Embodied Robot Models for Interdisciplinary Emotion Research

Lola Cañamero

Abstract—Due to their complex nature, emotions cannot be properly understood from the perspective of a single discipline. In this paper, I discuss how the use of robots as models is beneficial for interdisciplinary emotion research. Addressing this issue through the lens of my own research, I focus on a critical analysis of embodied robots models of different aspects of emotion, relate them to theories in psychology and neuroscience, and provide representative examples. I discuss concrete ways in which embodied robot models can be used to carry out interdisciplinary emotion research, assessing their contributions: as hypothetical models, and as operational models of specific emotional phenomena, of general emotion principles, and of specific emotion "dimensions". I conclude by discussing the advantages of using embodied robot models over other models.

Index Terms—Emotion models, robot models of emotions, embodied emotions, affective cognition, affective interaction, autonomous robots, human-robot interaction, embodied artificial intelligence, interdisciplinary research.

1 Introduction

> "This suggestion [...] would require tight collaborations between neuroscientists and roboticists in order to be refined, but because of the very peculiar characteristics of emotions, such an endeavor could lead to important advances in robot [...] designs. These advances in turn could lead to new insights on the functions of emotions and would suggest new avenues for research on their neural bases." (Jean-Marc Fellous, [35])

In this paper, I elaborate on the use of robots as models in emotion research—where such models can originate in different disciplines, have different levels of abstraction, be about different aspects—and their value to foster interdisciplinarity. Emotions provide an ideal framework for inter- and cross-disciplinary research since, due to their complex multi-faceted nature, they cannot be properly understood from the perspective of a single discipline. As I argued in [20], such interdisciplinarity is key in order to understand shared conceptual problems—such as mechanisms underlying the involvement of emotions in cognition and action, emotion elicitors, how emotions function as "cognitive modes", or the relation among emotions, value systems, motivation, and action—and shared challenges and goals for future research—what I called the "origins" and grounding problem of artificial emotions, dissolving the "mind-body" problem, untangling the "knot of cognition" (the links between emotion and intelligence), and measuring progress and assessing the contributions of emotions to our systems. What can robot models contribute to investigate such (and other) shared shared conceptual problems, and to address shared challenges and research goals? Far from attempting to provide a survey of robot and computational models of emotions (for surveys, see e.g., [64], [69], [82], [84], [85]), I will, on the contrary, address the use of robot models in interdisciplinary emotion research through the lens of my own research, focusing on embodied robot models, in a critical reflection on the interplay between my interest in interdisciplinary research and my actual robot models. However, some (but not all) of the discussions in this paper will also apply to other approaches to emotion modeling such as computer simulations or embodied virtual humans.

One of the key contributions that robot (and computational) models can make to emotion research is the possibility to implement, test, extract and analyze assumptions and consequences, and assess the scope and "usefulness" of different conceptualizations, models and theories of emotions. What can different approaches such as discrete "basic" emotions, emotion "dimensions", dynamical systems, etc, tell us about emotions? How do they "behave" when implemented in robots situated in, and in interaction with their environments? Such—at the same time scientific and philosophical—preoccupations have been at the basis of my own research on modeling emotions in autonomous robots for over two decades. Although in all cases from an embodied, situated and interactional perspective, at different points in time I have implemented (and combined) different emotion models, trying to understand what we, as scientists, and the robots, as embodied models, could (and could not) do with them. Here I provide a selection of models that, while not exhaustive, cover a wide range of approaches. Whereas I group them under different labels, I do not attempt to provide a classification of robot models of emotions (classifications can be found in the above-mentioned surveys), but to illustrate and discuss a number of ways in which robot models can be used to carry out interdisciplinary emotion research: as hypothetical models (Section 2), as models of specific emotional phenomena (Section 3), as models of general emotion principles (Section 4), and as models of specific emotion "dimensions" (Section 5). I conclude by discussing some of the features that make embodied robot models of emotion particularly appropriate to study emotions bridging interdisciplinary gaps.

Lola Cañamero is Head of the Embodied Emotion, Cognition and (Inter-)Action Lab, Dept. Computer Science, University of Hertfordshire, Hatfield, Herts, AL10 9AB, UK. L.Canamero@herts.ac.uk; www.emotion-modeling.info

2 Robots as Hypothetical Models

A sound theoretical speculation by neuroscientists or psychologists regarding whether a robot could have emotions [1], or how to build one that could have them [76], can be very valuable to roboticists as a guide to develop biologically-meaningful robots. I will argue that it can also be very useful to help neuroscientists understand not only the big, unresolved philosophical questions regarding the nature of emotions and cognition from a different perspective, but also more concrete aspects of their models and the criteria underlying them. Such models can provide theoretical desiderata and help identify and define key properties of the aspects of emotions being modeled. As neuroscientist Ralph Adolphs puts it [1, page 9]:

> "Could a robot have emotions? [...] we should attribute emotions and feelings to a system only if it *satisfies criteria* in addition to mere behavioral duplication. Those *criteria require in turn a theory of what emotions and feelings are* [...] I conclude with the speculation that robots could certainly *interact socially* with humans within a restricted domain (they already do), but that correctly attributing emotions and feelings to them would require that robots are *situated in the world* and *constituted internally* in respects that are *relevantly similar* to humans. In particular, if robotics is to be a science that can actually tell us something new about what emotions are, we need to engineer *an internal processing architecture that goes beyond merely fooling humans into judging that the robot has emotions*."

Adolphs raises many important issues—which I have highlighted in italics—regarding the characterization of emotions, their embodied and situated nature, and what "robot emotions" need to address in order to make relevant contributions to our understanding of emotions, and be part of the affective sciences. We will see these issues reappear in the other types of models—both theoretical and robot models—discussed in the remainder of the paper. They are important centuries-old philosophical (and scientific) questions and a potentially powerful tool to make emotion theorists critically think about what the key concepts of their theory could *really* mean—what they entail, which alternative types of mechanisms could underly them, what they *could have been like* given different evolutionary / developmental / socio-cultural histories, etc. However, while using sound criteria grounded in a theory is very important to ascertain what emotions can be, we should be careful to avoid a circular argument, defining emotions *a priori* as X, Y and Z, then calling something that shows X,Y and Z an emotion, to the exclusion of things that do not show X, Y and Z. A way to avoid such dangers would be to use, not only criteria stemming from the theory, but also from an interdisciplinary "meta-analysis" [31], [61].

3 Models of Specific Emotional Phenomena

Roboticists are interested in modeling specific emotional phenomena to elaborate biologically-inspired mechanisms that can be beneficial to improve the behavior of robots in ways that are similar to biological systems. Robots have also interested neuroscientists as potential testbeds for specific emotional phenomena. An example is offered by Joseph LeDoux, whose studies of fear processing in the amygdala have been very influential in the robotics community. The "processing approach" that he advocates [36] permitted him to make significant advances in the neuroscience of emotion by abandoning the old idea of the limbic system that attempted to explain all emotions. Instead, he focused on trying to unveil the mechanisms and circuits underpinning a single emotion—fear—and its involvement in different aspects of cognition, going beyond human emotions [36, page 82]: "The processing approach allows us to study unconscious emotional functions similarly in humans and other animals." And even in robots, as he draws on the experience gained through his research in neuroscience to offer advice for computational modeling [36, page 105]: "it might be fruitful for computational models to approach the problem of emotions by considering one emotion at a time and to focus on how the emotion is operationalized without losing the "big picture" of how feelings might emerge. This approach has led to the discovery of basic principles that may apply to other emotions as well as fear." Some of these principles are:

- P1 Emotions involve primitive circuits conserved across evolution.
- P2 In some circumstances, cognitive circuits can function independently from emotions.
- P3 There are two parallel routes of emotional processing of a stimulus, one fast, the other slower and modulates the fast route.

3.1 Robot Model Examples: Fear and the Amygdala

Several computational and robot models have been inspired by such (and similar) principles. Examples include models of fear conditioning [3], [67], [73], emotional learning [5], reinforcement learning [95], or second-order conditioning [68]. Such models present clear advantages from the robotics perspective, as they add functionalities to robots that are useful for their survival and interaction in dynamic, unpredictable, social environments inhabited by humans. What are the potential contributions of these models from the perspective of neuroscience?

