Emotion Understanding: Robots as Tools and Models

Lola Cañamero
Adaptive Systems Research Group
Department of Computer Science, University of Hertfordshire
College Lane, Hatfield, Herts AL10 9AB, UK
L.Canamero@herts.ac.uk

Philippe Gaussier
Neurocybernetics Team
Image and Signal Laboratory, CNRS UMR 8051
University of Cergy-Pontoise, ENSEA
6 av du ponceau, 95014 Cergy Pontoise, FRANCE
gaussier@ensea.fr

1   Is There a Place for Emotions in Robotics?

To many, the title of this chapter might sound like a complete nonsense — after all, what do emotions have to do with robots? How could robots possibly tell us anything about them, and more generally about us? Science fiction has typically portrayed robots as prototypes of the unemotional, and this vision largely prevails in science as well as for the layperson, in spite of the proliferation of "affective" toys and robots. Admittedly, many of these "affective devices" are little more than "bags of tricks" aimed at entertaining the young (and not so young) and technically-oriented sectors of the population, rather than pieces of scientific research trying to take a step towards a broader understanding of what emotional phenomena can be; yet these devices can be extremely efficient at achieving their goal, offering us good opportunities to reflect on the human tendency to anthropomorphize even with dealing with the dullest technology (Reeves and Nass 1996) and to perceive the world as emotionally colored. However, shedding light on human perception of affect is not the only — nor the main — contribution that robots can make to the study of emotions. Artifacts — computers or robots — can also be valuable tools to support scientists in their investigation of human emotion and, more fundamentally, they permit to model and test hypotheses about affective phenomena. The investigation of all these issues is the domain of a new research area generally known as "affective computing," which departs from and complements other disciplines traditionally concerned with the study of emotions in important ways, as Rosalind Picard points out (1997, page 3):

"This is different from presenting a theory of emotions; the latter usually focuses on what human emotions are, how and when they are produced, and what they accomplish. Affective computing includes implementing emotions, and therefore can aid the development and testing of new and old emotion theories. However, affective computing also includes many other things, such as giving a computer the ability to recognize and express emotions, developing its ability to respond intelligently to human emotion, and enabling it to regulate and utilize its emotions."

In this chapter, we focus on autonomous robots, as opposed to other artifacts such as computers, to sketch our views on how they can be meaningful tools and models for the study of emotion. Autonomous robots are especially interesting since they are physical devices moving around, making decisions, solving problems, and interacting in the real world   the same world humans and animals inhabit; therefore, they are particularly suited to study two fundamental aspects of emotions:

- Their role as mechanisms for adaptation to the environment and behavior control.
- Their role in social interaction and communication.

Robotic systems offer the possibility to approach these issues in different ways. On the one hand, models of some emotion "components" can be explicitly included in the architecture of the robot to give rise to behavior that appears to arise from an emotional system. On the other hand, autonomous robots allow us to investigate how emotions can be an emergent property of a dynamical system, without having to model an "emotional system" explicitly. In this latter case, the lack of an explicit emotional system does not imply that emotions are considered as phenomena that only exist in the eye of the beholder; on the contrary, we believe that understanding what kinds of problems can be solved without the involvement of an emotional system will help us delineate the functions of emotions in cognitive processing. We will therefore present some of the problems that autonomous robots currently tackle without any kind of emotion model, before discussing some cases in which it might make sense to bring emotions into play.

2 Some State-of-the-art Problems in Behavior-Based Robotics

As we mentioned previously, most of the state-of-the-art work in autonomous robotics does not use any emotional mechanism at all. Let us first briefly consider the design philosophy underlying autonomous robots, and then provide some illustrative examples of typical problems they have to solve.

2.1 The behavior-based paradigm in robotics

Robot architectures are based on the idea that real time reaction capabilities are crucial for good robot/environment interaction, in particular when the environment is dynamic or *a priori* unknown. To take into account the effect of the interactions with the environment, most robotic architectures use some kind of behavior-based approach (Brooks 1991, Arkin 1998). In this paradigm, which draws on ideas from biology, ethology and neuroscience, the control architecture is divided into different loops of processes running in parallel and corresponding to the different kinds of behaviors[1] and time scales the robot has to manage. For instance, it is important to introduce a fast reaction mechanism allowing robots to avoid obstacles while they are moving. At the same time (i.e., in parallel with this activity) other processes are being executed "independently," for example information from images can be analyzed to control much more complex activities like homing, object grasping or planning. All these processes can issue different motor commands, some of which might be incompatible and "compete" for the same actuator for their execution. Merging this asynchronous flow of parallel information can be performed via a prioritization mechanism.

As an example, in the subsumption architecture (Brooks 1986), one of the earliest types of behavior-based architectures, behaviors are organized in layers running in parallel, each of them with its own perception and action capabilities, and "loosely coupled," i.e., with minimal interaction limited to the inhibition of the (perceptual) input or the suppression of the (motor) output of one layer by another. Using these inhibition and suppression mechanisms, layers are arranged in a hierarchy with fixed priorities that reflect the importance that each layer (behavior) carries for the survival of the robot in that particular environment, and therefore determine which behavior(s) to execute given the presence of relevant stimuli in the environment. This allows for instance to decide that going to a planned position is less important for the survival of the robot than avoiding an important danger. In this case, the "warning" of a danger will suppress the output of the planning system. Therefore, the decision making process is completely decentralized and much more robust than the sequential "perception-reasoning-action" process used in classical artificial intelligence and robotics, since any low-level process can take over the control of the robot's behavior if necessary. The behavior of the robot can also be highly opportunistic.

Many other behavior-based architectures and "decision making" (or action selection) strategies can be used (see Arkin 1998, Pfeifer and Schreier 1999 for an overview), some of them including some sort of "internal goals" or "motivations" to drive the behavior of the robot, as we will see later. All of them share the basic philosophy of using parallel, loosely coupled perception-action loops, decentralized decision making, action-oriented perception, and lack of (or use of very minimal, coarse and distributed) representations. This way of designing control architectures attempts to avoid the symbol grounding problem (Harnad 1990) and other difficulties of the traditional approach (see Pfeifer and Schreier 1999).

---

[1] The term "behavior(s)" is used in a technical sense in behavior-based robotics to denote different competencies or activities that the robot is able to perform in its environment.

2.2 Sensory-motor and reinforcement learning

Let us suppose that an autonomous robot has to learn how to return to a particular location. If the robot is equipped with a visual mechanism to recognize particular objects in the visual scene, it can use this mechanism to localize where different objects are in a room. Hence, we can design a group of neurons to learn that a given location is characterized by a particular conjunction of landmarks and azimuths. When the robot is near the learned location, the mismatch between the learned configuration and the current situation will be very low. On the contrary, when the robot is far away from the learned location the mismatch will be high (Gaussier and Zrehen 1995, Gaussier *et al*. 1999). This place recognition mechanism can be used to build a homing behavior. The robot has to learn at least three locations/actions in the neighborhood of its home in order to reach it (see Figure 1).
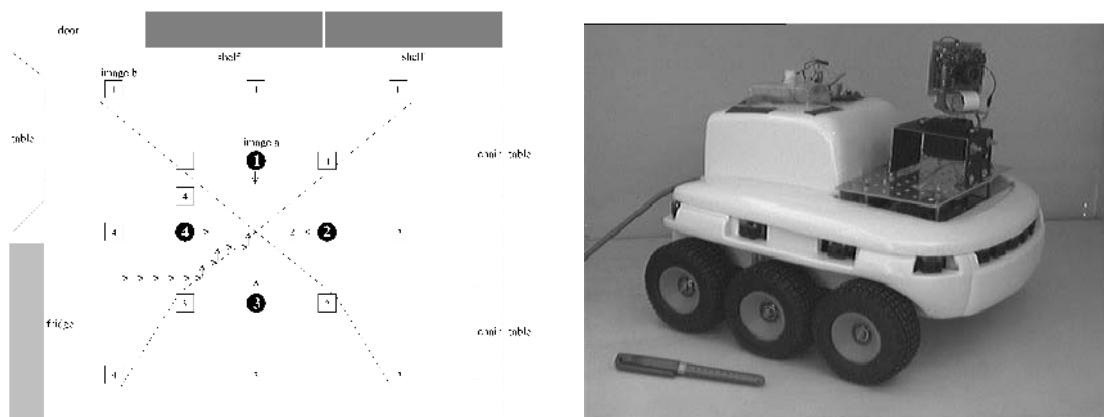

