Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 18 (2005) 445–455
Neural Networks
www.elsevier.com/locate/neunet

2005 Special Issue

Emotion understanding from the perspective of autonomous robots research

Lola Cañamero

Adaptive Systems Research Group, School of Computer Science, University of Hertfordshire, College Lane, Hatfield, Herts AL10 9AB, UK

Received 24 March 2005; accepted 25 March 2005

Abstract

In this paper, I discuss some of the contributions that modeling emotions in autonomous robots can make towards understanding human emotions—'as sited in the brain' and as used in our interactions with the environment—and emotions in general. Such contributions are linked, on the one hand, to the potential use of such robotic models as tools and 'virtual laboratories' to test and explore systematically theories and models of human emotions, and on the other hand to a modeling approach that fosters conceptual clarification and operationalization of the relevant aspects of theoretical notions and models. As illustrated by an overview of recent advances in the field, this area is still in its infancy. However, the work carried out already shows that we share many conceptual problems and interests with other disciplines in the affective sciences and that sound progress necessitates multidisciplinary efforts.

Keywords: Artificial emotions; Autonomous robots; Affective computing; Emotion understanding; Biologically inspired robotics

1. Introduction

Research on computational and robotic models of emotions has been very active over the last decade, which has also witnessed the proliferation of commercial 'affective' toys and robots. In her seminal book *Affective Computing*, MIT Professor Rosalind Picard characterized this research area and its scope (Picard, 1997, p. 3):

" […] computing that relates to, arises from, or deliberately influences emotions. 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."

Why should computers or robots have any affective capabilities or features? In the case of artifacts designed to interact with humans, the ability to display emotional expressions and to recognize and respond appropriately to the emotional states of the users can make them appear more 'life-like' and 'believable' (Bates, 1994; Ortony, 2003) to humans, and therefore make users more prone to accept them and engage in interactions with them (Cañamero & Gaussier, 2005; Stern, 2003). In the case of autonomous robots having to interact and make decisions in dynamic, unpredictable, and potentially 'dangerous' environments, mechanisms functionally equivalent to (some) emotions present in biological systems facing the same types of problems can greatly improve their performance and adaptation to the environment (Cañamero & Gaussier, 2005; Frijda, 1995).

Inspiration from human and animal emotions to include 'emotional' or emotion-like features and mechanisms in artifacts thus seems to help create 'better' engineering systems. However, to what extent are those features and mechanisms comparable to emotions in biological systems?

E-mail address: l.canamero@herts.ac.uk.

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.03.003

Can models and implementations of 'artificial emotions' contribute towards understanding human emotions or more generally what emotions are and how they work? Can they contribute to understanding emotions 'as sited in the brain'?

In this paper, I will focus on autonomous robots rather than computer simulations. Modeling emotions in autonomous robots offers the additional value of allowing to study how an operationalized model works when embodied in a physical entity with real (and therefore noisy and limited) sensing and actuation capabilities, and embedded in the 'real world'—the same world that we humans inhabit—with which it entertains complex dynamics of interactions.

Although the field is still in its infancy, I believe modeling emotions in autonomous robots can offer several valuable contributions to emotion research, principally regarding:

- *Human perception of emotions*. Expressive robots can be very efficient at engaging humans and eliciting emotional responses from them even when they only reproduce observable features of emotional expression (see e.g. Breazeal, 2002; Cañamero & Fredslund, 2001). Even such simple artifacts can be used as tools to investigate human perception of emotions and their influence in the human tendency to anthropomorphize (Reeves & Nass, 1996).
- *Tools to test and investigate theories*. 'Emotion-based' robots are often designed taking inspiration from theories of human emotions, and in some cases (examples of which we would like to see more often), in close collaboration between engineers and theorists—usually psychologists, less frequently neuroscientists. Autonomous robots constitute excellent tools not only to test theories, but also to investigate problems that would be difficult to study in humans, due for example to ethical implications, the difficulty of isolating the relevant elements, or the repetitious nature of the task. In this respect, artifacts can serve as 'virtual laboratories' for the study of emotions.
- *A synthetic approach*. The development of robotic architectures that model and operationalize different aspects of emotions and their involvement in cognitive and behavioral processes in artifacts carries many parallels with the problems investigated by emotion theories and models. Our approaches are however complementary, since we try to understand what emotions are and how they work by 'building' emotional systems, and this synthetic approach can complement studies of existing emotional systems.
- *Operationalization and simplicity*. To implement emotions in artifacts, and in particular in robots, theoretical concepts need to be operationalized very precisely and, very often, simplified. If over-simplification can be a risk, simplification can be very valuable to help us single out aspects of emotions that are important for the phenomenon under investigation.

What about contributing to the understanding of emotion 'as sited in the brain'? To many neuroscientists, robotic models of emotions might seem too unrelated to the human brain. Our models take for the most part inspiration from brain research, but are not as 'neuro-mimetic' as computational models developed by computational neuroscientists, partly due to constraints imposed by the computational and sensorimotor capabilities of the robot, partly due to the need to produce real-time behavioral responses adapted to the dynamics of the environment, and this is often very difficult to achieve with complex neural networks. The nervous systems of our robots are rather closer to those of much more simple animals. However, robotic models of emotions allow to study issues regarding for example the adaptive nature, development and evolution of emotions, issues that are not specific to the human brain but are also relevant to understand how emotions might have come to be what they are.