3.2 Discussion

While the fact that such models show that the (same) principles extracted from the analysis of living systems can be used for the purposes of synthesizing behavior in artificial systems is already a very interesting result, a fuller use of the robot models would target at "closing the loop" by making some unique contribution arising from the (interdisciplinary) use of robots. One such contribution would be linked to the fact that robot models permit to test different (e.g., alternative) hypotheses and compare the behavior generated by them. Let us take an example from the above principles—P3, the dual route of emotional processing of stimuli. This hypothesis [59] has been influential in computational and robot modeling of emotion—e.g., [67]. Although evidence is still found in favor of the dual route in terms of separate brain "circuits" [45], this hypothesis

is currently subject to debate, as alternative ones attribute the difference between "fast" and "slow" processing to differential processing of different ("low" or "high") spatial frequency information [94], or postulate "many-routes" [80], drawing a more complex picture of emotional processing that emphasizes the role of the cortex and its ability for fast processing. Robot models would permit particularly interesting and systematic comparisons of these hypotheses to try to understand their relevance, plausibility, and perhaps complementarity, when used to drive the behavior of a robot under different environmental circumstances, tasks and challenges.

An important issue that these models raise for interdisciplinary research is the correspondence between biological notions and computational constructs: which properties or features allow us to call a computational construct "amygdala", "orbito-frontal cortex", etc? Are the modeled properties sufficient to justify the use of the biological term? How useful can such practice be? How misleading?

4 Models of General Emotion Principles

Robots can also embed more abstract models of more general functional properties and principles of emotion and their interaction dynamics—modeling e.g., the roles of such properties and principles in emotional regulation of agent-environment interactions, or in different aspects of emotion-cognition (including motivation and behavior) interactions. This type of "functional modeling" using robots is not the exclusive realm of roboticists and has also attracted interest from psychologists and neuroscientists. Here I will consider one example from each field and illustrate the counterpart robot models with examples of my own work.

4.1 Regulation of Agent-Environment Interactions

Psychologist Nico Frijda adopts a functionalist view [40] that considers action, motivation for action and action control as the main role of emotion, and specific emotions as mechanisms to modify or maintain the relationships between an agent and its environment in different ways, e.g.: blocking influences from environment (anger); protecting the agent against these influences (fear); stopping or delaying an active relation when the agent is not prepared for it (sadness); diminishing risks of dealing with an unknown and potentially noxious environment (anxiety).

In [40], Frijda proposed guidelines to implement a functional model of emotions in a robot. From a functional point of view, we need to identify and model the properties of the structure of humans (e.g., autonomy, having limited energetic and processing resources, having multiple concerns, or the use of signals for interaction with the environment) and their environment (e.g., limited resources, uncertainty, partly social) that are relevant to the study of emotions, and that are shared by a structurally different "species". This would permit to build robots that "are situated in the world and constituted internally in respects that are relevantly similar to humans," borrowing Ralph Adolphs' words (cf. Section 2), and can therefore be relevant models of emotions. From such properties of humans and their environments, Frijda posits the following *functions of emotions*:

- F1 To signal the relevance of events for the concerns of the system.
- F2 To detect difficulties in solving problems posed by these events in terms of assuring the satisfaction of concerns.
- F3 To provide goals for plans for solving these problems in case of difficulties, resetting goal priorities and reallocating resources.
- F4 To do this in parallel for all concerns.

Such functions are valid across structurally different embodiments, architectures and environments. A system possessing mechanisms that fulfill (some of) these roles can thus be said to (partially) have emotions from a functional point of view. However, functional models remain underspecified regarding the underlying design and implementation mechanisms. Such underspecification raises many conceptual and design issues [18], [19], [20] that provide wonderful challenges and opportunities for theoretical and empirical exploration and interdisciplinary collaboration.

4.1.1 Example: Basic Emotions with Specific Functions

Despite the fact that the initial computational model designed by Frijda and his collaborators [42], [43] was developed within an appraisal and a classical artificial intelligence (AI) perspective, the fact that the model is underspecified has permitted to use the same principles in very different computational and robot models. For example, I used them in my early work [17] to model discrete basic emotions fulfilling the above functions in autonomous agents designed from the opposite perspective of embodied AI [14] and a "decentralized" view of intelligence [71]. This model also integrated elements from ethological and neuroscientific models of behavior, motivation and emotion, particularly [29], [53], [59], [77], to implement their underlying mechanisms and interaction dynamics. In this model, autonomous agents ("animats" or simulated robots) inhabiting a complex and dynamic two-dimensional action selection environment had to constantly make decisions in real time to interact and deal with the static and dynamic elements of the environment in order to survive. The environment (called "Gridland"), was populated by other moving agents (both friendly conspecifics and unfriendly agents that could attack, damage and "kill" the other agents) and containing (dynamic) survival-related consumable resources, obstacles obstructing availability of those resources, and other objects of various shapes. The complex action selection architecture of the agents included a (simulated) physiology—in effect a complex dynamical system—of homeostatically-controlled survival-related variables and modulatory "hormones", both generic (with global excitatory and inhibitory effects) and specific (acting differentially on specific receptors) that grounded the embodiment of the agents and other elements of the architecture such as motivations, consummatory and appetitive behaviors, and a number of basic emotions. In this system, emotions fulfilled the general functions F1–F4 in Frijda's above list. In turn, through its triggering conditions and its effects, each emotion was designed to have a specific function, in line with basic emotions theory and Frijda's view of basic emotions as mechanisms to control agent-environment interactions described above. Designed in such

way, each basic emotion would normally contribute to the good maintenance of homeostasis and to the agent's behavior selection; however, it could also have noxious effects when its intensity was too high or displayed in the wrong context—e.g., excessive anger could make the agent bump into things and harm itself, or modify negatively specific parameters of the ART neural network carrying out object recognition, creating a "confused" state that could make the robot interact with the "wrong" object.

4.1.2 Discussion

This kind of effort to synthesize and operationalize, in an artificial agent, elements from various conceptual models, offers opportunities for exploring the complementarities of neuroscientific and psychological models—basic emotions and an embodied approach that implements such emotions in terms of a physiology-based dynamical system in the Gridland example. It also allows us to manipulate numerous parameters to make concrete predictions regarding, for example, the adaptive value of various emotions in different situations and contexts, their effects on survival, motivation, behavior, perception, memory, decision making, etc., of agents interacting with their physical or social environment.

There are, however, questions that such models cannot answer due to the selection and explicit design of specific emotional subsystems with pre-defined functions. For example, this model cannot explain how some traits of emotions might emerge from, or be the side-effect of, other processes; or what could be the minimal set of mechanisms that would generate behavior that could be qualified (according to some criteria) as "emotion-like". Due to its complexity, this model also makes it difficult to understand the behavior of the system as a function of specific mechanisms or their interactions. A model of emotions as "emergent phenomena" would be more appropriate to study such questions.

4.2 Emotion Dynamics and Emergent Functionality

A very different type of functional approach is proposed by neuroscientist Jean-Marc Fellous, according to whom, understanding what emotions are involves understanding what they are for. In his view [35], "one of the main functions of emotions is to achieve the multi-level communication of simplified but high impact information." Critizicing models that posit specialized emotion brain centers, Fellous advocates a view of emotions as dynamical patterns of neuromodulations rather than patterns of neural activity [34]. Neuromodulation provides a useful framework to understand how the main function of emotions is achieved and, hence, how emotions arise, are maintained, and interact with other aspects of behavior and cognition, as well as other hard problems in emotion research. It also provides a common framework to study emotions across species—biological and artificial—from a functional stance, since the specific way in which the function of emotions is achieved depends on the specific details of the species and the emotion at hand. Emotions as patterns of neuromodulation affect the underlying neural circuitry in different ways and to different degrees depending on the complexity and properties of the circuitry. This theory has some important consequences regarding the conceptualization of emotions [35], for example:

- C1 Emotion is not the product of neural computations per se. The fact that some structures are more involved in emotions than others results from both, the fact that they are more susceptible to neuromodulation, and their anatomical position.
- C2 The coupling between an emotional (attractor) state and a cognitive process (e.g., a specific memory) does not presuppose that either has a predominant or causal role. Emotion and cognition are integrated and interdependent systems implemented by the same brain structures rather than two different sets of interacting brain structures. Emotion is related to the state of neuromodulation of these structures (pattern of activation of some neuromodulatory receptors), while cognition is related to the state of information processing (neural activity).
- C3 Neuromodulation occurs on a large range of time scales, while neural activity is restricted to the millisecond time scales. This difference accounts, at least in part, for the fact that emotions may significantly outlast their eliciting conditions.

