotion learning. In Dolores Cañamero, Chisato Numaoka, and Paolo Petta, editors, *Workshop: Grounding Emotions in Adaptive Systems*, pages 19–24. SAB'98: From Animals to Animats, August 1998.

[7] Jack Breese and Gene Ball. Bayesian networks for modeling emotional state and personality: Progress report. In Dolores Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 37–42, 1998.

[8] 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.

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

[10] Alastair Burt. Emotionally intelligent agents: The outline of a resource-oriented approach. In Dolores Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 43–48, 1998.

[11] Dolores Cañamero. Emotional and intelligent: The tangled knot of cognition. 1998 AAAI Fall Symposium Technical Report FS-98-03, AAAI, 1998.

[12] Dolores Cañamero, Chisato Numaoka, and Paolo Petta, editors. *Workshop: Grounding Emotions in Adaptive Systems*. SAB'98: From Animals to Animats, August 1998.

[13] Insook Choi. From motion to emotion: synthesis of interactivity with gestual primitives. In Dolores Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 61–67, 1998.

[14] Patricia S. Churchland and Terrence J. Sejnowski. *The Computational Brain*. The MIT Press, 1992.

[15] Scriptics Corporation. Scriptics — the tcl platform company. URL: `http://www.scriptics.com`, 1998.

[16] Mark Coulson and Simon Duff. Feedback loops in expression and experience: Emotion as cause and effect. In Dolores Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 68–69, 1998.

[17] Hartvig Dahl and Virginia Teller. What AI needs is a theory of the functions of emotions. In Dolores Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 70–75, 1998.

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

[19] Hanna Damasio. Uncovering neural systems behind words and concepts. Oral communication on The Foundations of Cognitive Science at the End of the Century, CCB, Lisbon, Portugal, May 1998.

[20] Ronald de Sousa. *The Rationality of Emotion*. The MIT Press, 1987.

[21] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3/4):169–200, 1992.

[22] Cynthia Breazeal (Ferrell). Early experiments using motivations to regulate human-robot interaction. In Dolores Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 31–36, 1998.

[23] Cynthia Breazeal (Ferrell). A motivational system for regulating human-robot interaction. In *Proceedings of AAAI-98*, pages 54–61. AAAI, 1998.

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

[25] Sandra Clara Gadanho and John Hallam. Emotion-triggered learning for autonomous robots. In Dolores Cañamero, Chisato Numaoka, and Paolo Petta, editors, *Workshop: Grounding Emotions in Adaptive Systems*, pages 31–36. SAB'98: From Animals to Animats, August 1998.

[26] 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.

[27] Janice Glasgow, N. Hari Narayanan, and B. Chandeasekaran, editors. *Diagrammatic Reasoning*. AAAI Press / The MIT Press, 1995.

[28] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley, 1988.

[29] Paul Graham. *ANSI Common Lisp*. Prentice-Hall, 1996.

[30] Bruno Haible, Michael Stoll, and Marcus Daniels. CLISP: common lisp implementation. URL: `http://clisp.cons.org`, 1997.

[31] Stuart R. Hamoeroff, Alfred W. Kaszniak, and Alwyn C. Scott, editors. *Toward a Science of Consciousness*. The MIT Press, 1996.

[32] Mardi Jon Horowitz. *Image Formation and Cognition*. Appleton-Century-Crofts, 1970.

[33] Guy L. Steele Jr. *Common Lisp the Language*. Digital Press, second edition, 1990.

[34] Joseph LeDoux. *The Emotional Brain*. Simon & Schuster, 1996.

[35] Jonathan Liu. A formal interpretation of the concept of emotion. In Dolores Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 116–121, 1998.

[36] John McCarthy. Programs with common sense. In *Mechanisation of Thought Processes, Proceedings of the Symposium of the National Physics Laboratory*, pages 77–84, London, U.K., 1958. Her Majesty's Stationery Office.

[37] John McCarthy. Ascribing mental qualities to machines. URL: `http://www-formal.stanford.edu/jmc/`, 1979.

[38] John McCarthy. Making robots conscious of their mental states. URL: http://www-formal.stanford.edu/jmc/consciousness-submit/consciousness-submit.html, 1995.

[39] Incorporated Merriam-Webster. Merriam-webster online. URL: `http://www.m-w.com/home.htm`, 1998.

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

[41] Linux Online. Linux online. URL: `http://www.linux.org`, 1998.

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

[43] John Ousterhout. *Tcl and the Tk Toolkit*. Professional Computing Series. Addison-Wesley, 1994.

[44] Jean Piaget and Bärbel Inhelder. *The Psychology of the Child*. Basic Books, Inc., 1969.

[45] Rosalind Picard. *Affective Computing*. MIT Press, 1997.

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

[47] 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.

[48] RoboCup. Robocup online. URL: `http://www.robocup.org/`, 1998.

[49] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Series in Artificial Intelligence. Prentice Hall, February 1995.

[50] K. R. Scherer. Appraisal theory. In T. Dalgleish and M. Power, editors, *Handbook of Cognition and Emotion*, pages 637–663. Wiley, Chichester, 1999.

[51] J. R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–457, 1980.

[52] Aaron Sloman and Monica Croucher. Why robots will have emotions. In *Proceedings IJCAI 1981*, June 1981.

[53] 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.

[54] Shimon Ullman. *High-level Vision.* The MIT Press, 1996.

[55] 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.

[56] Juan D. Velásquez. A computational framework for emotion-based control. In Dolores Cañamero, Chisato Numaoka, and Paolo Petta, editors, *Workshop: Grounding Emotions in Adaptive Systems*, pages 62–67. SAB'98: From Animals to Animats, August 1998.

[57] Juan D. Velásquez. Modeling emotion-based decision-making. In Dolores Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 164–169, 1998.

[58] Juan D. Velásquez. When robots weep: Emotional memories and decision-making. In *Proceedings AAAI-98*, pages 70–75. AAAI, AAAI Press and The MIT Press, 1998.

[59] Rodrigo Ventura, Pedro Aparício, Pedro Lima, and Carlos Pinto-Ferreira. Socrob — a society of cooperative mobile robots. In *Proc. of 1998 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 1998.

[60] Rodrigo Ventura, Luís Custódio, and Carlos Pinto-Ferreira. Artificial emotions — goodbye mr. spock! In *Progress in Artificial Inteligence, Proceedings of IBERAMIA'98, Lisbon, Portugal*, pages 395–402. Colibri, 1998.

[61] Rodrigo Ventura, Luís Custódio, and Carlos Pinto-Ferreira. Emotions — the missing link? In Dolores Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 170–175, 1998.

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

[63] Rodrigo Ventura and Carlos Pinto-Ferreira. Meaning engines — revisiting the chinese room. In Dolores Cañamero, Chisato Numaoka, and Paolo Petta, editors, *Workshop: Grounding Emotions in Adaptive Systems*, pages 68–70. SAB'98: From Animals to Animats, August 1998.

[64] Rodrigo M. M. Ventura and Carlos A. Pinto-Ferreira. Problem solving without search. In Robert Trappl, editor, *Cybernetics and Systems '98*, pages 743–748. Austrian Society for Cybernetic Studies, 1998. Proceedings of EMCSR-98, Vienna, Austria.

[65] Elias Vyzas and Rosalind W. Picard. Affective pattern classification. In Dolores Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 176–182, 1998.

[66] Ian Paul Wright. *Emotional Agents*. PhD thesis, Faculty of Science of the University of Birmingham, February 1997.