2. Bottlenecks

As we will illustrate in Section 3, numerous emotion-based robotic systems have been developed in recent years. In spite of the thriving growth of this area and the efforts by many to develop solid pieces of work grounded in sound research, a number of problems constitute, at present, 'bottlenecks' that can only be addressed by large-scale, long-term pluri-disciplinary efforts. Many questions need to be answered in order to achieve principled emotion-based architectures that at the same time provide (a) meaningful and robust solutions to problems arising in autonomous and interactive robots research, and (b) useful tools for emotion theorists to test their models or to gain insights towards emotion understanding. Some of these questions are:

- *Regarding models*: What is the scope of the different types of emotion theories and models? Do they explain the same phenomena/aspects of emotions? To what extent (and which ones) are they incompatible/can they be combined? What is the notion (definition) of 'emotion' underlying each of them and what sort of consequences and 'constraints' does each definition put regarding the operationalization and implementation of each model? Is a general definition of emotions possible/required for modeling?
- *Regarding emotion 'machinery'*: Which are plausible mechanisms underlying different aspects of emotions and their influence in cognition and action? What kinds of conceptual and computational mechanisms are better suited to explain and model the relations between emotion and cognition–action? How can computational mechanisms stemming from different conceptual traditions be integrated?
- *Regarding applications*: Which emotions/aspects of them can be meaningfully implemented in autonomous

and interactive robots? What guidance can emotion theories and models provide in the search for answers to this question? (e.g. focus on the adaptive value/functions/components of emotions?) How can different models be suitably operationalized/for which applications?

- *Regarding the assessment of the influences of emotion in cognition and action*: How can emotional states/processes be quantified and measured (a) from 'inside' the organism—robot architecture; (b) from observed behavior? To what extent does 'emotional behavior' respond to the effects of specific internal 'emotion machinery'? To what extent can it be explained as a 'side effect' of the interactions of the robot with its environment? (emotion as a notion in the eye of the beholder). To what extent does the analysis of observed behavior help us understand underlying emotional mechanisms?

I do not think that these questions need to or indeed can be (fully) answered from a theoretical perspective prior to computational modeling and implementation. I rather see the interplay between theory and computational modeling as a two-way, multidisciplinary enterprise towards improving our understanding of emotions in which modeling can provide support and insights in the search for those answers and in some cases even give rise to the reformulation of some theoretical questions and problems.

3. Approaches to modeling emotion in autonomous robots

Research in this area is devoted to the design and implementation of emotion-based control architectures to improve the adaptation capabilities of physically embodied agents that must act in and respond to changes in their environment autonomously (Cañamero, 2001b). In autonomous robotics, 'adaptation to the environment' is usually considered in two main senses (see Maes, 1995 for an extended discussion), depending on the temporal scale of environmental changes: Adaptation to short-term, smaller changes in the environment involves flexible and rapid decisions, and is the problem dealt with by action selection architectures; adaptation to more significant and long-term (lasting) changes in the environment involves the ability to modify and improve behavior over time and therefore learning. Let us review some of the main achievements in these areas, illustrating them by a few representative examples.

3.1. Emotion in action selection

Computational models of emotions in this area have followed two main approaches: Emotions designed as an integral part of the agent architecture or 'modeled' as emergent phenomena.

3.1.1. Designing emotions for behavior control

This approach postulates that, if emotions are to be meaningful to the robot, they must be an integral part of its architecture. However, to be meaningful, emotions must be grounded in an internal value system that is adaptive for the agent's (physical and social) environment, since 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 an example, the architecture proposed in Cañamero (1997) relies on both survival-related motivations and basic emotions to perform behavior selection by simulated robots.[1] 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 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. 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. Closely related architectures have been implemented in robots by Velásquez (1998) in the case of action selection and learning tasks and by Breazeal (2002) in a social robot.

This approach can show how emotions (or emotion-like mechanisms) modeled as evolutionary adaptations with specific survival-related functions can modulate the

[1] Initially implemented in simulated robots, this architecture is now being adapted and implemented in real robots (Avila-García & Cañamero, 2004, 2005).

'nervous system' and resulting overall behavior of the robot, improving its performance in and adaptation to specific environmental conditions. However, it cannot contribute to explaining why emotions might be adaptive mechanisms (i.e. the origin of such adaptations) since it takes the adaptive nature of emotions for granted. For this, developmental and evolutionary models are needed, and some early studies (so far in simulated agents, see for example Lowe, Cañamero, Nehaniv, & Polani, 2004) have started to investigate how different types of emotion-related displays and behavioral strategies evolve as a consequence of different environmental pressures.

3.1.2. Emergent emotions

The 'emergent approach' is typically adopted in artificial life computational models. In its simplest version, emotions are considered as pure epiphenomena in the eye of the beholder and the robot architecture does not contain any elements to which the production of emotion-like behavior can be related. As representative work in this tradition we can cite the thought experiments of Valentino Braitenberg (1984). His 'Vehicles'—later implemented in real robots in numerous occasions—are simple machines or 'robots' with very simple architectures consisting of direct connections between sensors and motors. They are situated in an environment containing a source of energy (e.g. heat, light) that can be detected by the sensors. By varying the connections between sensors and motors to be either lateral or counter-lateral and the way the motors are powered by the sensors (either positively or negatively), different variants of a basic architecture can be formed that give rise to different behaviors. By varying the position from which the Vehicles approach the energy source, the same architecture will produce different behaviors and the same behavior can be observed in Vehicles with different architectures. Although these architectures do not implement any 'emotional components' and the behavior of the Vehicles depends only on their morphology and interaction with the environment, the behavior displayed by some of these Vehicles can appear to an external observer as arising from internal states, and could be termed as 'love', 'fear', 'aggression', 'cowardice', or 'curiosity.' In the same vein, Pfeifer proposed an artificial life environment of 'Fungus Eaters' (Pfeifer, 1993) that also show emergent emotional behavior.

