# Building Emotional Artifacts in Social Worlds: Challenges and Perspectives

**Lola D. Cañamero**
Adaptive Systems Research Group
Dept. of Computer Science, University of Hertfordshire
College Lane, Hatfield, Herts AL10 9AB, UK
http://homepages.feis.herts.ac.uk/∼comqlc
L.Canamero@herts.ac.uk

**Abstract**

This paper discusses ideas relative to the construction of emotional artifacts that have to interact in a social world, and in particular with humans. It first examines some of the ways in which emotions can enhance social interactions with artifacts, and some of the challenges posed to the designer. After considering the debate that opposes "shallow" versus "deep" modeling, it sketches some ways in which we can anchor emotions in the architecture of artifacts in order to make emotional interactions meaningful not only to the human, but also to the artifact itself. It finally outlines some of the cognitive capabilities that artifacts should incorporate for their emotions to be properly grounded and to give rise to rich social exchanges with humans.

## Introduction

The social and the emotional are highly intertwined. For some researchers, emotions come into play (primarily for some, exclusively for others) as soon as we consider individuals in interaction with their social environment. For others, emotions are at the very heart of what being social means. Some opinions even establish a relation of "identity" between these notions, somewhat like the two sides of a coin (Dumouchel, 1999). In humans, emotions influence and shape the development of sociality as much as sociality influences and shapes the development of emotions.

This strong interdependence is being increasingly echoed in AI research. In this sense, the area of socially intelligent agents (Dautenhahn, 1998) has witnessed over the last years a growing interest in the involvement of emotions in social interactions. Likewise, the emotion modeling community has recently devoted much effort to the design and implementation of models of emotions for social interactions.

Artifacts are not humans, though, and one can always question whether social interactions involving artifacts need/can benefit from the consideration of emotional aspects. For artifacts interacting with other artifacts, my arguments would be very much in the line of those provided in (Cañamero, 1998) for individual emotional agents. In this paper, I will only consider interactions involving a human. Another question that naturally arises is what aspects of emotions are relevant to social interactions between artifacts and humans: Are observable behavioral features enough, or is an underlying model for emotion synthesis needed? In this latter case, how can we ground emotions in the architecture of the artifact so as to give rise to meaningful interactions? Can we take inspiration from human emotion research for this? This paper discusses some personal thoughts on all these questions.

## Emotions and Social Artifacts

What can artificial emotions contribute to social interactions between artifacts and humans? What are the major challenges that their design poses?

### Contributions of emotions

Intuitively, we can think of different roles that emotions can play in social interactions between artifacts and humans. Let us mention some of them.

- *Conveying intentionality.* People need to understand the behavior observed in their social partners—human or artificial—as resulting from causes or intentions that allow them to form coherent explanations of their observations. This coherence is necessary to interpret past or current relations, make predictions and establish expectations about future behavior. Emotions and personalities are often postulated as causes of behavior and as sources of intentions when explaining the behavior of other humans and animals. One could thus expect that they will be used in the same way when interacting with artifacts. Autonomous artifacts can, 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, artifact's emotions can be used with the same purpose, seeking responses that either match the artifact'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.* Artifacts able to express emotions and adapt their interactions to the emotional state of their partners can 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, though, is that emotional behavior and expressions make the artifact more believable, as it in a sense perceived as "closer" or more similar to ourselves (or at least to a living being).
- *Enhanced communication.* Emotional expression being a key element in non-verbal communication, endowing an artifact with emotional expressions can make communication cognitively less expensive for the human partner. If emotional artifacts are to achieve a sufficient level of sophistication at some point in the future, we could also dream of communicating with them at a "deeper" level of understanding. For example, provided that one day they could be able to interpret our subtle expressions and obtain relevant information from contextual clues, we would also like that they "understand" what we *mean*, not (only) what we say[1].

## Design challenges

Let us now consider some of the issues that set important challenges to the designer of emotional artifacts for social interactions.

**Which emotions?** Most projects implementing emotions or emotional expressions in artifacts have opted for the basic emotions subset, usually the set proposed by Ekman in (Ekman, 1992). These emotions seem to be generally better recognized by both humans and machines, and are also easier to synthesize in artifacts. Whereas basic emotions can be very adequate for simple social interactions (e.g. with children), they can be somewhat restrictive to achieve more complex interactions. Some projects with a basic categories approach allow artifacts to express emotion blends, such as (Cañamero & Fredslund, 2001), but in principle dimensional or componential approaches (Russell, 1997; Smith & Scott, 1997) allow to build artifacts with much more varied expressions, such as in (Breazeal, 2002). On the positive side, this wider variety makes it possible to engage humans in more socially rich interactions. The drawback can however be that an artifact with very rich emotional states and expressions could be too (cognitively and emotionally) demanding for the human partner.

Ideally and technology permitting, we could also wish that artifacts be able to recognize our subtle emotions, as human adults rarely express clear basic emotional states, or even to distinguish between "honest" and "fake" emotional displays in some cases. On the other hand, going too far on this raises privacy issues such as to what extent we want (and whether it is ethical) that artifacts recognize our subtle or purposefully hidden emotions.