Based on this theory and its consequences for understanding biological emotions, Fellous offers explicit guidelines and constraints to model emotions in robots [35]:

- G1 Emotions should not be implemented as separate, specialized modules computing an emotional value on some dimension. Such implementations cannot handle some of the key aspects of emotions such as the complexity of the emotional repertoire, its wide time scales, and its interactions with cognition.
- G2 Emotions should not simply be the result of cognitive evaluations, as not all emotions are generated cognitively, and such view does not capture the main function of emotions—efficient multi-level transmission of simplified but high-impact information.
- G3 Emotions should have their own temporal dynamics and should interact with one another. Implementing emotions as "states" does not capture those temporal characteristics, which are key aspects of emotions and may have functional consequences.

4.2.1 Robot Model Examples: Emergent Affect-Like Functionality through Hormonal Modulation

Models of neuromodulation in robots span different aspects, such as increasing the plasticity and flexibility of the robot controller [28], [90], improving its evolvability [33], [78], improving its adaptation and survival by affecting its action selection [38], [56], producing different emotion-related behaviors (different functionalities) from the same underlying controller [4], [22] or affecting its emotional and cognitive development [11], [65], [66]. Let us have a look at some of the models developed in my group that implement emotions as "emergent functionality" [91], relating them to the consequences and guidelines of the neuroscience model proposed by Fellous. These simple models do not intend to replicate the biological mechanisms in detail, but rather capture and study in a more abstract way the dynamics of (selected examples of) emotional modulation of the underlying "nervous system" of the robot. They hence model the dynamics of some aspects of "affect-cognition" interactions,

or rather of affective cognition, as I prefer to put it. I refer to the mechanism implementing emotional modulation more generally as "hormonal modulation" rather than strictly "neuromodulation". This choice reflects, on the one hand, the fact that the distinction between hormones, neurohormones and neuromodulators is not as clear-cut as traditionally thought and, on the other hand, the fact that, in line with [24], [25], I consider emotions as *embodied*, i.e., involving different aspects the whole body in interaction with the physical and social environment, rather than simply "sited" in the brain or the nervous system.

(a) Modulation of Exteroception

The study reported in [4], using Lego robots and a simplified version of the architecture and environment in [17], investigated the adaptive value of affect—specifically of motivation modulated by emotion—in decision making in a "two-resource problem" (TRP), considered to be the simplest action selection problem. In this scenario, an autonomous robot needs to choose between two resources that it needs to consume in order to survive, and do so in a timely manner. Our environment consisted of a walled arena with two types of static resources available—light and dark patterns located on the floor and that the robot could perceive when placed on top of them using infrared sensors. Our robot was endowed with two internal survival-related variables controlled homeostatically, giving rise to drives that motivated it to look for, and consume, the appropriate resource to satisfy the most urgent need. The most urgent need changed dynamically as a function of the robot's interaction with and perception of the environment.

While a solitary robot could easily survive in the initial environment containing static and easily-available resources, the inclusion of a second identical robot trying to solve its own TRP in the same environment made the problem harder for each robot, turning it into a "competitive" TRP (CTRP). This was due to the fact that, since robots had to be located on top of the resources to be able to consume them, they limited resource availability for the other robot when they were consuming or blocking access to them. The robots could not communicate with each other and were not even "aware" that another robot was present[1]. In such CTRP environment, robots endowed with the same internal architecture as in the TRP "died" easily, and the main causes that we identified were goal obstruction—access to resources blocked by the other robot—and overopportunism—excessive opportunistic consumption of the less needed resource to the detriment of the most pressing need, which could not be satisfied in time. Such problems arose because the robots did not have the capability to interact with the newly added dynamic element—each other—in an appropriate way—in this case, by having some sort of competition skills such as "flee-fight" behavior. Instead of endowing the robots with such competition skills explicitly (i.e., adding explicit competition behaviors), we opted for modulating the same architecture used in the TRP in order to give rise to different functionality adapted to the current situation. We also hypothesized that the problem might be regarded as an "attentional" one: the robot was not paying attention to what it needed to be paying attention to. Therefore, instead of directly modulating the behavior, we modulated the perception of the robot in order to change the overall behavior of the same "perception-action loop."

The addition of a simple synthetic hormone modulating perception—a parameter influencing the *incentive salience* [10] of the external stimuli—in the TRP architecture produced "flee-fight" behavior, providing some basic competition skills that allowed the robot to overcome the above-mentioned problems of the CTRP, greatly improved its adaptation and survival. The hormone was released as a function of the internal state of the robot[2], as follows. When in high need, if the robot's "risk of death" was high, the hormone affected the perception of the non-relevant stimulus (perceived through the infrared sensor) by diminishing its salience; this often caused the robot not to stop over the "wrong resource". When in high need, if the "risk of death" was low, tactile perception was affected—the readings of the bumper sensors were not taken into account, causing the robot *not* to reverse when bumping into something. The effects of the hormone on perception through the infrared sensor largely corrected the problem of overopportunism. Its effect on the contact sensor produced behavior akin to a "flee-fight" system, which addressed[3] the "goal obstruction" problem, since when a robot bumped into another robot that was blocking the targeted resource, if would "push it" out if it had enough "stamina" (low risk of death), due to the cancelation of the contact sensor, and would move away and "abandon the goal" if it didn't.

(b) Modulation of Interoception

Following an incremental approach, in [22] we increased the complexity of the problem by changing the relation between the two robots into an asymmetric "prey-predator" one. The prey robot was now fitted with an additional infrared sensor to detect the predator robot (which was fitted with a infrared-emitting device), and with an additional touch sensor that, when activated, caused "internal damage" measured by a new homeostatic variable and that made the robot move in a different direction. The robot could "heal" from internal damage inside a "nest" located in one of the corners. The architecture of the "predator robot" was identical to that used in the CTRP. Both robots had the same motor capabilities (e.g., other things being equal, they moved at the same speed). In this "Competitive 3-Resource Problem" (C3RP), the prey robot was thus confronted with three needs for which three resources were available, and with an active threat. The architecture that

1. The robots could only perceive each other as "obstacles" through their contact sensors when bumping into each other, exactly in the same way they perceived the wall of the arena or any other 3D object.

2. A combination of the intensity of the need and the risk of death from failure to keep the homeostatic variables within their viability limits, as captured by the quantitative metrics described in [4].

3. Note that the "pushing" behavior had a beneficial adaptive value only when it was directed at the other robot in cases of "goal obstruction"; however, due to the limited perceptual capabilities of the robot, such behavior could also happen at other points, either against the other robot or against the wall or any other obstacle encountered. In such cases, the behavior had no beneficial value, and could have either no effect or a negative one—e.g., if the robot was locked in a corner and kept bumping into the wall until it "died", unable to satisfy its need—and hence be dysfunctional, similarly to misdirected aggressive behavior or "anger" in biological systems.

successfully solved the CTRP was unsuccessful in this new C3RP environment, primarily due to the fact that, when the prey was attacked, it rarely managed to "get rid" of the predator (due to their similar speeds) in order to either reach the nest and heal, or attend to the other survival needs. Increasing the "escape" speed of the prey was not a satisfactory solution, since the prey would still be frequently damaged and interrupted when attending to its other needs.