Although very simplistic, this approach can nevertheless make valuable contributions towards understanding emotional phenomena. First, it warns us of the risks of 'over-attribution' by humans of mechanisms and functional capabilities that we might posses but robots do not have. Second, it points to the fact that even our apparent complexity might in some cases be the result of 'over-design' in our theories and models. Third, it stresses the importance of adopting an incremental synthetic approach—building systems incrementally—that starts by implementing very simple architectures that are systematically investigated and incrementally complexified as/if needed, in order to understand the mechanisms underlying (emotional) behavior. However, it also presents what can be seen as a major drawback (see Cañamero, 2003 for a discussion): Modeling and implementing emotions as purely emergent phenomena in the eye of the beholder can be an exciting challenge for the designer of a robot, but this view misses a potentially important contribution of artificial emotional systems—their ability to relate to and influence different behavioral and cognitive subsystems at the same time.

Trying to address the latter problem, a second strand within the 'emergent' approach explores how behavior that an observer could term as 'emotional' emerges from the interactions of different underlying mechanisms. The emphasis here is also in avoiding modeling emotions by explicitly implementing the same notions that we use to describe them (e.g. computational elements representing descriptive 'intentional' terms); we should rather model them using mechanisms at a level 'below' the phenomenon that we want to study (Pfeifer, 1993). Promising, although still early, research is being undertaken in this direction using simulated 'hormones' or 'neurohormones' that modulate different aspects of the underlying control architecture of the robot, and as a result, the overall behavior of the robot. Most work has concentrated in modulation leading to the production of behavior related to the 'flee/fight' response (e.g. Avila-García & Cañamero, 2004, 2005; French & Cañamero, 2005; Neal & Timmis, 2003). Such architectures typically combine homeostatic control and neural networks modeled at different levels of abstraction.

3.2. Learning

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. Computational models of emotion-related learning in autonomous agents have also generally adopted this type of approach. Examples of systems that use reinforcement learning are provided by the work of Gadanho that uses a neural network architecture and reinforcement learning to implement an adaptive robot controller that learns to navigate in environments of varying level of difficulty (Gadanho & Hallam, 2001). The MITB architecture of Ventura, Custódio, and Pinto-Ferrera (2001), closely inspired by the 'movie-in-the-brain' idea introduced by Damasio (1999), uses reinforcement learning to select courses of action aiming at obtaining desirable states for an agent that learns to perform a pendulum balancing task.

One of the main problems underlying traditional reinforcement learning models in robotics is how to make those signals truly meaningful to the robot so that the learning process is more autonomous and grounded in

the control architecture. 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 simple model of learning addressing this issue was proposed by Blumberg (1996), who developed an animated dog, Silas T. Dog, modeled closely after ethological models of animal behavior; this character possesses internal variables that represent emotions (as well as motivational states such as hunger or thirst) and that can influence what he learns. Changes in the dog's internal variables drive a learning process, e.g. when the 'fear' variable increases, Silas tries to determine which stimuli from his perception and short-term memory can best predict the change, and use that association to learn new ways to behave, e.g. avoid places where he previously perceived objects that caused 'fear'. In this work, however, emotions are modeled in a very simplistic way as single variables. A more complex approach is exemplified by the work of Velásquez, who implemented fear conditioning in his robot Yuppy (Velásquez, 1998) using an associative network model. The architecture of Yuppy contains, in addition to perceptual, behavior, motor, and drive systems, a set of basic emotions. Simulated 'pain' signals, triggered when a person disciplines Yuppy, allow it to learn 'secondary' emotions, and in particular to associate a new cognitive releaser (the sound of a flute) when this is paired with the releaser that caused 'pain'.

A more biologically plausible approach would need to include a mechanism rooted in an internal 'value system' 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 punishment and positive reward, and allowing to learn appropriate valenced reactions to them. The works of Andry, Gaussier, Moga, Banquet, and Nadel (2001), Cos-Aguilera, Cañamero, and Hayes (2003) and Lahnstein (2005) provide initial solutions in this direction, where 'pain' and 'pleasure' signals are rooted in discomfort or well-being related to the stability (or lack of it) of other internal (homeostatic) aspects of the robot architecture—reward/punishment based on the success in predicting the rhythm of motor actions of another agent in an imitation task in the case of Andry and colleagues, reward/punishment based on the ability to correct the errors of survival-related variables setting internal needs by interacting with objects in the environment in the architectures of Cos-Aguilera and colleagues and Lahnstein. In the first two cases, such 'pleasure' and 'pain' signals are used to learn affordances as a result of interactions with the environment.

Other studies investigate emotional learning taking a stronger 'neuro-inspired' approach, such as Balkenius and Morén (2001) and Morén (2002). These authors have elaborated a computational model of emotional learning that uses neural networks to implement different brain areas in the 'emotion circuit', such as the amygdala, thalamus, sensory cortex, and orbitofrontal cortex and their interconnections to model emotional conditioning. However, such complex models have not been implemented in (real or simulated) robots.

3.3. Memory

Management of memory is another major problem in autonomous robots that have to timely select appropriate courses of action. 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 can help to solve some of these problems and also provide generally coherent responses to a wide range of situations. The work of Araujo (1994) addressed this problem, although to our knowledge it has not been implemented in robots. His model, inspired by the theories of LeDoux about the dual route of emotion processing (LeDoux, 1989, 1996), tries to integrate low-level physiological emotional responses and their high-level influences on cognition by using two neural networks—'emotional' and 'cognitive'. Interactions between these two networks imitate mood-congruent memory retrieval and learning and the effects of anxiety on memory performance. A simpler robotic model addressing aspects of emotional memory is that of Velásquez mentioned in Section 3.2 (Velásquez, 1998), which uses an associative network with a modified Hebbian rule to form emotional memories or 'secondary emotions' related to the 'joy' and 'fear' subsystems when a person interacts with the robot by petting or disciplining it.