Figure 1. Left: Place cell-like segregation of space in a robotic experiment. Four panoramic images are learned (circles), others are associated with one of these learned panorama (boxes). As we can see, if the robot learns to reach the cross from each learned panorama it can reach the cross from any other panorama associated with the learned one (generalization). The set of arrows represents an example of robot trajectory (see Gaussier *et al*. 1999 for details). **Right**: Photo of the Koala robot used in the visual navigation experiments.

Such an architecture consists in a fairly direct coupling of perception and action, as shown in Figure 2. When the robot is facing its home base, it can learn to recognize the landmark-azimuth configuration and associate it to the direction of motion in order to reach its home (using a simple conditioning rule).
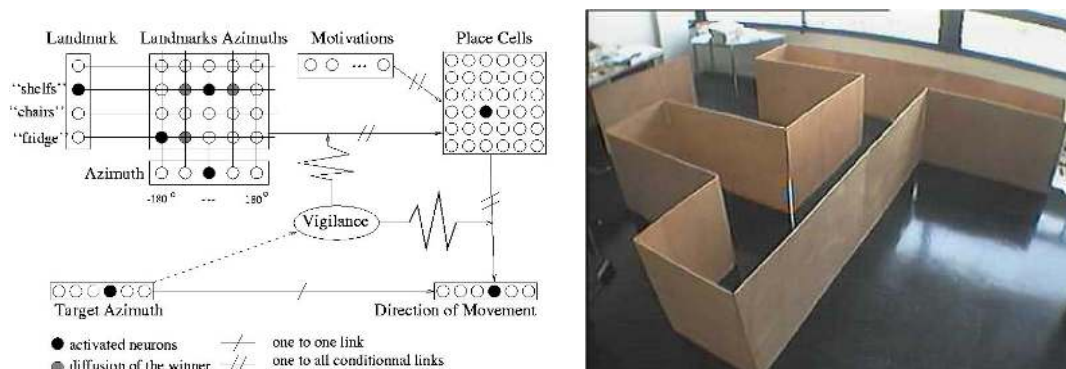

Figure 1. Left: Schematic representation of the PerAc block. From the perceived situation, the reflex system extracts information to control directly the actions. Concurrently, the recognition system learns

sensory input patterns and how to link them to actions by associative or reinforcement learning. The system adapts itself dynamically to the environment. **Right:** An example of maze learning where a delayed reward signal was used to learn to associate particular signs in the maze with an action allowing to reach a "goal" location where the reward was received.

If the robot is far from its home base—and supposing it cannot see it—it tries to recognize its current location. Due to the generalization capabilities of the place recognition system, the "winner-takes-all" mechanism used for the place recognition group will activate the neuron associated to the learned place that is nearest to the place where the robot is current located. This neuron will then trigger the learned motor action and the robot will move in a direction that decreases its distance to its home base. After several movements, the robot will enter in an area where another place cell wins, triggering another movement that will allow it to move in a more appropriate direction to reach its home. Hence, Perception-Action learning can be seen as a way to create an attraction basin: our robot is no more than a ball falling in the learned attraction basin.

Now, if we want our robot to move between places, for example going from one place where it can drink to another place where it can eat, using a simple Hebbian learning algorithm we can associate the place cells around each goal to a particular "motivation" or internal variable; the motivation and the place cell are supposed to be simultaneously active during the learning phase to allow an easy association. This motivation signal acts as a bias in the competition for place recognition   it changes the shape of the attraction basin and instead of going to the nearest learned goal, our robot becomes capable of going to another goal (location) due to the modification in the shape of the attraction basin (see Figure 3).
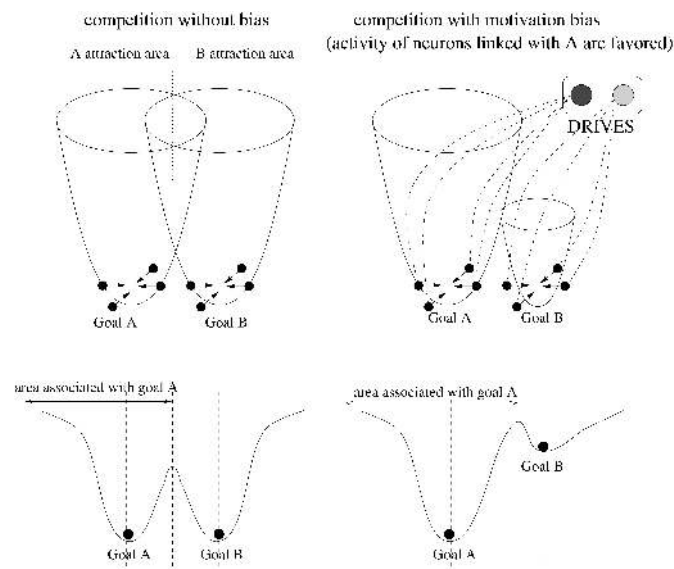

Figure 2: Bias introduced by motivations in goal selection.

In the case of an open environment, simple Hebbian learning is sufficient to build interesting behaviors. Obviously, if we want our robot to avoid a particular area because it is associated to a negative reward, it must be able to avoid the ongoing action. This learning needs to use the negative reinforcement signal as a modulator

factor of the Hebbian learning   in this case, to learn an inhibitory link between a place cell and the ongoing action. Other kinds of reinforcement learning rules (Barto *et al.* 1983, Watkins and Dayan 1992, Weaver *et al.* 1993, Gaussier *et al.* 1997) can be used to associate the recognition of any perceived information to an arbitrary action according to a reward signal that can be delayed in the most complex cases   for instance the robot is only rewarded if it succeeds in performing a particular sequence of actions (see Fig 2).

2.3   Action selection, planning and internal values

In more complex cases, for instance when the robot has learned several actions in a given place and has to choose one of them according to a particular goal, more complex mechanisms must be used. One solution consists in learning a cognitive map of the environment (Tolman 1948, Revel *et al.* 1998). This learning can be performed without any reinforcement signal (latent learning) since it relies on the learning of some kind of environment causality: "from here, I can go there." The almost co-activation of a departure node and an arrival node allows to reinforce the strength of the connections between these two nodes. After a while, the robot builds a graph or a map coding for the topological relationship between the different places it knows in its environment (see Figure 4).
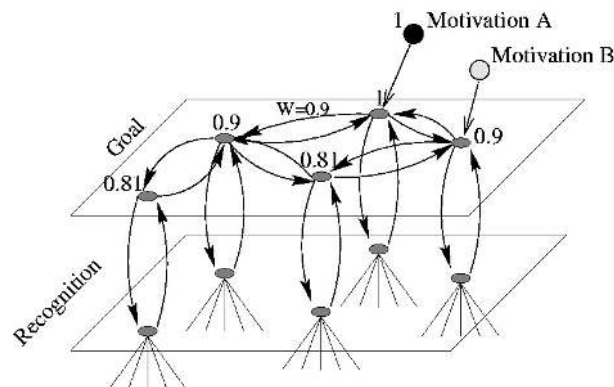

Figure 3: Global architecture of the planning system. When a motivation "A" activates a goal, a back-propagation of the information is performed in direction to all the nodes of the graph (all the weights are equal to 0.9). The recognition level allows to identify the situations when the robot arrives in the vicinity of a learned place. Those situations are directly linked to the goal level, which allows to plan a route from one attractor to the next.