**Inspiration from nature?** To what extent can we take inspiration from (theories and models of) human and animal emotions to design emotional artifacts? By this question I mean, on the one hand, whether this is possible, and on the other hand, whether it is desirable.

Is it possible? In spite the strong debates and disagreements among researchers of human emotions that might make one feel overwhelmed when exploring the literature in search of inspiration, I believe it is not only possible but also very useful to identify and isolate (fragments of) theories addressing problems that can inspire designs to solve problems in AI and robotics. Conversely, this practice can also provide feedback to human emotion researchers, as it offers them tools to test, analyze, and compare their theories, as argued in (Cañamero, 2001).

Is it desirable? My answer to this is that whereas taking nature and theories about natural emotions as sources of inspiration can be very inspiring and useful, as argued above, going beyond a certain point can be useless or even have negative consequences. Extreme realism is not devoid of dangers such as humans overattributing emotional capabilities to artifacts, and misunderstanding, delusion and frustrated expectations as possible consequences of this. A robot is a robot, not a human. We, designers, want humans to connect with emotional artifacts and feel comfortable in social interactions, but we should avoid fooling them into believing that they are interacting with another human. Besides these ethical considerations, other more methodological concerns are also at stake. For example, is it sensible to implement in artifacts the same methods that humans use when perceiving, displaying or recognizing emotions? In my opinion, only empirical investigation can provide answers to this.

**Human acceptance** A fundamental concern is whether humans are willing to accept and trust emotional artifacts as social partners, and how these artifacts should be designed to favor acceptance by humans. Two key factors for this seem to be believability, and that the human feels that s/he is (and that s/he *really* is) in control of the interaction at critical points. These two factors depend on many other variables, such as the personality of the human partner, the application domain where the artifact is put to work, the cultural and social context, etc. Designing "generic" emotional artifacts—i.e. artifacts embedding general emotion models that do not take into account individual differences—seems thus to have clear limitations. It could be more advisable (although much more difficult) to model emotions adapting them to types of "profiles" that take all these factors into account, and also reflect the evolution of emotional interactions over time—i.e. their "history".

[1] I owe this sentence to Jakob Fredslund

## Surface or Beyond?

> "...the concepts discussed in this article characterize the psychological basis of believability. Storytelling, empathy, historical grounding (autobiography), and "ecological grounding" in an environment are identified as factors relevant to the way humans understand the (social) world. [...] It is hoped that approaches to achieving believability based on these concepts can avoid the "shallowness" and "cheating" of approaches to believability that merely take advantage of the anthropomorphizing tendency in humans." (Dautenhahn, 1998, p. 574)

The same division of opinions that Dautenhahn portrays in the area of socially intelligent agents can be seen in the emotion modeling community: Should we model (only) the observable, surface aspects of emotions, or must we go for a "deep" modeling of a "true" emotional system beyond surface? Which is possible, necessary, and/or sufficient?

Whereas emotion modeling (synthesis) for individual agents has generally focused on the design of architectures to endow these agents with "true" emotional systems, the design of emotions for agents interacting socially has primarily paid attention to the external or "surface" features of emotional expression and interaction. Believability being a major issue as soon as humans are involved in (social) interactions with artifacts, the question can be rephrased as: What makes an emotional artifact believable for social interactions?

The features that allow humans to *perceive* the emotional displays of the agent as believable enough to elicit appropriate responses from them seem to be the most apparent answer. One would immediately think of general expressive features, such as something resembling a face (human or not) with some typical elements like (moving) eyes, eyebrows, or a mouth; patterns of movement; posture; timely responses; inflection of speech; etc. Other expressive features more related to social interaction include turn-taking during the interactions; direction of gaze; eye-to-eye contact; joint attention; etc. In theory, all these features could be modeled taking a "shallow" approach—which does not make their implementation more easy or trivial. However, such an approach makes it very difficult to maintain believability over prolonged interactions, as it is unlikely that the *behavior* of the artifact will remain coherent over sustained periods of time, given the complexity of these interactions. Coherence of behavior is an important issue, as humans want to understand and explain observed (expressive) behavior as the result of some underlying causality or intentionality for the emotional interaction with the artifact to be believable and acceptable to them. This can only be properly achieved if expressive behavior is guided by some underlying system of emotion synthesis (and ideally of personality as well).