4. Shared conceptual problems

The development of robotic architectures that model and operationalize different aspects of emotions and their involvement in cognitive and behavioral processes in artifacts carries many parallels with the problems investigated by emotion theories and models. Let us examine some of them.

4.1. Mechanisms underlying the involvement of emotions in cognition and action

A key issue that must obviously be addressed is an investigation of possible mechanisms that allow emotional phenomena to influence cognitive and behavioral processes and how they can be implemented in robots. Different emotion theories and models put the emphasis on diverse aspects of emotional phenomena and the different

mechanisms underlying them and this often implies different ways to conceptualize the link between emotion and cognition/action. For example, some models such as 'circuit models'[2] (e.g. Panksepp, 1998; Rolls, 1999) and 'adaptational models' (e.g. LeDoux, 1996) postulate specific 'emotion centers' or 'neural circuits' in the brain, particularly or primarily concerned with the processing of emotion-related information and the production of emotional responses. Other models such as Fellous (1999, 2004), related to 'peripheral feedback models' (e.g. Damasio, 1999), consider emotions as dynamical patterns of neuromodulations rather than patterns or circuits of neural activity. These approaches clearly entail different ways of conceptualizing the link between emotions and cognition and action. In the former case, this involves establishing connections between specialized 'emotion circuits' and other brain circuits and areas primarily involved with different cognitive and behavioral functions. In the latter, patterns of neuromodulations directly affect brain areas involved at all levels of (cognitive and behavioral) functions and the extent to which emotions can affect different aspects of cognition and action depends on the potential for neuromodulation of the neural substrate involved in those different aspects.

The operationalization of these approaches in robots gives rise to different computational models and mechanisms. For example, a computational model inspired by 'circuit' and 'adaptational' models such as Balkenius and Morén (2001) and Morén (2002) typically uses neural networks to explicitly implement different brain areas in the 'emotion circuit' such as the amygdala, thalamus, sensory cortex, and orbitofrontal cortex and their interconnections in a 'cognitive task' such as learning (emotional conditioning in this case). Such models, closely inspired from neuroscientific findings, seem very promising to study specific 'circuits' and the influence of particular emotional subsystems in individual skills such as (various types of) learning. However, integration of different 'circuits' and of emotion–cognition–action interactions at a more global scale is a very difficult task. On the contrary, a robotic architecture inspired by a 'neuromodulation' model such as those proposed by Brooks and Viola (1990) Avila-García and Cañamero (2004) and French and Cañamero (2005) will not try to recreate explicitly 'emotion centers' but it rather models different emotional processes in terms of simulated 'hormones' that affect the functioning of different elements of the architecture—the robot's 'nervous system'—modulating different cognitive and behavioral functions such as perception, attention, motivational priorities, behavior selection, and behavior execution. Such models are inspired from the functioning of human or animal brains at a more abstract level and therefore cannot make significant contributions regarding how emotion–cognition–action interactions are processed in precise areas of the human brain. However, they permit an easier integration of emotion–cognition–action interactions and a more systematic study of the 'global picture' (emotional influences in different aspects of the control architecture and how they affect the overall behavior of the robot); therefore, they can provide valuable insights towards understanding emotion–cognition–action interactions, even if it is not in a nervous system of human complexity but rather those of much more 'evolutionary primitive' species.

[2] In labeling these different emotion models, I follow the classification proposed by Scherer (2004).

4.2. Emotion elicitors

The study of mechanisms underlying the involvement of emotions in cognition and action in robotic models does not necessarily shed light regarding what mechanisms must be in place for 'external' and 'internal' influences to activate or produce an emotion in the first place. Researchers in this area often cite Izard's theory of four emotion elicitors (Izard, 1993)—neural/neurochemical, sensorimotor, motivational and cognitive—but very few have attempted to implement the four elicitors simultaneously (see e.g. Velásquez, 1996). Some of the problems that their implementation involves are, for example, establishing the causal relations among the different elicitors, or deciding which computational approaches and techniques are better suited to implement each of them and how they can be integrated. Appraisal theories are another approach that has become a very popular source of inspiration for computational models, in particular the more 'cognitive-oriented' ones. However, the two problems mentioned above reappear in most cases since, with few exceptions (e.g. Scherer, 2001; Smith, 2004) the gap between the level of abstraction of the theory and the concrete decisions needed for their implementation is too big, and engineers are left to their own intuitions in the search for solutions to bridge that gap—solutions that, understandably, are often ad hoc and driven by the particular needs of the application. Guidelines are needed for the operationalization of these theoretical models and their elaboration necessitates a joint collaborative effort between theorists and computer scientists. In particular, feedback from neuroscience is needed regarding how 'valence' is processed in the brain and how this can be meaningfully transposed to robots and regarding the interplay between emotion and attention.