When an interesting place is found, the robot can learn to associate the satisfaction of this goal to the node corresponding to that place on its cognitive map. Later, when the robot is at another place, if the previous goal is activated, it triggers a diffusion mechanism from node to node all along the cognitive map. Then, with the appropriate mechanism, it is easy to choose from the current robot location the action that will allow to go to the place corresponding to the most activated node in the direct neighborhood of the current node on the cognitive map. Gradually, the robot will rise in the gradient direction and reach the goal. If several goals are activated at the same time, the robot will choose the nearest goal location in terms of the minimum number of intermediate nodes on the map (cognitive distance).

This means that, in a first approximation, the robot should not be able to solve difficult choices such as going to a more distant place to satisfy two goals at the same time — therefore minimizing the global cost of the complete journey — instead of going to a nearer place where it can only satisfy one. However, the coupling of the dynamics between the robot's motivations (such as going regularly to eat and to drink) allows to reinforce particular links in the cognitive map. As a result, when the robot discovers randomly an "interesting" route (e.g., a route leading to a place where it can satisfy both goals simultaneously), it can reinforce this new route because it uses it more and more often, and because the timing in the motivations is modified, allowing to synchronize the activation of the motivations to the new pathway (Gaussier *et al*. 2000, Quoy *et al*. 1999). Figure 5 shows how the path that the robot follows to satisfy both goals changes when it tries by chance the new pathway and changes the activation timing of both motivations.
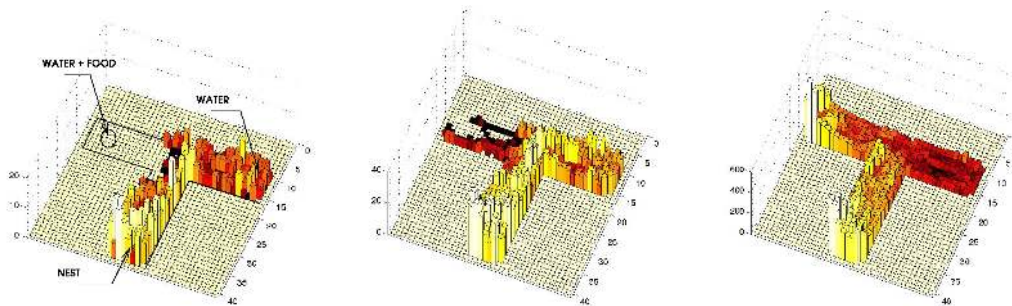

Figure 4: Histograms representing the evolution of the robot's behavior in a T-maze. At the beginning (left) the robot tries the right arm, finds the water and moves between the water and the nest (it does not succeed to explore the left arm). By chance, it begins to explore the right arm and finds water and food (center). Then very quickly it only explores that arm. At last (right), it almost never returns in the right arm except during the exploration phases. The histogram shows the cumulate number of crossing in a given area since the beginning of the exploration. The levels are normalized.

More complex action selection mechanisms have been used to manage and sequence complex behaviors (see e.g., Maes 1991, Tyrrell 1993, Donnart and Meyer 1996, Cañamero 1997) or to sum the effects of various motivations (e.g., Spier and McFarland 1997, Avila-García *et al*. 2003), but the principles underlying these mechanisms remain the same.

In all these cases, the use of a behavior-based approach that grants dynamical and continuous control of the robot also allow to simplify the work of the robot designer (Kelso 1995, Schöner *et al*. 1995). However, these architectures are not devoid of problems. For example, due to the use of local information, it may happen that the robot remains stuck in local minima and deadlock situations, such as switching rapidly and unproductively between alternative actions (dithering). As we will see in Section 3, this is a typical problem that an emotion-based mechanism could help to solve in a more satisfactory way.

3 Emotions in Autonomy, Adaptation, and Survival in the Solitary Robot

The previous section has illustrated some of the typical problems addressed in behavior-based robotics without resorting to any explicit "affective" mechanism in the architecture of the robot. However, for certain types of situations, it might make sense to use "emotion-based" mechanisms and strategies akin to those found in animals and humans when dealing with similar kinds of situations, in particular when problems of autonomy, adaptation and survival are at stake. Emotions are interesting to roboticists in various fundamental ways, by virtue of the main functions they play as regulatory systems at different levels, such as:

Mechanisms for (bodily) adaptation. By producing (rapid) bodily and cognitive changes, emotions constitute mechanisms that allow to deal better (faster or more appropriately) with events, both internal and external, which are important for the survival of the agent (and the species) in a particular environment, such as dangers, unexpected events, and opportunities.

Motivating and guiding action. As Frijda puts it (1995, page 506), "What emotions are about is action (or motivation for action) and action control." At the simplest level, the categorization of events as pleasant/unpleasant, beneficial/noxious, turns neutral stimuli into something to be pursued or avoided. Due to their generality of object, time, intensity, and density, emotions amplify the effects of motivations, which are stimulus-specific and need an urgent satisfaction (Tomkins 1984). As a consequence, they can change goal or motivation priorities (overriding "decisions" arising from the motivational state) to deal with situations that need an urgent response.

Expressive and communicative function. Emotions and their expression are also of paramount importance in communication and in the regulation of social interactions, as we will see in Section 5.

All these different functions could contribute to enhance the adaptation capabilities and autonomy of robots in one or another way (see Cañamero 2001c, Cañamero 2002a for a discussion of these issues). The next section provides some concrete examples regarding the roles of emotions in adaptation and motivation and guidance of action, and later Section 5 will touch upon applications of their expressive and communicative role in the context of human-robot interaction.

3.1 Where could emotions help?

Let us now consider some of the major problems faced by autonomous robots for which emotions could help provide better solutions.

Management of goals. Robots endowed with some sort of "motivational system" that sets and prioritizes "internal" goals to drive their decision making process enjoy greater autonomy than purely reactive robots that only respond to the presence of

external stimuli. However, in some cases, in particular in rapidly changing dynamic environments that presents threats to the survival of the robot, it can be disadvantageous to follow strictly the dictate of the motivational system, since goal priorities might need to be changed to address a new and urgent problem (e.g., to escape from a danger) before having satisfied the current need. As we saw at the beginning of this section, emotions can produce changes in goal priorities, and several models of emotions considered as "interrupts" (Simon 1967) have been implemented in robots.

Repetitive and inefficient behavior. Due to various features such as the use of local information and poor or noisy sensors and actuators, autonomous robots can engage in repetitive activities ("loops" or "deadlocks") that do not conduct to the achievement of the pursued goal and at the same time imply a cost (e.g., in terms of energy consumption) for the robot — for instance, repeatedly trying to get hold of an object that they cannot reach. A "second order" control mechanism that could monitor performance and goal achievement to detect inefficient behavior due to deadlocks would provide a solution to this problem. Rather than a burdensome "meta-reasoning" system, a more appropriate solution for these robots would be the use of internal states such as "boredom," "frustration" or "anger" arising from inefficient behavior and failure to achieve the current goal, and leading to goal or behavioral change that puts an end to the deadlock situation.