We again treated this problem on the perceptual side of the loop—interoceptive perception, in this case: the problem could be solved if perception had an effect on the robot's behavior *before* damage was caused. We explored the temporal properties of hormonal modulation and introduced another hormone that affected the perception of the level of damage, and had a slower decay. Following [74], a "gland" regulated release and amount of hormone present as a function of sensory inputs (distal perception of the predator via the infrared) and hormone decay. The perception of the level of damage, instead of being the "raw value", was mediated by the amount of hormone in the system. The effects and temporal dynamics of this hormone made the prey robot to start perceiving some level of internal damage (increasing as a function of the proximity of the prey and of the frequency and recency of previous attacks), and therefore start avoiding the predator, before the predator was too close, thus surviving more easily. This system can be regarded as a simple model of phenomena underlying the anticipatory role of "anxiety".

4.2.2 Discussion

These robot models meet many of the conceptual consequences and guidelines provided by Fellous' model, and notably the above-mentioned C1–C3, G1–G3. Due to their simplicity and transparency, this type of models provide an excellent framework to test neuroscientific hypotheses and related predictions incrementally in a highly controlled way. They foster identification and isolation of a small number of key variables to be studied and high control of the parameters used. This permits us to focus on detailed investigations of the *interactions* among different elements of the system—where system = robot + (physical and social) environment—and carry out a detailed quantitative and qualitative analysis of the system from "inside" and "outside" at different "levels" (e.g., "physiological", behavioral, interactional), as well as analyzing the interactions among such levels using methods and metrics from different disciplines. Finally, the use of autonomous robots interacting among themselves can be very valuable to single out key aspects of the interaction that cannot be isolated when one of the interaction partners is a human.

5 Models of Specific Emotion "Dimensions"

Modeling emotions in terms of dimensions [70] allows us to generate a very rich palette of affective states and expressions [8], [13], [69], well suited for subtle, varied and prolonged interactions with humans. However, this complexity can also pose problems if our goal is to understand the mechanisms and processes underlying emotions, since each of those dimensions are abstract "umbrella terms" that group—and blur differences among—a number of different mechanisms and phenomena. To overcome this problem, a possibility offered by robot models but generally not available with humans, are *single-dimension models*. Such models allows us to investigate questions such as how much (and which aspects of) emotion can be conveyed and perceived using arousal only or valence only.

Below, I discuss single-dimension embodied robot models based on specific roles of pleasure and arousal. These two dimensions are thought to provide different types of "information" about the world and our interactions with it: pleasure/valence is generally linked to information about value, whereas arousal would provide information about urgency or importance [92].

5.1 Pleasure-Based Robot Models

Like in biological systems, for autonomous robots that need to survive and interact in their environments, pleasure (or its functional robot equivalent) can provide signals to assess the positive or negative quality of the perceived stimuli, the behavior being executed, or the interaction with others. In my group, we have developed robot models of different roles of pleasure, where "pleasure" was used, for example, as a signal to learn object affordances in the context of decision making [26], to improve internal homeostasis and adaptation to the environment [27], [63], or to convey a positive message in human-robot interaction contexts [23]. Following the approach adopted in [17], in the above models, pleasure was modeled in terms of release of simulated hormones that affect different aspects of the underlying robot architecture—in the above examples, the learning algorithms, the motivational decision-making model, and the expressive behavioral elements, respectively. In its simplest form, in these models and other related work in robotics (e.g., [44], [48], [55]), pleasure, associated with positive valence, was a "signal" (a modulating parameter), linked to the satisfaction of (survival-related or social) needs. In humans, however, pleasure is much more complex and has multiple sources, roles and forms [41], to the extent that it would be more accurate to talk about "pleasures" [9]. Can robot models contribute to the understanding of the complexities underlying pleasure, and how? In the remaining of this section, I will discuss two main types of contributions that embodied robot models can make to our understanding of pleasure: (1) contributions to the clarification and discussion of cross-cutting interdisciplinary research questions, and (2) concrete examples of operationalization and experimental testing of specific models and hypotheses.

5.1.1 Robot Models and Research Questions

(a) The nature of pleasure: one or many?

Many definitions of pleasure have been provided, reflecting its multi-faceted nature, multiple meanings, and the multiple mechanisms underlying various types of pleasure [9]. Frijda [41], for example, distinguishes among categories such as sensory pleasure, non-sensory likings, pleasures of achievement, pleasures of gain and relief, social pleasure, activity pleasures, or esthetic pleasures. Overarching the different views, however, we can identify a common element, also found in lay and dictionary definitions: the "liking", "positive affect", or enjoyment that produces a tendency

to continue the ongoing activity causing the enjoyment. Can we reconcile these two meanings—general "liking" and specific types of pleasure—in robot models?

(b) The relation between pleasure and emotion

Some consider pleasure as an emotion—e.g., one of the earliest forms of emotion that evolved [77]—others as "a constituent quality of certain emotions as well as a trigger for certain emotions" [30, page 76]. The term "pleasure" is often used to denote a positive (subjective) hedonic quality, an affective sensation or feeling of pleasantness grounded in the body. To some, pleasure has both sensory and affective elements [72], whereas for most, pleasure is *the* affective component of sensing [16], [58], a "gloss" [41] of sensory processing. How can robot models help us to "dissect" these complex relations?

(c) The relation between pleasure, valence and usefulness

How do we go from "liking" something to "wanting" it? Although the interactions between hedonic quality, positive affect, and valence can be very complex [60], pleasure is often associated with positive valence and utility: pleasure would have biological "utility"—adaptive value, evolutionary usefulness—signaling that stimuli are beneficial [15], [77] and promoting their acceptance [39]. This view is common in homeostasis-based models [60], where pleasure is largely related to need satisfaction, and in models of conditioned learning that view pleasure as "reward", both in humans [86] and in robots [27], [44], [55]. However, robot models can also be used to implement and investigate other types of pleasure and their relation to valence and usefulness.

5.1.2 Example: Robot Model of "Pleasure(s)"

The study in [63] was designed to model and experimentally test the above issues in the context of decision making (action selection) in a motivationally autonomous robot that must survive in its environment, as follows.

(a) Concerning the *nature of pleasure*, the study combines the broad definition of pleasure as "liking" that produces a tendency for the robot to continue the ongoing activity causing the "liking", and the idea of the multiple meanings and roles of pleasure, by focusing on two different contexts in which such "liking" can take place in a survival-related decision making problem: linked to the satisfaction of survival-related physiological needs, or as a purely hedonic quality not directly linked to need satisfaction.

(b) The *relation between pleasure and emotion/affect* was conceptualized in terms of the affective component of sensing and, following my longstanding approach, modeled in terms of hormone release and modulation of motivation-related perception. This modulation of motivation-related perception through pleasure provides a link between "liking" and "wanting", as it changes the attentional effort [87], or the incentive salience [10], of the stimuli. In Pessoa's words [79, page 190]: "At the perceptual level, items with affective/motivational content act as if they had increased *salience*, which improves performance if they are task relevant but impairs it if they are task irrelevant."

(c) Finally, regarding the *link between pleasure and valence*, often conceptualized in terms of "usefulness" and "reward" in affective neuroscience, psychology, and robotics, in [63] we depart from the idea that pleasure is necessarily linked with reward—in the same way as value is not necessarily linked with reward [57]—or with signaling biological usefulness, opening the door to the investigation of the role of other types of pleasure not directly related with the satisfaction of needs [41], in addition to pleasure stemming from need satisfaction. At the same time, we set to investigate whether hedonic quality (just "liking", "pure pleasure" unrelated to need satisfaction) might also play a role in motivation [39], [41].

For our pleasure study, we used the humanoid robot Nao, since we developed our pleasure model with the intention to include it in the Nao-based Robin companion robot [23]. In a simple "two-resource problem" (TRP) action selection task, the robot had to timely alternate between searching for and consuming two survival-related resources—small colored balls representing "water" and "food"—distributed around a walled environment in different ways in the different experimental conditions tested. Consuming those resources had associated different levels and types of pleasure in the various experimental conditions that we investigated, as we will see below.