4.3. Emotions as cognitive modes

In humans, emotions entail distinctive integrated ways of perceiving and assessing situations, processing information, and modulating and prioritizing actions. In this respect, emotions can be seen as different 'cognitive modes' that have a 'global' and synchronized influence in our perceptual, cognitive, bodily and behavioral relation with the world. Achieving this in computational emotion

architectures involves a number of challenging problems, many of them largely unexplored, such as:

- Which aspects of cognition need to be in place in the architecture to be able to speak of a 'cognitive mode'?
- What mechanisms can be used to modulate different aspects of perception and cognition?
- What mechanisms are required to implement the different effects of various emotions?
- How can computational and robotic models take into account cultural and individual differences in the synthesis of emotions as 'cognitive modes'?
- Which are the causal relations between the different subsystems involved?
- Which (computational) mechanisms allow the integration of the 'fast' and 'slow' pathways in the processing of emotion-relevant information? How can this be done to achieve timely and emotionally adequate behavioral responses in autonomous robots?
- How to synchronize the effects of emotions on the different computational (cognitive, bodily and behavioral) subsystems involved in order to obtain the 'global' effect that emotions have in humans?
- How can we model the influence that emotions have in the perception of social partners?
- In biological systems, 'one of the main functions of emotion is to achieve a multilevel communication of simplified but high-impact information' (Fellous, 2004, p. 39). However, on what grounds can an autonomous robot assess what constitutes 'high-impact' information? In other words, what kinds of mechanisms (value systems) are needed to make information 'emotionally relevant' for an autonomous robot? How can 'value systems' be implemented as an integral part of the architecture in order to ground 'emotional meaning' for the robot?
- How can we model the relation between the 'cognitive modes' and the action tendencies involved by emotions in different architectures?

4.4. Emotions, value systems, motivation, and action

These notions are closely related in multiple ways and different control architectures have implemented these different aspects. In emotion synthesis, emotions play important roles in relation to the production of action in autonomous robots (see e.g. Cañamero, 2003), e.g.:

- The fact that emotions are related with ('general') goals or concerns rather than with particular behavioral response patterns (emotions versus goal-directed behavior), explains the fact that they allow to generate richer, more varied and flexible behavior.
- They can be conceptualized and modeled as 'second-order' control mechanisms that constantly monitor the internal and external environment to detect and respond to potential 'threats' of different sorts, therefore either sustaining or interrupting ongoing goal-directed behavior.
- They modify/amplify motivation, producing changes in motivational/goal priorities to deal more efficiently with certain types of relevant (survival-related) events.
- They can also constitute motivational factors and constitute 'value systems' that affect the selection of goals and goal-directed behavior.

The implementation of these functions in computational architectures poses, once more, non-trivial problems regarding the choice of underlying mechanisms and the integration of motivational, behavioral, and emotional components of the architecture. In addition, the link among these components must be grounded in some sort of 'value systems' to permit the autonomous generation of valenced reactions that characterize emotions and distinguish them from 'cognitions'. Another minor problem is that of assessing (measuring quantitatively) the benefits that the effects of emotions carry for the performance of the robot. This requires the development of different performance indicators (see e.g. Avila-García & Cañamero, 2002, Avila-García, Cañamero, & te Boekhorst, 2003 for initial performance indicators drawn from viability theory and ethology) and a systematic understanding of the problems relevant for the understanding of emotions posed by different types of environments. Regarding this latter point, some studies of properties of environments relevant for action selection activities in autonomous robots have been carried (e.g. Cañamero, Avila-García, & Hafner, 2002; Maes, 1995; Tyrrell, 1993) that could be extended to focus on environmental properties relevant to different aspects of emotions; here, robotic environments offer the advantage of permitting very systematic investigation of environments.

5. Challenges and goals for future research

Taking into account the need to dissolve the bottlenecks mentioned in Section 2 and extrapolating from the key developments and problems presented in previous sections, this section outlines some research directions that I consider as key development goals in the area in order to make more sound contributions towards a joint multidisciplinary effort to understand emotions. These 'challenges' reflect problems rooted in our philosophical tradition and are also present in neuroscience research. Approaching them from different perspectives should hopefully provide mutual insights among different disciplines.

5.1. The 'origins' and grounding problem of artificial emotions

To date, the design of most synthetic emotional systems is based on the intuitions of the designers of those systems

regarding the choice of primitives or building blocks (emotion 'components', 'modules' or 'features') that constitute the emotional system. In most cases those building blocks are 'hardwired' (designed manually) by the designer, taking inspiration from characteristics and functions of emotions in humans; such 'building blocks' are also labeled after terms used in the psychology and neuroscience of human emotions. While this approach, based on elements and terms already familiar to us, fosters understandability of the system, as put forward by Wehrle (2001), it carries two major dangers:

1. The risk of 'over-attribution' by human users of functions and capabilities implied by those terms that humans posses but robots do not have—a good example being the attribution of 'feelings' and conscious subjective states to robots.
2. The lack of 'grounding' of those emotion components, since they were included in the system on the grounds of their meaning for the human designer and user rather than due to their meaning for the robot in interaction with its (physical and social) environment.

Within this approach, avoiding the risk of over-attribution requires honesty and transparency by the designer regarding the mechanisms underlying these systems, in particular when presenting work to the layperson. As for emotion grounding, it requires a sound investigation and understanding of the reasons for the inclusion of those particular 'emotion primitives' in the architecture (in terms of their functions) and of the mechanisms for their integration with other elements of the architecture (in terms of their integration with different cognitive and behavioral subsystems).

Two other lines of research, little developed so far, can also greatly contribute to avoid those two dangers. The 'emergent' approach to emotion modeling that we reviewed in Section 3.1.2 and in which behavior that an observer could consider as 'emotional' but that arises from the interactions of the system with its environment, rather than from explicit 'emotion components' in the architecture, can contribute to avoid the risk of 'over-attribution'. It can also improve our understanding of emotional phenomena (and avoid 'over-design', i.e. the elaboration of systems and theoretical models unnecessarily complex) by uncovering some aspects of emotions that can be accounted for by simple cognitive and behavioral mechanisms and their interactions without the need to postulate specific 'emotion machinery'.