Autonomous learning. As we have seen in Section 2.2, learning in autonomous robots typically follows association or reinforcement models, and makes use of some kind of external signals (arising from the environment or supplied by a "critic") that provide positive or negative reward. Other models include learning by imitation (see below, Sections 4.1 and 4.2) the actions of another agent (human or robot) usually called the "demonstrator." One of the main problems underlying all these models is how to make those signals truly meaningful to the robot so that the learning process is more autonomous and grounded in the architecture of the robot. In other words, how can a robot make sense of the perceived signals by itself, as opposed to using reward information provided by some sort of "external teacher"? How can it decide what to learn and what *not* to learn? A mechanism rooted in an internal "value system" would be needed to provide internal signals regarding the "positive" or "negative" qualities of actions and stimuli, giving them a meaning with respect to the values, needs and goals of the robot beyond a metaphoric use of the terms "pain" and "pleasure" to refer to negative and positive reward, and allowing to learn appropriate valenced reactions to them. See (Andry *et al*. 2001), (Cos-Aguilera *et al*. 2003), and Section 4.2 of this chapter for initial solutions in this direction.

Cognitive overload. Management of memory is another major problem in autonomous robots. If the robot lacks appropriate criteria to filter out information, its memory is then too global, causing problems of cognitive overload and very long recall times. Mechanisms for selective memory inspired from emotional memory in humans (e.g., phenomena like mood-congruent recall of past memories) and the related notion of autobiographic memory, would help to solve some of these problems and also provide generally coherent responses to a wide range of situations.

It can be argued that it should be possible to conceive of different mechanisms to "fix" the above problems independently in a more or less *ad hoc* way without having to resort to any emotion-like system and without using the term "emotion" for those mechanisms. However, the interest of using an emotional system (or for that matter, implementing that metaphor) lies in the fact that the same system is intertwined with and can affect various other subsystems *simultaneously*. Although we will not attempt to define the term "emotion," we consider that in this context emotions can be appropriately characterized as processes of dynamic nature that integrate causally related processes of several subsystems (Mandler 1985): the physiological, the cognitive-evaluative, and the communicative-expressive[2]. Emotions then can be seen as "constructs which allow causally related complex dynamic states of the (four) subsystems to be characterized and identified simultaneously" (Pfeifer 1991).

Let us illustrate some of these ideas using the particular case of action selection.

3.2 Emotions for action selection in autonomous robots

What do robot emotions look like? Robotic emotional systems are no magic trick — like everything in robots, they consist of some form of "algorithm" or program. What is important in this "program" is how it affects the behavior and cognitive processing of the robot. Different types of models and approaches can be used to model emotions, as discussed in (Cañamero 2001c) (see also Cañamero 1998, Cañamero 2001b, for a representative collection of papers), but here we adopt the perspective of behavior-based robotics. Within this approach, if emotions are to be meaningful to the robot, they must be an integral part of its architecture and must be grounded in an internal value system that is adaptive for the robot's physical and social niche. It is this internal value system that is at the heart of the creature's autonomy and produces the valenced reactions that characterize emotions. As Wehrle puts it (2001, page 576), "grounding somehow implies that we allow the robot to establish its own emotional categorization which refers to its own physical properties, the task, properties of the environment, and the ongoing interaction with its environment."

As an example, the architecture proposed in (Cañamero 1997) and analyzed in (Cañamero 2002a) relies on both motivations and emotions to perform behavior selection. Initially implemented in simulated robots, this architecture is now being adapted to real robots as part of the PhD thesis of Orlando Avila-García at the University of Hertfordshire. The robots inhabit a typical action selection environment containing various types of resources, obstacles that hamper their activities, and predators, and they must choose among and perform different activities in order to maintain their well-being (the stability of their internal milieu) and survive (remain viable in their environment, following (Ashby 1952) as long as possible — their ultimate goal). The architecture of the robots is behavior-based, and consists of: a *synthetic physiology* of survival-related variables controlled homeostatically (e.g., blood sugar, vascular volume, energy, etc) and "hormones" that can alter the levels of the controlled variables; a set of *motivations* (aggression, cold, curiosity, fatigue, hunger, self-protection, thirst, and warm) activated by "errors" (deficit or excess) in

---

[2] When considering emotions in humans, the subjective experience subsystem is also included. However we prefer not to talk about subjective experience in the case of robots.

the levels of the controlled variables when these depart from their ideal values, therefore setting the internal needs of the robot; a repertoire of *behaviors* that can satisfy those internal needs or motivations (and also create new ones), as their execution carries a modification (increase or decrease) in the levels of specific variables; and a set of "basic" *emotions* (anger, boredom, fear, happiness, interest, and sadness) that can be activated as a results of the interactions of the robot with the world — the presence of external objects or the occurrence of internal events caused by these interactions — and release "hormones" when active.

Under "normal" circumstances, behavior selection is driven by the motivational state of the robot — at each point in time, normally the motivation with the highest need to be satisfied will be in charge of selecting the behavior that can best correct its related physiological error. Emotions constitute a "second order" control mechanism running in parallel with the motivational control system to continuously "monitor" the external and internal environment for significant events. They can alter motivational priorities and behavior execution through the effect of released hormones on the physiology, arousal, attention, and (internal and external) perception of the robot. This emotional system was specifically designed to overcome some of the major problems of reactive and motivated behavior selection architectures, such as lack of flexibility in overall behavior due to a rigid link between stimulus and response, repetitive and inefficient behavior, inefficient treatment of emergency situations, etc. Therefore the different emotions were designed to act as mechanisms for fast adaptation to particularly significant circumstances from the point of view of the survival of the robot in a highly dynamic environment, namely:

- *Anger*: A mechanism to block the influences from the environment by abruptly stopping the current situation. Its triggering event is the fact that the accomplishment of a goal (a motivation) is menaced or undone.
- *Boredom*: A mechanism to stop inefficient behavior that does not contribute to satisfy any of the robot's needs. Its triggering event is prolonged repetitive activity.
- *Fear*: A defense mechanism against external threats. Its triggering event is the presence of predators.
- *Happiness*: A form of re-equilibration triggered by the achievement of a goal.
- *Interest*: A mechanism for the robot to engage in interaction with the world. Its triggering event is the presence of a novel object.
- *Sadness*: A mechanism to stop an active relation with the environment when the robot is not in a condition to get a need satisfied; it acts by slowing down the metabolism and motor system of the robot.

This architecture was designed to show how the use of emotions can improve action selection in autonomous robots by solving some of the problems present in more traditional architectures. It could also be useful to study how emotions can be maladaptive by simulating (via modification of relevant parameters) various emotional disorders. However, since it was designed (hand-coded) to meet particular requirements, it cannot tell us why emotions are adaptive, or how they evolved to be so. Using evolutionary techniques to generate different emotional systems therefore seems a very interesting direction to explore, as it would also allow us to evaluate the

performance and adaptive value of emotions for different action selection tasks and environments. Different aspects of emotion-based learning need also be investigated in this context. Using a simplified version of this architecture, ongoing work as part of the PhD thesis of Ignasi Cos-Aguilera includes learning affordances in the context of behavior selection based on internal homeostatic and hormonal (valenced) feedback resulting from interaction episodes with objects in the environment (Cos-Aguilera *et al*. 2003).