Our robot's decision-making architecture builds on our model of "core affect" around a "physiology" that must be maintained within permissible limits for the robot to "survive"—remain "viable" and operational in its environment. Its main elements are two homeostatically-controlled essential variables (energy and hydration, replenished by consuming the appropriate resource), two motivations (hunger and thirst, modeled as functions combining the perception of internal physiological deficits and of relevant elements of the environment), and a number of behavioral systems that permit the robot to walk around, recover from falls, and look for and consume the resources to satisfy its motivations and correct physiological needs. In addition, this architecture includes a model of pleasure as a mechanism that acts on motivations—and hence on the decision making process—by modulating the perception of external stimuli. We modeled three kinds of pleasure:

1) A modulatory "pleasure hormone" dynamically released as a function of satisfaction of homeostatic needs, thus signaling an improvement in the interaction with the environment and fostering "openness"—i.e., increasing the interaction with the environment or continuing an interaction that is going well. Indirectly, this type of pleasure could be regarded as a signal of the "utility" of the elements of the environment.
2) In some of the experimental conditions, different fixed values of "pleasure hormone" were used as control conditions to compare with the "pleasure hormone" that fluctuated as a function of need satisfaction. Such fixed values can be thought of as background levels of hormones unaffected by interactions with the environment, as hormone-releasing chemicals (e.g., drugs) artificially added into the system, or as pathological conditions.
3) Additional hormone (a constant amount) was released linked to the execution of consummatory

behaviors of "eating" or "drinking" in one of the experiments. This additional release was unrelated to the satisfaction of needs, and corresponded to what we called "purely hedonic" pleasure.

We conducted three sets of experiments to assess the effect of these different types of pleasure under different environmental conditions. In the experiments, we varied the environments in terms of abundance, availability and symmetry of two resources (e.g., both resources equally abundant or scarce, one of the resources abundant, the other scarce, both equally or unequally accessible), as well as the pleasure associated with the resources (equally with both, and more, or different types of, pleasure associated with the abundant or with the scarce).

5.1.3 Discussion

We compared the behavior of robots whose pleasure hormones, released under different circumstances, play different roles, and measured them in terms of management of the viability of the robots "internal milieu" (in terms of maintenance of the homeostatically-controlled internal variables) and in terms of observable behavior. In all cases, the pleasure hormone acted on the "assignment of value" to the perceived resources, modifying their incentive motivational salience. In this way, pleasure also modulated how likely the robot was to interact with the perceived stimuli. Our results indicated that pleasure, including pleasure unrelated to need satisfaction, had adaptive value ("usefulness") for homeostatic management in terms of improved viability and increased flexibility in adaptive behavior.

Regarding improved viability, we found that the extent to which the different "types" of pleasure were adaptive or maladaptive depended on the features of environment and the demands it posed on the task, in addition to the "metabolism" of the robot. Whereas in some ("easy") environmental conditions, maximizing pleasure improved the viability of the robot, in others, a constant moderate level of pleasure (unrelated to need satisfaction) gave the best viability, yet in other, more "difficult" environments with asymmetric availability of resources, the addition of "purely sensory" pleasure associated with the scarce resource improved the viability of the robot.

Increased behavioral flexibility was observed particularly in terms of management of the trade-off between opportunism (taking advantages of the opportunities offered by the environment to satisfy a need that is not the most urgent at that point in time) and persistence (continuing working on a "goal" in order to satisfy a need). Notably, we found that pleasure allowed the robot to manage persistence and opportunism independently, and hence to display them in the appropriate context. This is important, for example, in situations where opportunism has a penalty but increased persistence is beneficial, and where an asymmetry in the availability of resources results in the need to consume each of the resources in different ways in order to achieve good management of homeostasis.

This study illustrates how a robot model can help analyze, operationalize and test, in a methodic and incremental way, hypotheses that cut across disciplines around the still ill-understood affective "dimension" of pleasure—or pleasures—and its different roles, underlying mechanisms, and links to different contexts. Further studies could contribute to a more systematic study of the roles of different "pleasures" and their underlying mechanisms in animals (including humans) and robots, moving beyond the view of pleasure as "reward" and as a signal of the utility of the stimuli that currently pervades neuroscience and robotics.

5.2 Arousal-Based Robot Models

Like pleasure, the notion of arousal is not univocal. In psychology, it is traditionally an abstract notion associated with general alertness, mobility, and readiness for action [39], [53], [77], [79]. As a "dimension" of emotions, affective arousal is generally associated with their level of "activity" or intensity [83], and the multiple phenomena and mechanisms grouped under this term have been blurred under the assumption of physiological uniformity, as discussed in [24], [39]. While physiological uniformity cannot be assumed, we are still far from understanding the precise roles of the different phenomena and mechanisms grouped under "arousal", and we cannot assume that they have clear-cut specific roles. From a modeling point of view, a general notion of affective arousal remains useful. However, when studying how "affect" and "cognition" are interrelated in affective cognition, we need to tease apart the different roles that arousal (or different aspects of it) might play in such interaction. A similar view is also espoused by some researchers in the brain sciences[4].

Let us discuss two of the aspects that we have explored using robot models[5] in collaboration with developmental psychologists, in the context of the development of attachment [12], [21], [93] between (human, chimpanzee and robot) "infants" and their human caregivers, and its influence on cognitive and affective development.

5.2.1 Affective Arousal and Cognition

Arousal, and specifically affective arousal, influences cognitive processes such as attention, memory, problem-solving and learning [92]. Traditionally, this relation is thought to have an inverted U-shape relation, in line with the "Yerkes-Dodson law": too low or too high levels of arousal are thought to interfere negatively, whereas middle levels have an "energizing" effect.

In embodied robots, arousal can be modeled as a parameter that indicates the level of "activity" of different elements of the architecture and embodiment of the robot. For example, the activity of a neural network as the robot is learning can be used as an indication of alertness, the activity of the sensors (highly or poorly stimulated) can

4. For example, Ferreira-Santos, discussing the role of arousal from the perspective of the affective predictive coding paradigm, argues that, to explain how arousal effects are triggered, different kinds of prediction errors need to be considered [37].

5. To take into account different interaction dynamics, as well as to avoid possible unwanted effects on the human perception of a specific type of embodiment, these models were implemented in various robots with very different embodiments and sensory-motor capabilities, such as the wheeled Khepera and Koala robots from K-Team (www.k-team.com), both of them fitted with rings of infra-red sensors around their bodies, the legged dog-like robot Aibo developed by Sony (older "third generation" versions from 2004) and humanoid Nao developed by Aldebaran, both of which rely primarily on vision and touch sensors.

contribute to the general arousal and alertness of the robot, the "mismatch" between a prediction and a perception (a prediction error) can be taken as a source of affective arousal (e.g., "anxiety") that will trigger corrective actions, in line with affective and interoceptive predictive coding approaches [6], [7], [37], [88]. In some contexts, such parameter can be taken as an indication of the "affective arousal" of the robot, e.g., of its "agitation", "uneasiness" or "distress" that can affect, for example, the learning rate (or other parameters) of a neural network or some other element of the architecture of the robot. The "arousal" parameter can then modulate the working of other elements of the robot architecture, such as the learning rate of a neural network.

Regarding learning in the context of attachment, we have, for example, studied the effects of novelty-induced arousal in the learning of a "baby" (Aibo) robot that explored its environment with the help of a human "caregiver", and learned to categorize and memorize the features of objects that it came across in this exploration [47]. In our robot model, novelty (objects with as-yet unknown features such as new colors or shapes, or presenting strong differences with previously encountered objects) increased the arousal—modeled as an internal parameter—of the robot. Two different mechanisms decreased arousal: (1) the "comforting" feedback provided by the human "caregiver"—either tactile, by stroking the robot, or visual, by showing her/his face in the visual field of the robot—decreased it to a greater extent (at a faster rate), and (2) a slower decay via self-regulation in the absence of human feedback for a prolonged period of time. Humans with different caregiving and interaction styles would provide different types of feedback (e.g., primarily tactile, primarily visual, or both), at different points (e.g., when "stress" caused by too much novelty was signaled by the robot, or at specific moments chosen by the human, such as when the robot seemed "stuck" in front of an object, or as "reward" for apparent progress in its exploration), and to different degrees (e.g., only initially to "bootstrap" the exploration of the robot or through the entire exploration episode). The study reported in [47] compared the effects of the feedback provided by humans using different caregiving styles ("attentive" or highly comforting and responsive to distress calls, versus "independent", letting the robot control its arousal by itself) in the pace and type of learning of environments with various degrees of novelty.