Emotion grounding can be better achieved by developing computational models of emotions that take developmental and evolutionary perspectives, in which emotional systems form (or 'grow') in the course of the interactions of the artifact with its (physical and social) environment over the life time of an individual (development) or a 'species' (evolution). In these cases, emotions acquire a meaning not only for the human designer and user, who would be able to track and understand the reasons behind those particular emotional systems, but also for the robot itself.

5.2. Dissolving the 'mind–body' problem

Related to the grounding problem of emotions in artifacts is the problem of investigating and establishing well-founded links between 'higher' and 'lower' levels of cognition and action and the influences of emotions in both. Different conceptual traditions in AI, which can be roughly classified into symbolic and embodied approaches, put the emphasis on these different 'higher' and 'lower' levels. Symbolic AI models emotion as perceived by introspection, using rule-based symbol systems, and taking inspiration from psychological theories. Embodied AI focuses on 'lower-level' aspects of emotion related to their embodiment, takes inspiration from biology and neuroscience, and uses tools from dynamical systems theory, neural networks, behavior-based robotics, etc. At present, embodied approaches are not enough developed to be able to model 'higher-level' aspects of emotions, and current attempts to build 'complete' emotion architectures use 'hybrid' models that integrate a 'deliberative' component that models emotion-based reasoning and appraisal as accessible to introspection, and an embodied 'reactive' component that generates behavior arising from 'fast' emotional response (see e.g. Gratch & Marsella, 2004; Petta, 2003). However, sound progress towards solving this problem needs to go beyond the type of solutions that current state-of-the-art research allows to provide, and that often reminds us too much of the 'pineal gland' that Descartes postulated as a link between 'body' and 'mind'. Ideally, future systems should not provide 'hybrid' solutions but rather a true integration that departs from a dualist stance and dissolves the 'mind–body' problem. This task is not an easy one, since this problem has been pervasive in philosophy for centuries and is still open. However, this goal should be regarded as an ideal to guide research. In particular, a number of problems need to be addressed if we want to progress towards more sound solutions, such as:

- The roles that emotions play in the synchronization of numerous cognitive, behavioral and bodily subsystems.
- The mechanisms needed to bridge the gap between the 'internal' and 'external' aspects of emotions in order to synthesize expressive behavior truly grounded in the architecture of the robot (see Cañamero & Gaussier, 2005 for a discussion of this latter problem).
- The integration of multiple levels of emotion generation.

An interesting research avenue to explore in this respect would be the use of computational models of emotions that try to bridge the gap between 'lower-level' and 'higher-level' aspects of cognition and action, e.g. by using emotional aspects of the architecture to synchronize

the functioning of different behavioral and cognitive components.

5.3. Untangling the 'knot of cognition': The links between emotion and intelligence

The surge of research in computational models of emotions and affective computing in general is often conceptualized as a consequence of a change in paradigm regarding the role of emotions in intelligence; this surge was to a big extent triggered by the popularization of neuroscientific models such as those of Damasio and LeDoux. Overcoming the tradition that regarded emotions as undesirable consequences of our embodiment that impaired reasoning and decision making, nowadays, emotions are considered as pervasive in many aspects of cognition and action and an essential element of intelligence. Even if we agree with the latter position, this view should not become an unquestioned assumption that engineers use as an argument by authority, as this would give rise to rather superficial computational models of emotions and poor understanding of our achievements. On the contrary, sound progress in the area necessitates 'dissecting' or 'untangling' this 'Gordian knot' by investigating and modeling the mechanisms underlying different aspects of the involvement of emotions in cognition and action; it also needs carefully assessing which aspects, among the many possible, can be meaningfully modeled in our robots and which ones are specific to human or animal emotional systems, and therefore meaningless in the case of robots. Mechanisms underlying emotional modulation of different aspects of cognition and action need to be singled out, in close collaboration with emotion scientists (e.g. neuroscientists, psychologists), and implemented in robots.

This is not sufficient, however, as emotions are not 'isolated' or 'independent' entities or modules within the brain or the robot architecture, but are deeply intertwined and rely on many other aspects and subsystems; therefore, these elements must be investigated and modeled in parallel with emotions. Some of the relevant notions are, for example (see Cañamero, 2001a for a discussion): The notion of 'self', of which only some rudiments around the ideas of 'bodily self' (e.g. proprioceptive feedback) and 'autobiographic self' (e.g. autobiographic and emotion-related memory) can be meaningfully implemented in robots; mechanisms for social motivation; or mechanisms underlying simple forms of 'empathy' (e.g. mechanisms of emotional contagion such as sensory–motor coordination and synchronization, mimicry, feedback, etc.).

5.4. Measuring progress: Which are the contributions of emotions to our systems?

An important requisite for the advancement of the area is the ability to assess the benefits that the inclusion of emotional systems brings to our robots. The fact that emotions seem to fulfill a number of important functions in humans and provide increased complexity and flexibility of behavior in natural systems cannot be used to justify the value of artificial emotional systems. The inclusion of emotional elements in the architecture of our robots does not make them more valuable per se. On the contrary, we must be able to show accurately and precisely that (or rather whether) our results allow us to conclude that emotions improved the performance or the interaction capabilities of our robot and how. An obvious way of doing this is by running control experiments in which the robot performs the same task 'with' and 'without' emotions and comparing the results. Although some criteria and indicators of performance have been proposed, much research is still needed to develop different criteria, performance indicators, scenarios, testbeds, etc., to achieve not only qualitative but also quantitative evaluations of our robots adapted to different types of applications and problems.