4   Issues in Human-Robot Interaction

The models and architectures seen in previous sections lack any mechanisms to allow social interactions with other entities — humans or robots. To interact socially, many other aspects and cognitive capabilities have to be incorporated in the robot. In some cases, interaction can result as a side-effect of clever sensory-motor couplings, as we will see in the next section. In other cases, we might want to endow our robots with explicit expressive capabilities that make humans attribute them emotional states and react to them.

4.1   Side effects of sensory-motor systems

Robot learning has been studied for a long time in the case of an isolated robot. Algorithms have become more and more efficient but after a while it became obvious that an isolated system, even with the best algorithms, cannot be able to discover very complex strategies or behaviors by its own. The combinatorial explosion is simply to huge and time consuming. Imitation appears as a powerful tool for learning. The problem is that most of the works on imitation (e.g., Meltzoff and Moore 1977) assume the imitator as a model of the demonstrator, and this implies the presence of a complex internal (and innate) mechanism to manage the imitation process.

Our idea at the Neurocybernetics Team was to develop a simple sensory-motor regulation system, somewhat like a homeostat (Ashby 1952) to learn the correspondence between the visual position of the robot arm and its proprioception. In this architecture, the robot simply detects the area in the image where more motion is occurring and tries to learn to associate this location with the different possible proprioceptions. If the robot perceives a difference between its proprioception and its vision, it tries to reduce it. An interesting side effect is that if a human is moving his arm in the robot field of view, the robot perceives the motion and tries to reduce the error between its proprioception and the visual information. As a result, the robot arm begins moving, following the movement of the human demonstrator. Therefore, an external observer has the impression that our robot is imitating, while all it tries to do is remaining in a fixed position, and it is just moving to reduce the error between its proprioception and its vision (see Figure 6).
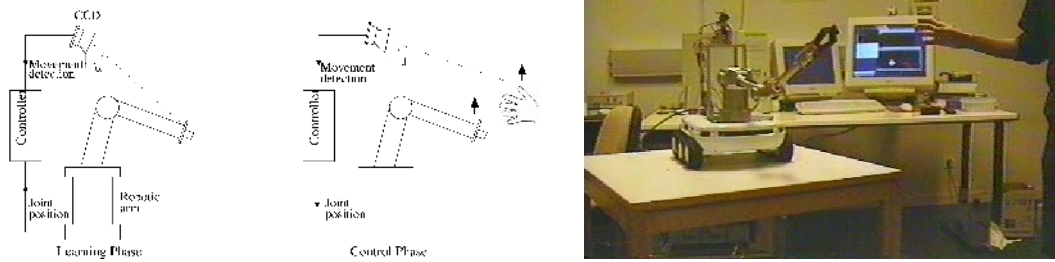
Figure 5: Low-level imitation principle applied to a robotic arm.

Here, perceptual ambiguity is used to bootstrap the imitative behavior, which only relies on a very simple sensory-motor strategy. The neural network used for this is not so obvious, though, since it works with a single camera, a five-degree-of-freedom arm, and it allows to build a multi-modal map (Gaussier *et al*. 1998, Andry *et al*. 2000, Andry *et al*. 2002). Learning the complete movement is easy since the robot does not try to learn the trajectory of the demonstrator's arm trajectory, but it simply needs to store its own sequence of actions. Unfortunately, an external reward is needed to decide if the trajectory has to be stored or not — was it a good imitation? Was it a real demonstrator or just a shadow our robot arm was following? We will see latter how to avoid the use of an explicit learning signal.

Another interesting example is the work of (Takanishi *et al.* 1997). They proposed the design of a robotic head with a neck and an active vision system allowing to control the gaze of two CCD cameras. The system was tested in different situations showing that active movements of the robot camera and active movements of the head/neck in the direction of the target allowed a better measure of the target localization. An interesting side-effect was how much this simple sensory-motor coupling was appealing to the observers. We had the feeling the robot "wanted" to look at something. Moreover, when the object was too near to the head, the head was moving back and the observer had the impression that the robot was trying to avoid an unpleasant situation. It was a clear case of a robotic system able to trigger some empathic behaviors, due to our tendency to assign intentions to others.

This system was able to trigger an emotional response in the human observers by the only use of a simple hardwired sensory-motor mechanism. This result opens the door to the development of non-conventional human/machine interfaces showing how powerful the evocation of emotional feeling can be to interact with a human — even if the robot does not have any emotional system at all.

4.2   Imitation as a communication tool

As seen in the previous section, passive imitation is interesting to trigger some basic social interactions. However, in baby/parents relations there is no need for an explicit reward to learn a particular behavior. Consequently, in the case of a robot, the question is: how could internal reward be produced that does not depend on a

symbolic signal? The solution came from the study of baby/parent interaction and more precisely on the fact that the synchrony of the exchange was crucial to maintain the attention of the baby (Nadel 2000). We proposed a simple mechanism merging all the input data in order to try to predict the rhythm of the global incoming flow of sensory data. If the system succeeds to predict the rhythm, we consider that the robot is in the rhythm of the interaction and that this later is successful: the robot generates by itself a positive reward that reinforces its current behavior. On the contrary, if the robot fails to predict the rhythm of the exchange we consider that the robot is not in the rhythm of the exchange; a negative reward is then generated, inhibiting the current behavior.

We were able to show that this kind of mechanism can be used to teach an arbitrary set of sensory-motor rules to a computer without giving it any reward (Andry *et al*. 2001). It was the frequency of the trials of the human demonstrator that was directly used by the computer to infer if it has or not to reinforce its current behavior. In order to close this "interaction loop" between the robot (or computer) and the human, we would need to endow the artifact with the ability to display the "internal state" induced by this reward in a way that is not only efficient but also intuitive and acceptable to humans. Let us now consider how roboticists approach the design of expressive artifacts.

5   Emotions in Human-Robot Interaction

Whereas emotion modeling for individual robots has focused on the design of architectures, i.e., the "inner" aspect of emotions, the design of emotions for robots interacting socially has primarily paid attention to the "external" features of emotional expression. Emotional expression is certainly a key factor in social communication and interaction, since the external manifestations of emotions can play a major role as "signaling" mechanisms, at several levels. For example, the emotional expression of an individual can be used by another as social reference, to "assess" the type of situation it is confronted with and the appropriate response to it. In some cases, emotional expression carries some "information content," and it can then be said to play a communicative role; it can be controlled to some extent and be intentionally used to let others know one's emotional state so that they can make expectations about our own behavior and adapt theirs accordingly. Emotions also contribute to the construction of intersubjectivity.

Ideally, we would like to see these "internal" and "external" aspects fully integrated in the same robotic architectures in a not too distant future. For the time being, however, the enormous challenges that each of these problems poses in robotics has caused that attempts at bringing them together have so far been very simple and rather "ad hoc." In this section we will therefore deal only with the "external manifestations" of emotions in the context of human-robot interaction.

5.1   Why expressive robots?

Work on expressive robots for interaction with humans is receiving increasing attention. Intuitively, we can think of different roles that emotional expression can play in social interactions between humans and robots:

Conveying intentionality. People need to understand the behavior observed in their social (human or artificial) partners as resulting from causes or intentions that allow them to form coherent explanations to interpret past behavior, make predictions and establish expectations about future behavior. Emotions and personalities are often postulated as such causes of behavior and sources of intentions. Autonomous robots might, in addition, use emotions and their expressions to convey intentions or needs to humans.

Eliciting emotions. In the same way as other people's emotions elicit emotional responses from humans, emotions in robots can be used with the same purpose, seeking responses that either match the robot's emotional state (e.g., a pilot assistant that tries to bring the pilot to an alert state) or are instrumental to it (e.g., a robot expressing sadness due to its inability to accomplish a task can receive the help of a "moved" human).