I adhere thus to the opinion that the believability of emotional displays and interactions can be better achieved, for non trivial social interactions, if it is sustained by a "deeper" emotion system grounded in the architecture of the artifact. This is not to say, however, that I consider "shallow" approaches as uninteresting or "cheating". The fact that the human tendency to anthropomorphize is so pervasive and compelling that makes us treat our TV set and computer like people (Reeves & Nass, 1996) makes it worth studying it with the possibilities that emotional artifacts offer. Much can be learned about human emotions and emotional interactions from projects that heavily rely on the human tendency to anthropomorphize, such as the expressive robots Sparky (Scheeff *et al.*, 2001), Kismet (Breazeal, 2002) or Feelix (Cañamero & Fredslund, 2001; Cañamero, 2001). On the human side, they can help us identify key elements that shed light on what triggers in humans the tendency to anthropomorphize and what makes emotional behavior and displays believable to the human eye. On the artifact side, they can provide very valuable feedback for the design of artifacts (e.g. their morphology) that can interact socially and emotionally in a way that is better adapted and more natural to humans.

The choice between one approach and the other will thus depend on what we are interested in learning about emotions, and on the application foreseen for the artifact.

## Anchoring Emotions in Artifacts

If we want our emotional artifacts to go beyond surface emotional features and displays to "have" emotions (in some sense) and deeper emotional interactions, we must find ways to anchor emotions in them that are suited to the artifact's structure and dynamics of interactions, as well as to the other partners (including humans) in their social world. This anchorage can be understood in a weaker or a stronger sense.

In its weaker sense, it can be taken to mean endowing artifacts with some components or modules that explicitly represent some elements of emotions. This set of components produces, given appropriate inputs, outputs (e.g. behavior, textual or graphical displays) appearing to arise from emotions because they are similar to emotional behaviors and responses observed in biological systems under equivalent circumstances. This corresponds to the "black-box" approach to emotion modeling in the classification proposed by Wehrle and Scherer (Wehrle & Scherer, 1995). As pointed out by Wehrle, "although such models provide little information concerning the mechanisms involved, they are very useful for practical decision-making and for providing a sound grounding for theoretical and empirical studies. In particular, they can help to investigate the necessary and sufficient variables. System performance (e.g. the quality of classification and computational economy), as well as cost of data gathering, are important criteria for assessing the quality of the chosen computational

model" (Wehrle, 2001, p. 565). This form of anchoring emotions can thus be very valuable to solve AI and robotics problems such as speeding up the system's responses to certain stimuli. It can also be of great help to do a systematic analysis of significant variables. However, the emotional system of the artifact does not, *per se*, shed much light on the underlying emotional mechanisms and processes, nor is it fully meaningful to the artifact itself, as it has been engineered by the designer. Only in a very restricted or metaphorical sense would I say that this form of anchorage allows the artifact to "have" emotions.

In its stronger sense, the anchorage of emotions can be seen as emotion grounding, in the same sense of the term "grounding" as it is given when talking about the *grounding problem* (Harnad, 1990). In this sense, emotions must be modeled in such a way as to be rooted in, and intertwined with, the perception and action processes of the artifact so that emotions and their consequences can have an intrinsic meaning for it. To put it in Wehrle's words, "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" (Wehrle, 2001, p. 576). Process modeling (Wehrle & Scherer, 1995), which attempts to simulate naturally occurring processes using hypothesized underlying mechanisms, is a more appropriate approach than the black-box one in this case. In this stronger sense, emotions being intrinsically meaningful to the artifact itself, I would argue that the artifact "has" emotions in a broad sense that implies the adoption of a functional view on emotion modeling (Frijda, 1995; Cañamero, 1998). This position is not devoid of methodological problems, though, as discussed in (Cañamero, 1998; Wehrle, 2001). For example, if we take inspiration from models of existing (biological) emotional systems (e.g. emotion categories, dimensions, action tendencies) to design the artifact's emotional system, one can question to what extent its emotions are really grounded. On the contrary, if we let the artifact develop its own emotions, these might not be understandable to us. As a partial way out of this dilemma, Wehrle proposes the use of appraisal dimensions borrowed from psychology as a basis for the value system of the artifact, in order to benefit from the possibility of describing the resulting emotional behavior in known terms.

In the following sections, I will sketch some elements of an alternative view to grounding emotions in embodied artifacts, in line with the emotion architecture proposed in (Cañamero, 1997). I will first discuss some elements necessary to ground emotions in individual artifacts, to consider later additional elements to ground emotions in artifacts interacting socially.

## Grounding emotions in the individual

> "Our interest in emotion in the context of AI is not an interest in questions such as "Can computers feel?" or "Can computers have emotions?" [...] our view is that the *subjective experience* of emotion is central, and we do not consider it possible for computers to experience anything until and unless they are conscious. Our suspicion is that machines are simply not the kinds of things that can be conscious." (Ortony *et al*, 1988)

Although the use of terms such as "vision" or "memory" seems to be generally well accepted when applied to artifacts, in spite of the fundamental differences between human and artificial vision or memory systems, many (artificial and human) emotion researchers remain skeptical about the use of the term "emotion" applied to artifacts and about the possibility for artifacts to "have" emotions and emotional interactions with humans and other agents. Arguments commonly heard stress the impossibility for artifacts to implement notions such as self, feelings, subjective experience, or consciousness. These phenomenological aspects of experience and emotions are most apparent to (healthy) humans, and we seem to be particularly keen on using them to define the realm of what is uniquely human.