6. Conclusion

In this paper, I have discussed some of the contributions that modeling emotions in autonomous robots can make towards understanding human emotions—'as sited in the brain' and as used in our interactions with the environment—and emotions in general. Such contributions are linked, on the one hand, to the potential use of such robotic models as tools and 'virtual laboratories' to test and explore systematically theories and models of human emotions, and on the other hand to a modeling approach that fosters conceptual clarification and operationalization of the relevant aspects of theoretical notions and models. As illustrated by my overview of the main approaches towards modeling emotions in robots, the field is still in its infancy. However, the work carried out already shows that we share many conceptual problems and interests with other disciplines in the affective sciences. Sound progress that allows us to build not only robust engineering systems but also emotion systems that can provide feedback to emotion theorists necessitates collaborative efforts. I would thus like to conclude by inviting neuroscientists, psychologists, philosophers, and other affect theorists to join forces with roboticists in a common effort to understand emotions—human, animal, and robot emotions.

Acknowledgements

This research is partly funded by the EU FP6-IST Network of Excelence HUMAINE under contract 507422. This paper reflects solely the views of the author and not those of the consortium and it is partly based on her contributions to the HUMAINE workshop 'Theories and Models of Emotion' and to the project technical report FP6-2002-IST-1-507422-D7b. I am grateful to

the workshop participants for fruitful discussions and to Orlando Avila-García, Rene te Boekhorst and Ignasi Cos-Aguilera for their contributions to parts of the research presented here.

References

Andry, P., Gaussier, P., Moga, S., Banquet, J. P., & 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.

Araujo, A. (1994). Memory, emotions, and neural networks: Associative learning and memory recall influenced by affective evaluation and task difficulty. PhD Thesis, University of Sussex, UK, May 1994.

Ashby, W. R. (1952). *Design for a brain*. London: Chapman & Hall.

Avila-García, O., & Cañamero, L. (2002). A comparison of behavior selection architectures using viability indicators. In *Proceedings of international workshop on biologically-inspired robotics: The legacy of W. Grey Walter* (pp. 86–93). HP Labs, Bristol, UK, 14–16 August, 2002.

Avila-García, O., & Cañamero, L. (2004). Using hormonal feedback to modulate action selection in a competitive scenario. In S. Schaal, A. J. Ijspeert, A. Billard, S. Vijayakumar, J. Hallam, & J.-A. Meyer (Eds.), *From animals to animats 8: Proceedings of the 8th international conference on simulation of adaptive behavior (SAB'04)* (pp. 243–252). Cambridge, MA: The MIT Press, 243–252.

Avila-García, O., & Cañamero, L. (2005). Hormonal modulation of perception in motivation-based action selection architectures. In L. Caamero (Ed.), *Agents that want and like: Motivational and emotional roots of cognition and action. Papers from the AISB'05 symposium* University of Hertfordshire, UK, April 14–15, 2005. SSAISB Press.

Avila-García, O., Cañamero, L., & Boekhorst, R. (2003). Analyzing the performance of 'Winner-Take-All' and 'Voting-Based' action selection policies within the two-resource-problem. In *Proceedings of seventh European conference in artificial life (ECAL03)*, (pp. 733–742). Berlin: Springer.

Balkenius, C., & Morén, J. (2001). Emotional learning: A computational model of the amygdala. *Cybernetics and Systems*, *32*(6), 611–636.

Bates, J. (1994). *The role of emotion in believable agents TR CMU-CS-94-136*. Pittsburgh, USA: School of Computer Science, Carnegie Mellon University.

Blumberg, B. (1996) Old tricks, new dogs: Ethology and interactive creatures. PhD thesis, MIT Media Lab, September 1996.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: The MIT Press.

Breazeal, C. (2002). *Designing sociable machines*. Cambridge, MA: The MIT Press.

Brooks, R. A., & Viola, P. A. (1990). Network based autonomous robot motor control: From hormones to learning. In R. Eckmiller (Ed.), *Advanced neural computers* (pp. 341–348). Amsterdam: Elsevier.

Cañamero, L. (1997). Modeling motivations and emotions as a basis for intelligent behavior. In W. L. Johnson (Ed.), *Proceedings of first international conference autonomous agents* (pp. 148–155). New York, NY: ACM Press.

Cañamero, L. (2001a). Building emotional artifacts in social worlds: Challenges and perspectives. In L. Cañamero (Ed.), *Emotional and intelligent II: The tangled knot of social cognition. Papers from the 2001 AAAI Fall Symposium* (pp. 22–30). Menlo Park, CA: AAAI Press.

Cañamero, L. (2001b). Emotions and adaptation in autonomous agents: A design perspective. *Cybernetics and Systems*, *32*(5), 507–529.

Cañamero, L. (2003). Designing emotions for activity selection in autonomous agents. In R. Trappl, P. Petta, & S. Payr (Eds.), *Emotions in humans and artifacts* (pp. 115–148). Cambridge, MA: MIT Press.

Cañamero, L., Avila-García, O., & Hafner, E. 2002. First experiments relating behavior selection architectures to environmental complexity. In *Proceedings of 2002 IEEE/RSJ international conference on intelligent robots and systems (IROS 2002),* (pp. 3024–3029). IEEE Press.

Cañamero, L., & 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.

Cañamero, L., & Gaussier, P. (2005). Emotion understanding: Robots as tools and models. In J. Nadel, & D. Muir (Eds.), *Emotional development: Recent research advances* (pp. 235–258). New York, NY: Oxford University Press.

Cos-Aguilera, I., Cañamero, L., & Hayes, G. (2003). Learning object functionalities in the context of behavior selection. In U. Nehmzow, & C. Melhuish (Eds.), *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.

Damasio, A. (1999). *The feeling of what happens: Body, emotion and the making of consciousness*. London: Vintage.

Fellous, J.-M. (1999). The neuromodulatory basis of emotion. *The Neuroscientist*, *5*, 283–294.