Human comfort. Robots able to express emotions and adapt their interactions to the emotional state of their social partners can be expected to make humans feel more comfortable during interaction. One obvious reason is that this interaction is tailored to meet the emotional needs of the human. Another important reason, is that emotional behavior and expressions make the robot more believable, as in a sense it is perceived as being "closer" to ourselves — or at least to a living being.

Enhanced communication. Emotional expression being a key element in non-verbal communication, endowing a robot with emotional expressions can make communication cognitively less costly for the human partner. If emotional robots are to achieve a sufficient level of sophistication to interpret our subtle expressions and obtain relevant information from contextual clues at some point in the future, we might also want them to "understand" what we *mean*, not (only) what we say.

Giving the impression of life, believability, and interaction adapted to humans seem thus to be the main features that expressive robots should show to be accepted as "social partners" by us. How can they be actually implemented in robots?

5.2   Designing emotional expression for robots

Building a "believable" expressive robot that can interact with humans poses many challenges that, in our opinion, need to be approached from a multi-disciplinary perspective. In addition to the more technical aspects, robot designers are confronted with a number of conceptual problems for which psychological theories and models, conceived to analyze existing (biological) systems rather than to build artificial ones, do not provide much guidance. Let us examine some of them.

Features

Human emotional expression is multi-modal and highly complex, very difficult to reproduce in robots in its entirety. Researchers have to select a subset of features to convey emotion in a way that makes expression believable—although not necessarily realistic—to the human observer. Inspiration and data are usually sought from psychology, classical animation (e.g., Thomas and Johnston 1981), and very often empirical testing. Faces are the most commonly used means of expression, usually in a simplified (and caricaturized) presenting eyes, a mouth, eyebrows, and eyelids in the most complex versions. Very few attempts exist to produce highly realistic faces, including elements such as artificial skin (see Menzel and d'Aluisio 2000 for examples). Posture and movement (of the elements of the face and sometimes of other body parts, e.g., a neck) are regarded as highly important for expression; in particular movement must be coherent, well synchronized and have the right timing. Some researchers (e.g., Breazeal 2002) put forward the use of features activating the "baby scheme"—features that make adults react to them as they would to a baby—although it has been discussed that the use of those features is not necessary for emotional expression (Cañamero 2002b) but rather might make the robot itself more appealing or "cute" to the human eye. Expressive faces are sometimes combined with vocal inflexion (see e.g., Picard 1997, Breazeal 2002), although current speech synthesis systems still cannot produce voice of human quality and tend to sound too artificial.

Underlying model

Is facial emotional expression better modeled as a discrete set of prototypical expressions or as points in a space of continuous dimensions? The choice of one or the other model has an impact on the type and quality of the expressions produced, but again the literature does not provide much guidance and roboticists must generally resort to the use of their own intuition and much empirical testing. Most robotic faces follow a discrete categories approach and use few degrees of freedom (typically four to six) and few expressive features (typically two to four) to produce highly stereotyped but easily readable expressions, such as Sparky (Scheeff *et al.* 2002), shown in Figure 7, left; some of these robots also allow to blend prototypical expressions to form more complex and "chimeric" ones, such as Feelix (Cañamero and Fredslund 2001), depicted in Figure 7, center. The fact that stereotyped expressions are easily readable makes these robots suitable for simple interactions in which it is important to avoid putting strong cognitive demands on the human user, such as when interacting with children or people with cognitive disabilities, and when the user must pay attention in priority to some other task and the face is used as an expressive "interface." Other robots designed to engage in more complex interactions and to produce more subtle expressions follow a dimensional approach. This if for example the case of Kismet (Breazeal 2002), a robotic head with many degrees of freedom, many expressive features, and carefully timed movement (Figure 7, right), designed as a testbed to investigate infant-caretaker interactions. The expressions of this robot are not always easily interpretable; however, it can produce more subtle expressions that engage humans to interact with it as they would with an infant.
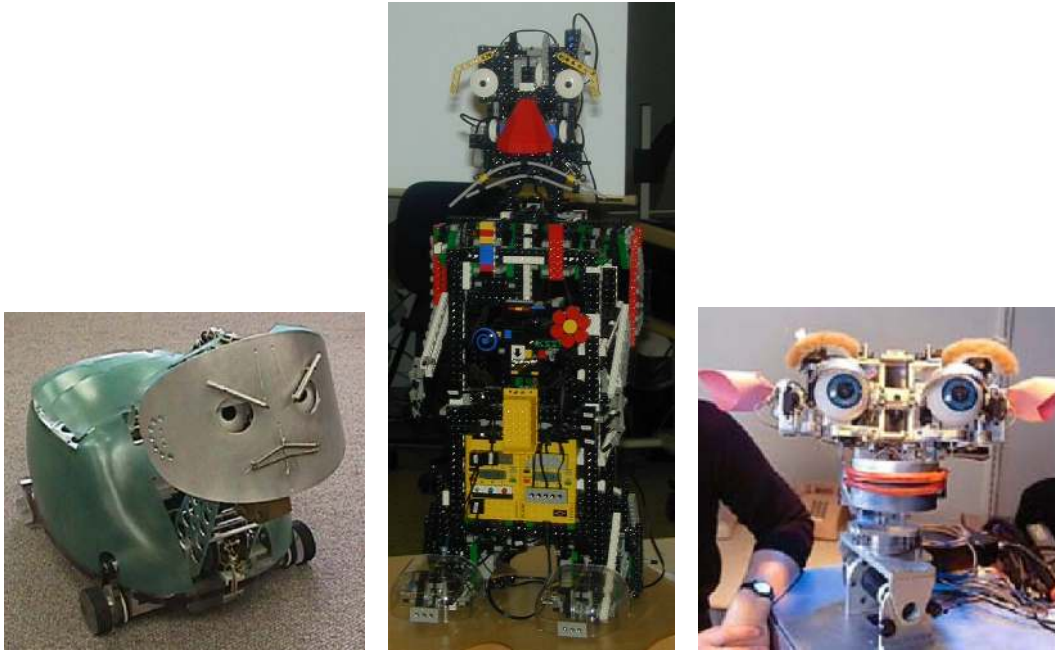
Figure 7. Expressive robots. **Left**: Sparky. **Center**: Feelix. **Right**: Kismet.

Level of complexity

The fact that the emotional expression of the robot has to be believable to humans does not imply that it must be realistic. As a matter of fact, most expressive robots have caricaturized faces rather than detailed reproductions of human faces. Although some researchers struggle to achieve realistic human-like faces, there seems to be a good reason to prefer simplicity unless a nearly perfect level of realism can be attained. This idea is reflected in the "uncanny valley" hypothesis put forward by the Japanese roboticist Masahiro Mori and summarized in (Reichard 1978): The emotional reaction of humans to human appearance and movement in robots is not linear. It increases positively with similarity, but a chasm is found, when similarity is high enough to appear human but imperfections in shape or movement are perceived as very disturbing, producing a negative emotional reaction. A strong positive emotional reaction is found again when similarity is perfect. According to Mori, movement seems to have more weight in this reaction than appearance. A cartoon-like, very caricaturized face with rudimentary movement can thus be more effective than a very sophisticated robotic face, since the state of the art in robotics is still far from producing the required level of complexity.

"Shallow" or "deep" modeling?