I share with these criticisms the skepticism about the possibility for robots and other artifacts to have selves, feelings, subjective experience, consciousness, and emotions *in the same way as humans do*. Artifacts are not biological systems, they are made from a totally different sort of material, have different bodies, actuators, perceptual and cognitive capabilities, experiences, and niches, and in a sense they can be regarded as different, "new" types of species. However, I do believe that endowing artifacts with some form of (at least some of) these notions—or rather their functional counterparts—is necessary for them to interact with humans in a way that humans can understand and accept, and can likewise enhance many aspects of their behavior and performance. Rudimentary implementations of some of these concepts that ground emotions have already been proposed by practitioners of AI approach and robotics, as we will see below. We must keep in mind, however, that these are only the first attempts of a nascent endeavor, and that at this stage we can only aim at laying foundations rather than at witnessing accomplished results.

As already mentioned in the previous section, this "strong" view raises the question whether (and why) we are willing to/should use terms borrowed from models of human cognition and emotion applied to these artificial systems. Arguments can be set forth against and in favor of this practice. The use of these terms rather than newly invented ones lets *us*, humans, explain behavior and phenomena in terms we already know and understand. However, I also think that great care has to be paid to make this fundamental difference between artificial and human emotions (and the other notions as well) very clear, in particular when presenting our work to the general public, to avoid the dangers of overattribution and frustrated expectations. These dangers

might make it advisable to avoid the use of these terms in some contexts. Perhaps with time and habit one day people will talk about "robotic selves" and "emotions" as naturally as they already do about "computer vision".

Below, I will consider the particular case of an adaptive embodied autonomous artifact—a robot—but many of the arguments can be easily applied to other artifacts such as software agents or virtual characters. I will only sketch some ideas around the notion of "self", and deliberately avoid the more slippery grounds of feelings and consciousness. I think that, for these latter notions, it would be too premature to even give an opinion as to whether we will ever be able to implement (a rudimentary notion of) them in the future, given on the one hand to the lack of agreement and partial knowledge that the different disciplines have today, and on the other hand to the current state-of-the-art in AI and robotics, still at a too early stage for this. I refer the reader to (Damasio, 1999) for his excellent account of "the feeling of what happens" (i.e. the notions of feelings and consciousness grounded in the body) from a neurobiological perspective, also full of insights for the design of artificial systems.

## Embodiment

For the argument's sake, I will begin by placing myself within an embodied AI perspective (Brooks, 1991) and claim that this approach is better suited than symbolic AI to ground emotions in artifacts. The emphasis of situated AI on complete creatures in closed-loop bodily interaction with their (physical and social) environment allows for a more natural and coherent integration of emotions (at least the "non-cognitive" or perhaps the "non-conscious" aspects of them) within the global architecture and behavior of the agent. This view has important implications for emotions grounding:

- Emotion grounding requires that our model of emotions clearly establish a link between emotions, motivation, behavior, and perception, and how they feed back into each other. This link makes that emotion (as well as motivation, behavior, and perception) can affect and be affected by the other elements in a way that can be either beneficial—e.g. energize the body to allow the individual to escape faster from a predator—or noxious—e.g. cause perceptual troubles that lead to inappropriate behavior—for the agent.
- This link must be grounded in the body of the agent—for instance, by means of a synthetic physiology as in (Cañamero, 1997)—since it is through the body that agents interact with the physical and social world. I am thus implying that emotions—natural or artificial—cannot exist without a body[2].
- Since we are talking about complete autonomous artifacts, emotions must be an integral part of their architecture. 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.

A last remark before proceeding. "Embodiment" in embodied AI does not only mean that the artifact has a physical body through which it senses and acts on the world. As it is apparent to the practitioners of this field, embodiment has the fundamental implication that intelligence—cognition and, let's add, emotion—can only develop through the interactions of an embodied nervous system (or, for the matter, brain or mind) with its physical and social world.

## Robotic "selves"

Our subjective experience or the notion of "(our-)self" is the result of many different factors involving higher- and lower-level mechanisms and processes. Let us consider here two of these elements that have already received some attention in robotic research.

**Bodily self.** The experience of the own body is perhaps the most primary form of a notion of "self" and the starting point to endow an embodied artifact with some rudiments of this notion. Oliver Sacks eloquently illustrates how the perception of the own body grounds the experience of the self in his account of the case of "the disembodied lady" (Sacks, 1986). The sense of the body is given, following Sacks, by various mechanisms working together that give rise to different body models in the brain:

- Vestibular feedback, which provides us with a sense of balance.
- Exteroceptive feedback such as visual feedback that gives rise to body-image (the brain's visual model of the body), and auditory feedback.
- Proprioception: the perception of the elements of our body (limbs, facial muscles, etc.) that makes us feel our body as belonging to us.