In other studies around arousal and learning, using Nao humanoid robots this time, we have investigated, for example, the role of arousal and its regulation to learn "affective landmarks" (secure or dangerous areas) [48], to learn tasks interactively [49], or as a "stress" signal for the robot to assess the difficulty of a learning task and solicit help from a human when the task was beyond its current abilities [51].

These examples illustrate how robot models allow us to assess the role(s) of arousal in "providing information" about urgency or importance [92] in different tasks and contexts.

5.2.2 Arousal Regulation in Interaction with Humans

Looking at the human side of the interaction, we have investigated the effect that different ways of expressing and regulating arousal by an autonomous robot can lead to the formation of very different perceptions, ideas, narratives, and "caregiving styles" in humans. For example, using the above-mentioned playmat scenario from [47], we carried out a 3–day study at the London Science Museum [46], [50] to assess human responses to robots behaving according to different attachment profiles: "needy"—soliciting the human often in a highly expressive manner when aroused while exploring the playmat, using multiple modalities such as flashing its LEDs, looking around for and at faces, and barking—versus "independent"—not soliciting the human and minimally expressive when aroused, and "focused" on exploring the playmat.

Adding *adaptation* to the equation, and based on the role of sensitivity in attachment [32], in a later study [52] we used the responsiveness of the humans to the expression of needs and distress of the robot as a signal that permits the robot to self-adapt—in terms of frequency and modality of its responses—the regulation and expression of its arousal as a function of the perceived interaction style of the human. To test this model, we carried out three experiments using three variants of a learning task in a complex environment containing novel and incongruent objects, in which a Nao robot had to learn the features of several objects located on a table, as follows.

A first experiment investigated how different caregiving styles can be suited to different characteristics of the strategies used to regulate stress in the "needy" and "independent" robot profiles in an exploration task. Our results showed that, to achieve the same results with the two robot profiles, different caregiving styles were needed: the "independent" robot needed less interaction with the human caregiver to progress in its exploration, whereas the "needy" robot needed an almost constant presence and "comfort" of the caregiver to progress in its learning with comparable dynamics.

A second experiment investigated how different types of interaction in a more demanding—more complex and difficult to learn—and stressful environment might affect differentially the cognitive and affective development of the "infant" (robot)—namely its regulatory, exploratory and learning patterns. Our results showed that the two robot profiles exhibit different behavioral dynamics: the "needy" robot, for which the comfort provided by the human decreased the arousal faster, stopped more often and spent more time learning, whereas the "independent" robot, for which the comfort provided lowered the level of arousal for a longer time, showed longer exploration episodes.

A third experiment investigated the use of adaptation to the responsiveness of the human caregiver as a suitable mechanism for the robot to deal with real-time variations in the caregiver's availability to respond to regulatory behaviors, and to adapt to different caregivers. The adaptation capability added to the architecture modulated the effect of the comfort provided by the human by modulating the parameters used to process the comfort received. Our results showed that the robot could modify its own profile autonomously along the "needy" – "independent" dimension: more comfort made the robot lean toward the "needy" profile, whereas unattended requests made the robot move towards a more "independent" profile.

5.2.3 Discussion

These studies, largely carried out in collaboration with developmental emotion psychologists Kim Bard and Jacqueline Nadel as part of interdisciplinary projects investigating emotion development, provide examples of robot models explicitly developed for interdisciplinary emotion research. Our aim, beyond "designing agents to understand infants" [81], was to develop *cross-disciplinary models* that could be used to both, analyze emotional development in biological systems, and synthesize affective robots. We have already illustrated how models from psychology permitted us to develop "useful" mechanisms for robots to learn and adapt autonomously in interaction with humans. Let us now illustrate how robot models fed back into psychology by allowing to investigate novel topics, as well as "classical" topics in novel ways.

Whereas the substantial developmental and comparative psychology literature on attachment has mostly studied "negative episodes" (distress, confusion, or fear) elicited by the introduction of an element external to the dyad under the "strange situation" paradigm [2], our models allowed psychologists to investigate: (a) other, less studied, situations—learning and exploration episodes—potentially stressful due to the novelty of the environment and the complexity of the objects and agents the infant can interact with; and (b) how a human can *positively* influence the exploration patterns and learning outcomes of a developing "infant"—in this case a robot endowed with an attachment subsystem.

A change of focus was also possible, beyond the traditional emphasis on the classification of "attachment styles" and their supposedly distinctive features. Whereas the resulting attachment bond can have different qualities (typically ranging from "secure" to "insecure" in attachment theory), our interdisciplinary work goes in the direction that there is no universal "golden standard" regarding a caregiving style to achieve an attachment bond of high quality. Different styles can be more or less suited to different characteristics of the infant (e.g., in terms of strategies used to regulate stress) and viceversa, and are strongly influenced by society and culture [54].

6 Concluding Remarks

In this paper, I have discussed the use of embodied robots as models for interdisciplinary emotion research. To conclude, I would like to reflect on some of the advantages that embodied robot models offer over other models.

As models, robots present very different features to other types of models such as computational models. Herbert Simon [89] distinguished between models that simulate a system by predicting its behavior and deriving consequences from premises (e.g., a system for weather prediction), and models that are a simulation of a system by embodying a few key features of that system and being put to behave in the same environment, governed by the same laws (e.g., a satellite is not a simulation of a moon, it is a moon, the "real thing"). I would argue that computational models of emotions fall in the first category (e.g., a computational model of emotions that predicts, and possibly generates, behavior given a number of premises and appraisal operations), and embodied autonomous robot models, such as those I have discussed in this paper, in the second. According to Simon, the first type of models are appropriate for understanding systems with many parameters, for which it is difficult to predict behavior without complex or extensive calculations, whereas the second type can also be used as a source of new knowledge to understand, by synthesis, the behavior of poorly known systems. The choice between one or the other type of model for (interdisciplinary) emotion research will depend on the type of research questions under investigation. However, when interaction is at stake, embodied autonomous robots present clear advantages.

As physical entities, autonomous robots engage humans in more natural interactions than simulations and virtual models. By "more natural" I mean that humans can use the sensory-motor modalities and interactions that we normally use with other humans and animals—such as tactile contact, moving together, physically holding the robot—and in the same physical space, in addition to other modalities—such as voice or vision—that can also be used with virtual agents, which do not share the same space with us.

Compared to other robot models of emotions, the specific type of embodied robot models that I have used in my research—with "internal" as well as "external" embodiment—address Adolphs' recommendation (cf. Section 2) that "if robotics is to be a science that can actually tell us something new about what emotions are, we need to engineer an internal processing architecture that goes beyond merely fooling humans into judging that the robot has emotions." They also open the door to investigating issues concerning the relations between "mind" and "body" in emotion research, in line with LeDoux's "processing approach" and enactivist accounts of emotion [24].

Finally, as affectively autonomous agents with their own needs, motivations, affective processes and interactions, that engage and disengage in interaction with humans in a coherent trade-off between attending to the human, being social, and being independent, provide a meaningful interaction partner that is more easily perceived and treated as an agent by us [23]. Such affectively autonomous robots could also make contributions to psychology and neuroscience well beyond the use of robots as passive perceptual stimuli, allowing us to study emotions in interacting agents in systematic and controlled ways.

Acknowledgments

I am grateful to the members of the Embodied Emotion, Cognition and (Inter-)Action Lab, past and present, for their contributions to various aspects of the research discussed here, and to three anonymous reviewers and the editors of this special issue for their useful suggestions to improve this manuscript. Funding was provided partly by the European Commission through grants HUMAINE (FP6-IST–507422), FEELIX GROWING (FP6-IST–045169), and ALIZ-E (FP7-ICT–248116), and partly by the University of Hertfordshire through various PhD studentships. The opinions expressed are solely the author's. This paper grew out of a presentation given at the 2014 Human-Robot-Interaction (HRI) Workshop "HRI: a Bridge between Robotics and Neuroscience", www.macs.hw.ac.uk/ kl360/HRI2014W/index.html.