Is it enough to model the external features of emotional expression to achieve believable interactions with humans, or should these external manifestation be produced by an underlying emotional system? As we have seen in the case of the active vision head described in Section 4.1, simple but cleverly engineered sensory-motor coupling can result in very believable emotional reactions even if the system had not been designed with that purpose in mind. Therefore, in the case of sporadic or short-term interactions, a "shallow modeling" approach can be enough and very effective. However, to achieve believable long-term interactions, the external behavior of the robot must be not only coherent, but also flexible and adapted to that of its

partner—features that can only be achieved if external behavior results from a "deep" internal model.

5.3   Useful tools?

In Section 5.1, we have sketched some of the reasons that make expressive robots interesting devices for interaction with humans from a practical and engineering perspective. However, expressive robots can also be useful tools to facilitate and support the investigation of research questions regarding human emotions. In this respect, we have developed a very simple expressive head (Figure 8) in the context of an ongoing collaboration with psychologist Philippe Brun of University of Rouen. This head is being used as a tool to study imitation and evocation of emotional expressions by typical and autistic children. Therefore, the features and expressions of the head need to be very simple—only five degrees of freedom, three in the mouth and two in the eyebrows, are used to display a small subset of emotional expressions. However, even a simple robot like this presents a number of features that make it an interesting tool for research, such as:
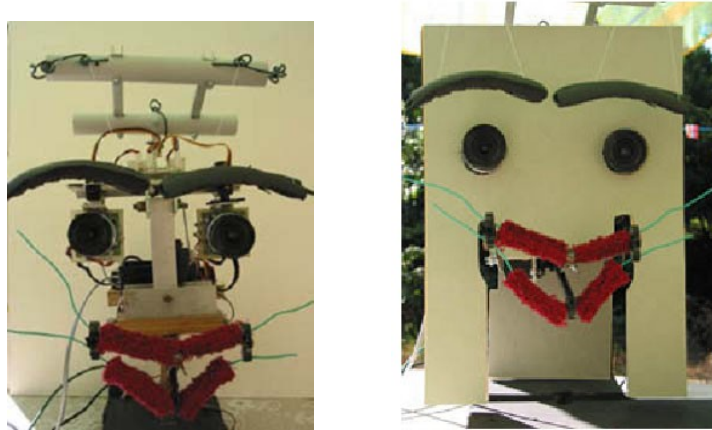

Figure 6: Our expressive head. Left: bare version. Right: covered with a mask.

- A few parameters can be modified to control movement, speed, timing, etc., in order to vary the dynamics of the expressions.
- A wide range of expressions, both typical and atypical, can be formed.
- The robot can display still or dynamic emotional expressions.
- It can be used as a tool to display expressions only, or in interaction with the human.
- The degree of similarity with a human face can be easily varied—always remaining a caricature of it—by adding other features, in order to assess the degree of complexity needed for a "humanoid" to be perceived as "human."
- A physical device is usually perceived as more appealing and life-like than a video image. However, it is not clear whether this is always the case or whether this always a positive feature—for example, this might be questionable with autistic people.

Such a simple robot allows to investigate a number of questions that would be difficult to study otherwise. A more interesting tool, which would also be a model of theories

on emotion understanding, should be capable of reacting to human expressions and behavior, therefore closing the interaction loop—an issue in our agenda for future research.

6   Concluding Remarks

In this chapter, we have reviewed state-of-the art research in behavior-based robotics relevant to the investigation of the contributions of emotions to adaptive behavior and social interaction, and therefore to the use of robots as tools and models for emotion research. The emphasis that the behavior-based paradigm puts on complete creatures in closed-loop bodily interaction with their environment has important implications for the design of artificial emotional systems (Cañamero 2001a), particularly:

- To be meaningful to the robot, emotions must be an integral part of its architecture rather than an *ad hoc* appendage. This means that emotions must be grounded in an internal value system that is meaningful (adaptive) for the robot's physical and social niche. It is this internal value system that is at the heart of the creature's autonomy and produces the valenced reactions that characterize emotions.
- Emotion grounding requires that our model clearly establish a link between emotions, motivation, behavior (including expressive behavior), perception, and various aspects of "cognition," so that these elements can affect and feed back into one another.
- This link must be rooted in the body of the agent, since it is through the body that agents interact with the physical and social world.

This approach does not imply, however, that only by including explicit "emotion machinery" in the architecture of robots can these produce emotion-oriented behavior or provide useful hints for the investigation of emotional phenomena. On the contrary, we have seen through several examples in active vision, reinforcement learning and imitation, that the use of simple sensory-motor couplings induces emergent behaviors or side effects that create emotional reactions in the human observer. These systems can also provide valuable feedback for modeling purposes, since the emergent properties introduced by the effects of dynamical interaction can simplify the design process and the emotion model itself. Nevertheless, to achieve long-term and flexible adaptation to the environment and to the dynamics of social interactions, emotional systems appear as a fundamental constituent of the control architectures of robots confronted to the same types of situations for which emotions provide useful adaptation mechanisms in humans and other animals. Building such artificial emotional systems is an example of synthetic psychology in which robots constitute efficient tools to simulate and investigate the dynamic effects of cognitive models of the brain and in particular of emotional phenomena.

Acknowledgments

References

Andry, P., Gaussier, P., Moga, S., Banquet, J.P., and Nadel, J. (2001). Learning and Communication in Imitation: An Autonomous Robot Perspective. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 31(5): 431—444.

Andry, P., Gaussier, P., and Nadel, J. (2002). From sensory-motor coordination to low-level imitation. In G. Prince, Y. Demiris, Y. Marom, and H. Kozima, editors, *Proc. Second Intl. Workshop on Epigenetic Robotics*. Lund University Cognitive Studies, 94: 7—15.

Andry, P., Moga, S., Gaussier, P., Revel, A., and Nadel, J. (2000). Imitation: learning and communication. In *Proc. Sixth International Conference on Simulation for Adaptive Behavior* (SAB'2000), pages 353—362. Cambridge, MA: The MIT Press.

Arkin, R. (1998). *Behavior Based Robotics*. Cambridge, MA: The MIT press.

Ashby, W. (1952). *Design for a Brain: The Origin of Adaptive Behavior*. London: Chapman and Hall.

Avila-García, O., Cañamero, L. and te Boekhorst, R. (2003). Analyzing the Performance of "Winner-Take-All" and "Voting-Based" Action Selection Policies within the Two-Resource Problem. In W. Banzhaf, T. Christaller, P. Dittrich, J.T. Kim, and J. Ziegler, editors, *Advances in Artificial Life. Proceedings of the 7th European Conference (ECAL 2003)*, LNAI 2801, 733—742. Berlin Heidelberg: Springer-Verlag.

Barto, A., Sutton, R., and Anderson, C. (1983). Neuronlike adaptive elements that can solve difficult control problems. *IEEE Transactions on Systems, Man and Cybernetics*, 13(5): 834—846.

Breazeal, C. (2002). *Designing Sociable Robots*. Cambridge, MA: The MIT Press.

Brooks, R.A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1): 14—23.

Brooks, R.A. (1991). Intelligence Without Representation. *Artificial Intelligence*, 47(2): 139—159.

Cañamero, L.D. (1997). Modeling Motivations and Emotions as a Basis for Intelligent Behavior. In W.L. Johnson, editor, *Proc. of the First International Conference on Autonomous Agents*, 148—155. New York, NY: ACM Press.

Cañamero, L.D., editor (1998). *Emotional and Intelligent: The Tangled Knot of Cognition. Papers from the 1998 AAAI Fall Symposium*. Menlo Park, CA: AAAI Press.

Cañamero, L.D. (2001a). Building Emotional Artifacts in Social Worlds: Challenges and Perspectives. In *Emotional and Intelligent II: The Tangled Knot of Social Cognition. Papers from the 2001 AAAI Fall Symposium*, 22—30. Menlo Park, CA: AAAI Press.