If one of them fails, the others can compensate to some extent. Due to a sensory polyneuritis, Christina, aged 27, lost the perception of her own body (proprioception), feeling "disembodied" or with a "blind body". As a consequence, she initially lost the capacity to walk or move her limbs; her voice became flat, as vocal tone and posture are proprioceptively controlled; her face became also flat and expressionless (although her emotions remained of full and normal intensity), due to the lack of proprioceptive tone and posture; and she lost her sense of corporeal identity, leaving her with the impression that she could not 'feel'. With strong self-motivation and rehabilitative support, she developed compensatory forms of feedback—very artificial-looking at the beginning, more natural with time and practice—that allowed her to become operational again;

[2]This is not the case in programs or agents that reason about emotions.

for example, she used attention and visual feedback to move her limbs, and learned to talk and move as if she was on stage. However, she could never recover the sense of bodily identity.

The elements that inform the sense of the body—vestibular system, visual feedback and proprioception—have been implemented in one form or another in different robotic projects, like for example in the humanoid robot Cog (Brooks *et al.*, 1998). Proprioceptive feedback, with applications in robotics such as detecting self-position or controlling self-movement (including expression), has for example being used in the humanoid robot Infanoid (Kozima, 2001) in conjunction with a value system to evaluate (in a valenced way) the proprioceptive information that the robot is receiving. Although still at its earliest stages, robotics research is thus starting to implement some rudimentary elements involved in the notion of bodily self. However, how these elements work together and interact with other subsystems to give rise to the *sense* of bodily identity belongs more to the realm of feelings and subjective experience, and therefore is out of our possibilities given the current state of the art in AI and robotics, and the still very partial picture that the different disciplines can offer about these notions.

**Autobiographic self.** Another key element defining the notion of self is the 'autobiographic self', which Damasio (Damasio, 1999) presents as a necessary ingredient of the notions of identity and personality. Following Damasio, the autobiographic self is formed by the reactivation and coherent organization of selected subsets of autobiographic memories—past experiences of the individual organism. These memories are not static, but modified with experience during the individual's lifetime, and are also affected by expectations about the future. The autobiographic self and autobiographic memories are deeply related to emotions in several ways. First, some of these autobiographic memories are emotionally loaded. Second, emotions facilitate mood-congruent recall of past memories (Bower, 1981). Third, as Damasio points out, the existence of an autobiographic memory and self allows organisms to provide generally coherent emotional and intellectual responses to all sorts of situations.

The area of socially intelligent agents has for some years acknowledged the importance of autobiographic memories to found the construction of social identity and social interaction in artifacts. For example, Dautenhahn has proposed a dynamic systems approach to model autobiographic memories, and the notion of autobiographic agent as the embodied realization of an agent that dynamically reconstructs its individual history over its life-time (Dautenhahn, 1996). Nehaniv has used ideas from algebra (semigroup theory) to propose a representation of histories as autobiographies of social agents (Nehaniv, 1997).

The artificial emotion community has also made some attempts at implementing simple versions of notions relevant to the autobiographic self. Velásquez, taking inspiration from (Damasio, 1994), has for example implemented emotional memories in a learning pet robot (Velásquez, 1998), with the purposes of permitting the learning of secondary emotions as generalizations of primary ones, and of providing markers that influence the robot's decisions. As another example, Ventura and colleagues have applied Damasio's concept of the "movie-in-the-brain" (Damasio, 1999) to implement a mechanism that allows an agent to establish and learn causal relationships between its actions and the responses obtained from the environment, and to decide courses of action accordingly (Ventura *et al.*, 2001). The "movie-in-the-brain" mechanism influences decisions on courses of action as follows: the agent stores chunks of sequences of perceptions and actions, together with a measure of their corresponding desirability. When a similar situation is encountered in the future, the agent can make decisions based on its personal experience.

## Social Grounding

In addition to the notions previously mentioned (among many others), stemming from an individual's point of view, robots and other artifacts must incorporate many other mechanisms in order to behave, cognize and emote socially. Let us consider some of the elements that researchers have already started to implement in social artifacts (in particular in robots) which are fundamentally related to social emotions.

### Social motivation

Although different in nature and roles, motivation and emotion are highly intertwined and cannot be considered in isolation from each other. On the one hand, emotions can be seen as "second order" behavior control mechanisms that monitor the motivational system in achieving its goals. On the other hand, emotions modulate motivation (e.g. its intensity) and provide very strong motivation for action.

As in the case of "emotion", the term "motivation" spans a wide range of phenomena as varied as physiological drives, search for internal (psychological) or external (cultural, social) rewards, or incentives for self-regulation. The mechanisms underlying these different aspects are likely to be very different in fundamental ways, as they have to deal with factors of very diverse nature: biological, psychological, cultural, social, etc. Whereas physiological motivation (drives) are generally explained (at least partially) using a homeostatic regulation approach, models for motivation in self-regulation and social interaction and their connection to emotion are not so well established.