REFERENCES

[1] R. Adolphs. Could a robot have emotion? theoretical perspectives from social cognitive neuroscience. In J.-M. Fellous, M. Arbib, eds., *Who Needs Emotions? The Brain Meets the Robot*, pp. 9–25. OUP, 2003.

[2] M. Ainsworth, M.C. Blehar, E. Waters, and S. Wall (1978). *Patterns of Attachment: A Psychological Study of the Strange Situation*. Hillsdale, NJ: Lawrence Erlbaum, 1978.

[3] J. L. Armony, D. Servan-Schreiber, J. D. Cohen, and J. E. LeDoux. Computational modeling of emotion: explorations through the anatomy and physiology of fear conditioning. *Trends in Cognitive Science*, 1(1): 28–34, April 1997.

[4] O. Avila-García and L. Cañamero. Using hormonal feedback to modulate action selection in a competitive scenario. In S. Schaal, A. Ijspeert, A. Billard, S. Vijayakumar, J. Hallam, J.-A. Meyer, eds., *From Animals to Animats 8: 8th Intl. Conf. on Simulation of Adaptive Behavior (SAB'04)*, pp. 243–252, Cambridge, MA, 2004. MIT Press.

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

[6] L.F. Barrett and M. Bar (2009). See it with feeling: affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364, 1325–1334.

[7] L.F. Barrett and W.K. Simmons (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, 16, 2015, 419–429.

[8] A. Beck, B. Stevens, K.A. Bard, and L. Cañamero (2012). Emotional Body Language Displayed by Artificial Agents. *ACM Transactions on Interactive Intelligent Systems*, 2012, 2(1): 2:1–2:29.

[9] K.C. Berridge, M.L. Kringelbach, eds (2010). *Pleasures of the Brain*. New York, NY: Oxford University Press, 2010.

[10] K.C. Berridge and T.E. Robinson (1998). What is the role of dopamine in reward? Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews* 28(3), 1998: 309–369.

[11] A. Blanchard and L. D. Cañamero. Developing affect-modulated behaviors: Stability, exploration, exploitation or imitation. In F. Kaplan, P.-Y. Oudeyer, A. Revel, P. Gaussier, J. Nadel, L. Berthouze, H. Kozima, C. Prince, and C. Balkenius, editors, *6th Int Workshop on Epigenetic Robotics (EpiRob'06)*, pages 17–24, Paris, France, September 2006. Lund University Cognitive Studies.

[12] J. Bowlby (1988). *A Secure Base: Parent-Child Attachment and Healthy Human Development*. Basic Books, 1988.

[13] C. Breazeal. *Designing Sociable Robots*. The MIT Press, 2002.

[14] R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47(2):139–159, 1991.

[15] M. Cabanac. Physiological role of pleasure. *Science* 173(4002), 1971: 1103–1107.

[16] M. Cabanac. The dialectics of pleasure. In: Kringelbach ML and Berridge KC (eds.), *Pleasures of the Brain*. Oxford University Press, New York, NY, 2010, pp. 113–124.

[17] L. Cañamero. Modeling motivations and emotions as a basis for intelligent behavior. In W. L. Johnson, ed., *First Intl. Conf. Autonomous Agents (Agents'97)*, pp. 148–155, NY, 1997. ACM Press.

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

[19] L. Cañamero. 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, 2005.

[20] L. Cañamero. Emotion understanding from the perspective of autonomous robots research. *Neural Networks*, 18(4):445–455, 2005.

[21] L. Cañamero, A.J. Blanchard, J. Nadel. Attachment Bonds for Human-Like Robots. *Intl. J. Humanoid Robotics* 3(3), 2006, pp. 301–320.

[22] L. D. Cañamero and O. Avila-García. A bottom-up investigation of emotional modulation in competitive scenarios. In A. Paiva, R. Prada, and R. W. Picard, editors, *Affective Computing and Intelligent Interaction, 2nd Intl. Conf., (ACII 2007)*, pages 243–252, Berlin Heidelberg, 2007. LNCS, Springer-Verlag.

[23] L. D. Cañamero and M. Lewis M. Making New "New AI" Friends: Designing a Social Robot for Diabetic Children from an Embodied AI Perspective. *Intl. Journal of Social Robotics*, 8(4): 523–537, 2016.

[24] G. Colombetti. *The Feeling Body: Affective Science Meets the Enactive Mind*. The MIT Press, Cambridge, MA, 2014.

[25] G. Colombetti. The Embodied and Situated Nature of Moods. *Philosophia* (2017) 45: 1437–1451.

[26] I. Cos, L. Cañamero and G.M. Hayes. Learning Affordances of Consummatory Behaviors: Motivation-Driven Adaptive Perception. *Adaptive Behavior*, 2010, 18(3–4): 285–314.

[27] I. Cos, L. Cañamero, G.M. Hayes, and A. Gillies. Hedonic Value: Enhancing Adaptation for Motivated Agents. *Adaptive Behavior*, 2013, 21(6): 465–483.

[28] B. Cox and J. L. Krichmar. Neuromodulation as robot controller. *Robotics and Automation Magazine*, 16(3):72–80, 2009.

[29] A. Damàsio. *Descartes' Error*. Avon Books, New York, NY, 1994.

[30] A. Damàsio. *The Feeling of what Happens: Body and Emotion in the Making of Consciousness*. Vintage, London, U.K., 1999.

[31] L. Damiano, A. Hiolle, and L. Cañamero. Grounding synthetic knowledge: An epistemological framework and criteria of relevance for the scientific exploration of life, affect and social cognition. In T. Lenaerts, M. Giacobini, and H. Bersini, editors, *Advances in Artificial Life, ECAL 2011: Proc. of the Eleventh European Conf. on the Synthesis and Simulation of Living Systems*, pages 200–207, Cambridge, MA, 2005. The MIT Press.

[32] M. De Wolf and M.H. van Ijzendoorn. Sensitivity and attachment: a meta-analysis on parental antecedents of infant attachment. *Child Development* 68, 1997, 571–591.

[33] P. Duerr, C. Mattiusi, A. Soltoggio, and D. Floreano. Evolvability of neuromodulated learning for robots. In *ECSIS Symposium on Learning and Adaptive Behaviors in Robotics Systems*, pages 41–46. IEEE Computer Society, IEEE Press, 2008.

[34] J.-M. Fellous. The neuromodulatory basis of emotion. *The Neuroscientist*, 5:283–294, 1999.

[35] J.-M. Fellous. From human emotions to robot emotions. In E. Hudlicka and L. Cañamero, editors, *Architectures for Modeling Emotions: Cross-Disciplinary Foundations. Papers from the 2004 AAAI Spring Symposium*, pages 37–47, Menlo Park, CA, 2004. AAAI Press.

[36] J.-M. Fellous and J. LeDoux. Toward basic principles for emotional processing. what the fearful brain tells the robot. In J.-M. Fellous and M. A. Arbib, eds., *Who Needs Emotions? The Brain Meets the Robot*, pages 79–115. Oxford University Press, 2003.

[37] F. Ferreira-Santos. The role of arousal in predictive coding. Commentary on Mather, M., Clewett, D., Sakaki, M., and Harley, C.W. (2016). Norepinephrine ignites local hot spots of neuronal excitation: How arousal amplifies selectivity in perception and memory. *Behavioral and Brain Sciences*, Volume 39, 2016, e207

[38] R. French and L. Cañamero. Introducing neuromodulation to a Braitenberg vehicle. In *IEEE Intl. Conf. on Robotics and Automation—Robots get Closer to Humans (ICRA 2005)*, pages 4199–4204, Barcelona, Spain, April 2005. IEEE Press.

[39] N.H. Frijda. *The Emotions*. Cambridge University Press, 1986.

[40] N.H. Frijda. Emotions in robots. In H. L. Roitblat and J.-A. Meyer, editors, *Comparative Approaches to Cognitive Science*, pages 501–517, Cambridge, MA, 2005. The MIT Press.

[41] N.H. Frijda. On the nature and function of pleasure. In *Pleasures of the Brain*, K. C. Berridge, M. L. Kringelbach, eds. New York, NY: Oxford University Press, 2010. chapter 6, pp. 99–112.