Cañamero, L.D., editor (2001b). *Emotional and Intelligent II: The Tangled Knot of Social Cognition. Papers from the 2001 AAAI Fall Symposium*. Menlo Park, CA: AAAI Press.

Cañamero, L.D. (2001c). Emotions and Adaptation in Autonomous Agents: A Design Perspective. *Cybernetics and Systems*, 32(5): 507—529.

Cañamero, L.D. (2002a). Designing Emotions for Activity Selection in Autonomous Agents. In R. Trappl, P. Petta and S. Payr, editors, *Emotions in Humans and Artifacts*. Cambridge, MA: The MIT Press.

Cañamero, L.D. (2002b). Playing the Emotion Game with Feelix: What Can a LEGO Robot Tell Us about Emotion? In K. Dautenhahn, A. Bond. L.D. Cañamero and B. Edmonds, editors, *Socially Intelligent Agents: Creating Relationships with Computers and Robots*, 69—76. Norwell, MA: Kluwer Academic Publishers.

Cañamero, L.D. and Fredslund, J. (2001). I Show You How I Like You—Can You Read it in My Face? *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 31(5): 454—459.

Cos-Aguilera, I., Cañamero, L. and Hayes, G. (2003). Learning Object Functionalities in the Context of Behavior Selection. In U. Nehmzow and C. Melhuish, editors, *Proceedings of Towards Intelligent Mobile Robots (TIMR'03): 4th British Conference on Mobile Robotics*. University of the West of England, Bristol, UK, 28—29 August, 2003.

Donnart, J.-Y. and Meyer, J.-A. (1996). Hierarchical map building and self-positioning with MonaLysa. *Adaptive Behavior*, 5(1): ** PHILIPPE, pages ??

Frijda, N.H. (1995). Emotions in Robots. In H.L. Roitblat and J.-A. Meyer, editors, *Comparative Approaches to Cognitive Science*, 501—516. Cambridge, MA: The MIT Press.

Gaussier, P., Joulain, C., Banquet, J., Leprêtre, S., and Revel, A. (1999). The visual homing problem: an example of robotics/biology cross-fertilization. *Robotics and Autonomous Systems*, 30: 155—180.

Gaussier, P., Leprêtre, S., Quoy, M., Revel, A., Joulain, C., and Banquet, J. (2000). Experiments and models about cognitive map learning for motivated navigation. In J. Demiris and A. Birk, editors, *Interdisciplinary approaches to robot learning*, volume 24, pages 53—94. Robotics and Intelligent Systems Series, World Scientific, ISBN 981-02-4320-0.

Gaussier, P., Moga, S., Quoy, M., and Banquet, J. (1998). From perception-action loops to imitation processes: a bottom-up approach of learning by imitation. *Applied Artificial Intelligence*, 12(7/8): 701—727.

Gaussier, P., Revel, A., Joulain, C., and Zrehen, S. (1997). Living in a partially structured environment: How to bypass the limitation of classical reinforcement techniques. *Robotics and Autonomous Systems*, 20:225—250.

Gaussier, P. and Zrehen, S. (1995). Perac: A neural architecture to control artificial animals. *Robotics and Autonomous System*, 16(2-4): 91—320.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42: 335—346.

Kelso, J.S. (1995). *Dynamic patterns: the self-organization of brain and behavior*. Cambridge, MA: The MIT Press.

Maes, P. (1991). A Bottom-Up Mechanism for Behavior Selection in an Artificial Creature. In J.A. Meyer and S.W. Wilson, editors, *Proc. First Intl. Conf. on Simulation of Adaptive Behavior*, 238—246. Cambridge, MA: The MIT Press.

Mandler, G. (1985). *Mind and Body*. New York, NY: W.W. Norton.

Meltzoff, N. and Moore, M.~K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198: 75—82.

Menzel, P. and d'Aluisio, F. (2000). *Robo sapiens: Evolution of a New Species*. Cambridge, MA: The MIT Press.

Nadel, J. (2000). The functional use of imitation in preverbal infants and nonverbal children with autism. In A. Meltzoff and Prinz, W., editors, *The Imitative Mind: Development, Evolution and Brain Bases*. Cambridge: Cambridge University Press.

Pfeifer, R. (1991). A Dynamic View of Emotion with an Application to the Classification of Emotional Disorders. Vrije Universiteit Brussel, AI Memo 91—8.

Pfeifer, R. and Schreier, C. (1999). *Understanding Intelligence*. Cambridge, MA: The MIT Press.

Picard, R.W. (1997). *Affective Computing*. Cambridge, MA: The MIT Press.

Quoy, M., Gaussier, P., Leprêtre, S., Revel, A., Joulain, C., and Banquet, J. (1999). A neural model for the visual navigation and planning of a mobile robot. In *Advances in Artificial Life, ECAL99*, LNAI volume 1674, pages 319—323. Berlin Heidelberg: Springer-Verlag.

Reeves, B. and Nass, C. (1996). *The Media Equation. How People Treat Computers, Television and New Media like Real People and Places*. New York, NY: Cambridge University Press/CSLI Publications.

Reichard, J. (1978). *Robots: Fact, Fiction + Prediction*. London: Thames & Hudson Ltd.

Revel, A., Gaussier, P., Leprêtre, S., and Banquet, J. (1998). Planning versus sensory-motor conditioning: what are the issues? In *Proc. SAB'98: From animals to animats 5*, pages 129—138. Cambridge, MA: The MIT Press.

Scheeff, M., Pinto, M., Rahardja, K, Snibbe, S., and Tow, R. (2002). Experiences with Sparky, a Social Robot. In K. Dautenhahn, A. Bond. L.D. Cañamero and B. Edmonds, editors, *Socially Intelligent Agents: Creating Relationships with Computers and Robots*, 173—180. Norwell, MA: Kluwer Academic Publishers.

Schöner, G., Dose, M., and Engels, C. (1995). Dynamics of behavior: theory and applications for autonomous robot architectures. *Robotics and Autonomous System*, 16(2-4): 213—245.

Simon, H.A. (1967). *Motivational and Emotional Controls of Cognition*. Psychological Review, 74(1): 29—39.

Spier, E. and McFarland, D. (1997). Possibly Optimal Decision Making under Self-Sufficiency and Autonomy. *Journal of Theoretical Biology*, 189: 317—331.

Takanishi, A., Matsuno, T. and Kato, A. (1997)., Development of an Anthropomorphic Head-Eye Robot with Two Eyes – Coordinated Head-Eyes Motion and Pursuing Motion in the Depth Direction (Wasada University Japan). IEEE RSJ, IROS 97 Grenoble France, *video proceedings*.

Thomas, F. and Johnston, O. (1981). *Disney Animation: The Illusion of Life*. New York: Abbeville Press.

Tolman, E. (1948). Cognitive maps in rats and men. *The Psychological Review*, 55(4): 189—208.

Tomkins, S.S. (1984). Affect Theory. In K.R. Scherer and P. Ekman, eds., *Approaches to Emotion*, 163—195. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tyrrell, T. (1993). The Use of Hierarchies for Action Selection. *Adaptive Behavior*, 1(4): 387—419.

Watkins, C. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3): 279—292.

Weaver, S., Klopf, A., and Morgan, J. (1993). A hierarchical network of control systems that learn: modeling nervous system function during classical and instrumental conditioning. *Adaptive Behavior*, 1(3): 263—319.

Wehrle, T. (2001). The Grounding Problem of Modeling Emotions in Adaptive Systems. *Cybernetics and Systems*, 32(5): 561—580.