Different models of emotion synthesis for individual artifacts have integrated motivation as one of their elements, taking a homeostatic regulation approach to motivation, for example (Cañamero, 1997; Velásquez, 1998). Inspired from these to a large extent, and partly due to the lack of a better model, architec-

tures that have included motivation for social interactions and emotions have also taken a homeostatic regulation approach to model social motivation (Cañamero & Van de Velde, 2000; Breazeal, 2002). Motivation for social interactions is thus modeled as a set of drives such as 'social interaction', 'attachment', 'fatigue', 'stimulation level', etc. This simplified model can be very productive from a pragmatic point of view. At the architectural level, it has allowed to use the same type of mechanisms to couple motivation and emotion as those used in architectures of non-social agents—a complementary second-order monitoring mechanism, or an element of the homeostatic loop itself, depending on the approach. With respect to social interactions, it has proved capable of producing behavior that engages humans in socially rich emotional interactions and that regulates these interactions (Breazeal, 2002).

However, if we want to reach a deeper understanding of the relationships between social motivations and emotions, and a more sound and better grounded coupling of these notions in our artifacts, a finer-grained approach seems to be necessary. The relationships of social motivations and emotions to concepts like competence (in problem-solving, social, etc.), control (external and self-control), self-esteem, coherence and predictability of the environment, self-regulation, and different kinds or rewards, to name a few, need to be explored.

## Theory of mind

The term "theory of mind" is used in developmental psychology to refer to a set of metarrepresentational abilities that allow an individual to understand the behavior of others within an intentional framework—or as Tomasello puts it, to understand others as mental agents (Tomasello, 1999). Theory of mind is thus a theory of other people's minds. It relies on the ability to understand oneself as an intentional agent and to perceive others as being "like me", to use Tomasello's expression. A theory of mind thus allows us to correctly attribute "internal states"—percepts, beliefs, wishes, goals, thoughts, etc.—to others.

How would an artifact's theory of mind affect its social and emotional interactions? Scassellati draws a very eloquent picture:

> "A robotic system that possessed a theory of mind would allow for social interactions between the robot and humans that have previously not been possible. The robot would be capable of learning from an observer using normal social signals in the same way that human infants learn; no specialized training of the observer would be necessary. The robot would also be capable of expressing its internal state (emotions, desires, goals, etc.) through social interactions without relying upon an artificial vocabulary. Further, a robot that can recognize the goals and desires of others will allow for systems that can more accurately react to the emotional, attentional, and cognitive states of the observer, can learn to anticipate the reactions of the observer, and can modify its own behavior accordingly." (Scassellati, 2000)

We are still very far from achieving this full picture but, recognizing the importance of this notion of social and emotional interactions with artifacts, various robotic projects have started to implement some basic elements. Kozima, for example, is working on a mechanism for acquisition of intentionality in his humanoid robot Infanoid (Kozima, 2001), to allow the robot make use of certain methods for obtaining goals. Beginning with "innate" reflexes, Infanoid explores a range of advantageous cause-effect associations through its interactions with the environment, and gradually becomes able to use these associations spontaneously as method-goal associations. Scassellati has been working for several years on elements of a theory of mind for a robot in the frameworks of the humanoid robot Cog (Brooks *et al.*, 1998). Taking inspiration from the models proposed by Leslie (Leslie, 1994) and Baron-Cohen (Baron-Cohen, 1995), he has started to specify the perceptual and cognitive abilities that a robot with a theory of mind should employ, focusing initially on the implementation of two abilities: distinguishing between animate and inanimate motion, and identifying gaze direction for shared attention (Scassellati, 2000).

## Sympathy and empathy

A set of related mechanisms are relevant to the capacity that humans and other social animals have to "connect" with the emotional state of others and to "adopt" it to varying degrees: phenomena like emotional contagion (related to imitation), sympathy, empathy, perspective taking, and prosocial behaviors like helping. The literature has long debated the differences among these phenomena and whether they belong more to the "cognitive" or to the "emotional" realm, in particular in the case of empathy. Recently, a framework has been proposed (Preston & de Waal, 2001) that explains all these phenomena within a "perception-action model" applicable to both the cognitive and the emotional domains, and sees empathy as a superordinate category. The cognitive and emotional mechanisms involved in these phenomena vary. Whereas in emotional contagion the observer's state results "automatically" from the perception of the other's state, in sympathy—"feeling with"—the observer "feels sorry" for the other, focusing more on his situation than on his physical state; attention is fundamental in empathy—"feeling into"—, where the observer's state results from the attended perception of the other's state and, in the framework proposed in (Preston & de Waal, 2001), arises from a projection on the part of the observer rather than from a perception of the other's state. Besides attention, effects of familiarity/similarity, past experience, learning and cue salience are all fundamentally involved in empathy.