Fellous, J.-M. (2004). From human emotions to robot emotions. In E. Hudlicka, & L. Cañamero (Eds.), *Architectures for modeling emotions: Cross-disciplinary foundations. Papers from the 2004 AAAI Spring Symposium* (pp. 37–47). Menlo Park, CA: AAAI Press.

French, R., & Cañamero, L. (in press). Introducing neuromodulation to a Braitenberg vehicle. In *Proceedings of the IEEE international conference on robotics and automation—robots get closer to humans (ICRA 2005)*, (pp. 4199–4204) Barcelona, Spain, April 18–22. IEEE Press.

Frijda, N. (1995). Emotions in robots. In H. L. Roitblat, & J.-A. Meyer (Eds.), *Comparative approaches to cognitive science* (pp. 501–516). Cambridge, MA: The MIT Press.

Gadanho, S. C., & Hallam, J. (2001). Emotion-triggered learning in autonomous robot control. *Cybernetics and Systems*, *32*(5), 531–559.

Gratch, J., & Marsella, S. (2004). A domain-independent framework for modeling emotion. *Cognitive Systems Research*, *5*(4), 269–306.

Izard, C. E. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. *Psychological Review*, *100*(1), 68–90.

Lahnstein, M. (2005). The emotive episode is a composition of anticipatory and reactive evaluations. In L. Cañamero (Ed.), *Agents that want and like: Motivational and emotional roots of cognition and action. Papers from the AISB'05 Symposium,* (pp. 62–69) University of Hertfordshire, UK, April 14–15, 2005. SSAISB Press.

LeDoux, J. (1989). Cognitive–emotional interactions in the brain. *Cognition and Emotion*, *3*, 267–289.

LeDoux, J. (1996). *The emotional brain*. New York: Simon & Schuster.

Lowe, R., Cañamero, L., Nehaniv, C., & Polani, D. (2004). The evolution of affect-related displays, recognition and related strategies. In J. Pollack, M. Bedau, P. Husbands, T. Ikegami, & R. A. Watson (Eds.), *ALIFE IX: Proceeding of the 9th international conference on the simulation and synthesis of living systems* (pp. 176–181). Cambridge, MA: The MIT Press.

Maes, P. (1995). Modeling adaptive autonomous agents. In C. G. Langton (Ed.), *Artificial life: An overview* (pp. 135–162). Cambridge, MA: The MIT Press.

Morén, J. (2002). Emotion and learning: A computational model of the amygdala. PhD thesis, *Lund University Cognitive Studies*, 93.

Neal, M., & Timmis, J. (2003). Timidity: A useful emotional mechanism for robot control? *Informatica*, *27*, 197–203.

Ortony, A. (2003). On making believable emotional agents believable. In R. Trappl, P. Petta, & S. Payr (Eds.), *Emotions in humans and artifacts* (pp. 189–211). Cambridge, MA: MIT Press.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. New York, NY: Oxford University Press.

Petta, P. (2003). The role of emotions in a tractable architecture for situated cognizers. In R. Trappl, P. Petta, & S. Payr (Eds.), *Emotions in humans and artifacts* (pp. 251–287). Cambridge, MA: The MIT Press.

Pfeifer, R. (1993). Studying emotions: Fungus eaters. In *Proceedings of the first european conference on artificial life (ECAL'93)*, (pp. 916–927). ULB, Brussels, Belgium, May 24–26.

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

Reeves, B., & 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.

Rolls, E. T. (1999). *The brain and emotion*. New York: Oxford University Press.

Scherer, K. (2001). Appraisal considered as a process of multilevel sequential checking. In K. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion* (pp. 92–120). New York, NY: Oxford University Press.

Scherer, K. (2004). *Preliminary plans for exemplars: Theory*. HUMAINE deliverable FP6-2002-IST-1-507422-D3c. Available at http://emotion-research.net/deliverables.

Smith, C. A. (2004). A functional perspective on emotion elicitation: Some considerations for the development of emotional architectures. In E. Hudlicka, & L. Cañamero (Eds.), *Architectures for modeling emotion: Cross-disciplinary foundations. Papers from the 2004 AAAI spring symposium* (pp. 135–143). Menlo Park, CA: AAAI Press.

Stern, A. (2003). Creating emotional relationships with virtual characters. In R. Trappl, P. Petta, & S. Payr (Eds.), *Emotions in humans and artifacts* (pp. 333–365). Cambridge, MA: MIT Press.

Tyrrell, T. (1993). Computational mechanisms for action selection. PhD Thesis, Centre for cognitive sciences, University of Edinburgh. Available at http://www.cs.bham.ac.uk/sra/People/Stu/Tyrrell.

Velásquez, J. (1998). Modeling emotion-based decision making. In L. D. Cañamero (Ed.), *Emotional and intelligent: The tangled of knot of cognition. Papers from the 1998 AAAI Fall Symposium* (pp. 164–169). Menlo Park, CA: AAAI Press.

Velásquez, J.D. (1996). Cathexis: A computational model for the generation of emotions and their influence in the behavior of autonomous agents. Master's thesis, MIT Media Lab, September 1996.

Ventura, R., Custódio, L., & Pinto-Ferrera, C. (2001). Learning courses of action using the 'movie-in-the-brain' paradigm. In L. Cañamero (Ed.), *Emotional and intelligent II: The tangled knot of social cognition. Papers from the 2001 AAAI Fall Symposium* (pp. 147–152). Menlo Park, CA: AAAI Press.

Wehrle, T. (2001). The grounding problem of modeling emotions in adaptive artifacts. *Cybernetics and Systems: An International Journal*, *32*(5), 561–580.