Again, researchers in the field of social artifacts have acknowledged the importance of these phenomena for social understanding and social/emotional interactions (in humans, animals, and artifacts), in particular of empathy as a form of emotional communication that favors the perception of social signals, as discussed for example in (Dautenhahn, 1997). Some elements involved in these phenomena are being investigated by different researchers, such as agent architectures to achieve a "sympathetic coupling" (Numaoka, 1997), or underlying perceptual mechanisms such as gaze direction for joint attention or cue salience (Scassellati, 2000; Breazeal, 2002), but we are a long way from artifacts capable of showing empathy. For the moment, all that our emotional and expressive artifacts can achieve is to produce (something resembling to) "empathic" reactions from humans, as for example reported in (Scheeff *et al.*, 2001; Cañamero, 2001; Breazeal, 2002).

## Conclusion

Starting from the idea that the social and the emotional are highly intertwined, this paper has discussed issues around the construction of emotional artifacts that have to interact in a social world, and in particular with humans. I have first examined some of the ways in which emotions can enhance social interactions with artifacts, and some of the challenges that they pose to the designer. After considering the debate that opposes "shallow" modeling of observable emotional features versus "deep" modeling that includes an underlying emotional system, I have sketched some ways in which we can anchor emotions in the architecture of artifacts in order to make emotional interactions meaningful not only to the human, but also to the artifact itself. I have finally discussed some of the cognitive capabilities that artifacts should incorporate for their emotions to be properly grounded and to give rise to rich social exchanges with humans.

A final comment concerning approaches to design. Most of the projects mentioned take a developmental path to the design and construction of artifacts with social and emotional capabilities. Ontogenetic development is no doubt necessary for an (embodied) artifact to build up its cognitive and emotional abilities through its interactions with the physical and social environment. Again, human development is what we know better (or closer) and it is natural to take it as a model to inspire the design of an artifact that can develop and learn from experience. Designers are facing an enormous challenge here, though, given the complexity of human beings and their developmental process. Starting the artifact from scratch is impossible, and the designer has to decide at what level (or from what "primitives") and according to what theory(-ies) s/he will start to work. Each of the elements discussed in previous sections relies on many other (physical, perceptual and cognitive) capabilities, the implementation of which is equally challenging and not always possible. We thus seem confronted with a dilemma. On the one hand, a developmental approach is necessary both to ground emotions in artifacts and to gain a better understanding of how emotions develop and interact with other aspects of our cognition and sociality. On the other hand, trying to follow too closely the steps proposed for human development by psychological theories, besides being possible only in a very limited way, can introduce many biases in our model and lead to deadlocks. Although devoid of the richness of these models, complementing this endeavor with an investigation of the development of, and interactions between emotions, cognition, and sociality using (much) simpler models (more on the bacteria side than on the human side, so to say) could also provide very valuable insights for the understanding of emotions in social species.

## Acknowledgments

Some of the ideas presented in this paper benefited from discussions with colleagues in the working group on "Emotions in Social Interactions" during the AAAI 2000 Fall Symposium Socially Intelligent Agents: I thank the contributions of Alan Bond, Cristina Conati, Nadja De Carolis, Jakob Fredslund, Eva Hudlicka, Christine Lisetti, and Valery Petrushin. I am also grateful for the feedback provided by recent discussions with colleagues of the "human emotion" community: Nico Frijda, George Mandler, Keith Oatley, Andrew Ortony, Jaak Panksepp, Klaus Scherer, and Robert Zajonc.

## References

Baron-Cohen, S. 1995. *Mindblindness*. Cambridge, MA: MIT Press.

Bower, G.H. 1981. Mood and Memory. *American Psychologist*, 36: 129–148.

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

Brooks, R.A. 1991. Intelligence without Representation, *Artificial Intelligence*, 47: 139–159.

Brooks, R.A., Ferrel (Breazeal), C., Irie, R., Kemp, C., Marjanovic, M., Scassellati, B., and Williamson, M. 1998. Alternative Essences of Intelligence. In *Proc. 15th National Conference on Artificial Intelligence (AAAI–98)*, pp. 961–967. Menlo Park, CA: AAAI Press.

Cañamero, L.D. 1997. Modeling Motivations and Emotions as a Basis for Intelligent Behavior. In W. Lewis Johnson, ed., *Proceedings of the First International Conference on Autonomous Agents*, pp. 148–155. New York, NY: ACM Press.

Cañamero, L.D. 1998. Issues in the Design of Emotional Agents. In *Emotional and Intelligent: The Tangled Knot of Cognition. Papers from the 1998 AAAI Fall Symposium*. Technical Report FS–98–03, pp. 49–54. Menlo Park, CA: AAAI Press.

Cañamero, L.D. and Van de Velde, W. 2000. Emotionally Grounded Social Interaction. In K. Dautenhahn, ed., *Human Cognition and Social Agent Tech-*

*nology*, pp. 137–162. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Cañamero, L.D. 2001. Playing the Emotion Game with Feelix: What Can a LEGO Robot Tell Us about Emotion?. In K. Dautenhahn, A. Bond, L. Cañamero, and B. Edmonds, eds., *Socially Intelligent Agents: Creating Relationships with Computers and Robots*. Kluwer Academic Press (*to appear 2001*).

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

Damasio, A. 1994. *Descartes' Error. Emotion, Reason, and the Human Brain*. New York: Putnam's Sons.

Damasio, A. 1999. *The Feeling of What Happens. Body and Emotion in the Making of Consciousness*. New York: Harcourt.

Dautenhahn, K. 1996. Embodied Cognition in Animals and Artifacts. In *Embodied Cognition and Action. Papers from the 1996 AAAI Fall Symposium*. Technical Report FS–96–02, pp. 27–32. Menlo Park, CA: AAAI Press.

Dautenhahn, K. 1997. I Could Be You—The Phenomenological Dimension of Social Understanding. *Cybernetics and Systems*, 28(5): 417–453.

Dautenhahn, K. 1998. The Art of Designing Socially Intelligent Agents: Science, Fiction, and the Human in the Loop, *Applied Artificial Intelligence*, 12(7/8): 573–617.

Dumouchel, P. 1999. *Emotions : essai sur le corps et le social*. Le Plessis-Robinson, France: Institut Synthélabo/PUF.

Ekman, P. 1992. An Argument for Basic Emotions. *Cognition and Emotion* 6(3/4): 169–200.

Frijda, N.H. 1995. Emotions in Robots. In H.L. Roitblat and J.-A. Meyer, eds., *Comparative Approaches to Cognitive Science*, pp. 502–516. Cambridge, MA: MIT Press.

Harnad, S. 1990. The Symbol Grounding Problem, *Physica D*, 42: 335–346.

Kozima, H. 2001. Infanoid: A Babybot that Explores the Social Environment. In K. Dautenhahn, A. Bond, L. Cañamero, and B. Edmonds, eds., *Socially Intelligent Agents: Creating Relationships with Computers and Robots*. Kluwer Academic Press (*to appear 2001*).

Leslie, A.M. 1994. ToMM, ToBY, and Agency: Core Architecture and Domain Specificity. In L.A. Hirschgeld and S.A. Gelman, eds., *Mapping the Mind: Domain Specificity in Cognition and Culture*, pp. 119–148. Cambridge University Press.

Nehaniv, C.L. 1997. What's Your Story? — Irreversibility, Algebra, Autobiographic Agents. In *Socially Intelligent Agents. Papers from the 1997 AAAI Fall Symposium*. Technical Report FS–97–02, pp. 150–153. Menlo Park, CA: AAAI Press.

Numaoka, C. 1997. Innate Sociability: Sympathetic Coupling. In *Socially Intelligent Agents. Papers from the 1997 AAAI Fall Symposium*. Technical Report FS–97–02, pp. 98–102. Menlo Park, CA: AAAI Press.

Ortony, A., Clore, G.L. and Collins, A. 1988. *The Cognitive Structure of Emotions*. New York: Cambridge University Press.

Preston, S.D. and de Waal, F.B.M. 2001. Empathy: Its Ultimate and Proximate Bases. Accepted for publication in *Behavioral and Brain Sciences*.

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

Russell, J.A. 1997. Reading emotions from and into faces: Resurrecting a dimensional-contextual perspective. In J.A. Russell, J.M. Fernández-Dols, eds., *The Psychology of Facial Expression*, pp. 295–320. Cambridge University Press.

Sacks, O. 1986. *The Man Who Mistook His Wife for a Hat*. London, UK: Picador.

Scassellati, B. 2000. Theory of Mind. . . for a Robot. In *Socially Intelligent Agents: The Human in the Loop. Papers from the 2000 AAAI Fall Symposium*. Technical Report FS–00–04, pp. 164–168. Menlo Park, CA: AAAI Press.

M. Scheeff, J. Pinto, K. Rahardja, S. Snibbe and R. Tow. Experiences with Sparky, a Social Robot. In K. Dautenhahn, A. Bond, L. Cañamero, and B. Edmonds, eds., *Socially Intelligent Agents: Creating Relationships with Computers and Robots*. Kluwer Academic Press (*to appear 2001*).

Smith, C.A. and Scott, H.S. 1997. A Componential Approach to the meaning of facial expressions. In J.A. Russell, J.M. Fernández-Dols, eds., *The Psychology of Facial Expression*, pp. 229–254. Cambridge University Press.

Tomasello, M. 1999. *The Cultural Origins of Social Cognition*. Cambridge, MA: Harvard University Press.

Velásquez, J.D. 1998. Modeling Emotion-Based Decision-Making. In *Emotional and Intelligent: The Tangled Knot of Cognition. Papers from the 1998 AAAI Fall Symposium*. Technical Report FS–98–03, pp. 164–169. Menlo Park, CA: AAAI Press.

Ventura, R., Custódio, L., and Pinto-Ferreira, C. 2001. Learning Course of Action using the "Movie-in-the-Brain" Paradigm. *This volume*.

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

Wehrle, T. and Scherer, K. 1995. Potential Pitfalls in Computational Modeling of Appraisal Processes: A Reply to Chwelos and Oatley, *Cognition and Emotion*, 9: 599–616.