[42] N. Frijda and D. Moffat. Modeling emotion. *Japanese Journal of Cognitive Studies*, 1(1):5–15, 1994.

[43] N. Frijda and J. Swagerman. Can computers feel? theory and design of an emotional system. *Cognition and Emotion*, 1:235–258, 1994.

[44] S.C. Gadanho and J. Hallam. Emotion-triggered learning in autonomous robot control. *Cybernetics and Systems: An International Journal*, 2001, 32(5): 531–559.

[45] M.I. Garrido, G.R. Barnes, M. Sahani, R.J. Dolan. Functional Evidence for a Dual Route to Amygdala *Curr Biol.* 2012, Jan 24; 22(2-2): 129–134.

[46] A. Hiolle, K.A. Bard and L. Cañamero. Assessing Human Responses to Different Robot Attachment Profiles. In *Proc. 18th Annual IEEE International Symposium on Robot and Human Interactive Communication (IEEE RO-MAN 2009)*, pp. 251–256.

[47] A. Hiolle, L. Cañamero. Conscientious Caretaking for Autonomous Robots: An Arousal-Based Model of Exploratory Behavior. In *Proc. 8th International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems (EpiRob 2008)*, Lund University Cognitive Studies, 139. Lund: LUCS pp. 45–52.

[48] A. Hiolle, L. Cañamero. Learning Affective Landmarks. In *Proc. 9th International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems (EpiRob 2009)*. Lund University Cognitive Studies, 146. Lund: LUCS, pp. 211–212.

[49] A. Hiolle, L. Cañamero, P. , Andry, A. Blanchard, P. Gaussier P. Using the Interaction Rhythm as a Natural Reinforcement Signal for Social Robots: A Matter of Belief. In *Proc. Intl. Conf. Social Robotics, ICSR 2010*, Ge, S S., Li H., Cabibihan J-J., Tan Y K., Ed., pp. 81–89.

[50] A. Hiolle, L. Cañamero, M. Davila-Ross, and K.A. Bard. Eliciting Caregiving Behavior in Dyadic Human-robot Attachment-like Interactions. *ACM T. Interactive Intelligent Systems*, 2(1), 3:1–24, 2012.

[51] A. Hiolle, M. Lewis M., and L. Cañamero. A Robot that Uses Arousal to Detect Learning Challenges and Seek Help. In *Proc. 14th*

Conference on the Synthesis and Simulation of Living Systems (ALIFE 2014), pp. 864–871. Cambridge, MA: The MIT Press.

[52] A. Hiolle, M. Lewis M., and L. Cañamero. Arousal Regulation and Affective Adaptation to Human Responsiveness by a Robot that Explores and Learns a Novel Environment. *Frontiers in Neurorobotics*, 8(17), 2014.

[53] E.R. Kandel, J.H. Schwartz, and T.M. Jessell. *Essentials of Neural Science and Behavior*. Norwalk, CT: Appleton & Lange, 1995.

[54] H. Keller. Attachment and Culture. *Journal of Cross-Cultural Psychology*, Volume: 44 issue: 2, 2012, pp. 175–194.

[55] H. Kitano. A model for hormonal modulation of learning. In: *Proc. Fourteenth Intl. Joint Conf. on Artificial Intelligence*, volume 1. IJCAI, Montreal, Canada, 1995,, pp. 532–540.

[56] J. L. Krichmar. The neuromodulatory system: A framework for survival and adaptive behavior in a challenging world. *Adaptive Behavior*, 16(6):385–399, 2008.

[57] J. L. Krichmar and F. Röhrbein. Value and reward based learning in neurobots. *Frontiers in Neurorobotics* 7(13), 2013: 1–2.

[58] M.L. Kringelbach. The hedonic brain: A functional neuroanatomy of human pleasure. In: Kringelbach ML and Berridge KC (eds.) *Pleasures of the Brain*. Oxford Univ. Press, NY, 2010, pp. 202–221.

[59] J. LeDoux. *The Emotional Brain*. Simon & Schuster, New York, 1996.

[60] S. Leknes and I. Tracy. Pain and pleasure: Masters of mankind. In: Kringelbach ML and Berridge KC (eds.) *Pleasures of the Brain*. Oxford University Press, New York, NY, 2010, pp. 320–335.

[61] M. D. Lewis. Bridging emotion theory and neurobiology through dynamic systems modeling. *Beh. Brain Sciences*, 22:169–245, 2005.

[62] M. Lewis and L. Cañamero. Are discrete emotions useful in human-robot interaction? feedback from motion capture analysis. In *Affective Computing and Intelligent Interaction (ACII) 2013*, pages 97–102. IEEE Computer Society, IEEE Press, September 2013.

[63] M. Lewis and L. Cañamero. Hedonic quality or reward? A study of basic pleasure in homeostasis and decision making of a motivated autonomous robot. *Adaptive Behavior*, 2016, 24(5): 267–291.

[64] J. Lin, M. Spraragen, M. Zyda. Computational Models of Emotion and Cognition. *Advances in Cognitive Systems* 2(2012): 59–76.

[65] J. Lones, M. Lewis, and L. Cañamero. Epigenetic adaptation in action selection environments with temporal dynamics. In P. Lió, O. M. andG. Nicosia, S. Nolfi, and M. Pavone, editors, *Advances in Artificial Life, ECAL 2013: Proc. 12th European Conf. on the Synthesis and Simulation of Living Systems*, pp. 505–512. The MIT Press, 2013.

[66] J. Lones, M. Lewis, and L. Cañamero. A Hormone-Driven Epigenetic Mechanism for Adaptation in Autonomous Robots. *IEEE Trans. on Cognitive and Developmental Systems* 10(2), 2018: 445–454.

[67] R. Lowe, T. Ziemke, and M.D. Humphries. The dual-route hypothesis: Evaluating a neurocomputational model of fear conditioning in rats. *Connection Science* 21(1):15–37. March 2009.

[68] F. Mannella, S. Zappacosta, M. Mirolli, and G. Baldassarre. A computational model of the amygdala nuclei's role in second order conditioning. In M. Asada, J. Hallam, J.-A. Meyer, and J. Tani, editors, *From Animals to Animats 10 (SAB)*, pages 321–330, Berlin, Heidelberg, 2008. LNCS 5040, Springer-Verlag.

[69] S. Marsella, J. Gratch, and P. Petta. Computational models of emotion. In K. Scherer, T. Banzinger, and E. Roesch, editors, *Blueprint for Affective Computing: A Sourcebook*, pp. 21–46. OUP, 2010.

[70] A. Mehrabian and J. A. Russell. *An Approach to Environmental Psychology.* Cambridge, MA: The MIT Press, 1974.

[71] M. Minsky. *The Society of Mind*. Simon & Schuster, NY, 1985.

[72] J.P. Nafe. An experimental study of the affective qualities. *The American Journal of Psychology*, 1924, 35(4): 507–544.

[73] N. Navarro-Guerrero, R. Lowe, and S. Wertmer. A neurocomputational amygdala model of auditory fear conditioning: A hybrid system approach. In *Proc. of the IEEE Intl. Joint Conf. on Neural Networks (IJCNN)*, pages 214–221, 2012.

[74] M. Neal and J. Timmis. Timidity: A useful mechanism for robot control? *Informatica*, 27(4):197–204, 2003.

[75] J. K. O'Regan. How to build a robot that is conscious and feels. *Minds and Machines*, 22:117–136, 2012.

[76] J. Panksepp. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, New York, NY, 1998.

[77] A. Philippides, P. Husbands, T. Smith, and M. O'Shea. Flexible couplings: Diffusing neuromodulators and adaptive robotics. *Artificial Life*, 11:139–160, 2005.

[78] L. Pessoa. *The Cognitive-Emotional Brain: From Interactions to Integration.* Cambridge, MA: The MIT Press, 2013.

[79] L. Pessoa, R. Adolphs. Emotion processing and the amygdala: from a "low road" to "many roads" of evaluating biological significance *Nature Reviews Neuroscience*, Vol. 11, Nov. 2010, pp. 773–783.

[80] D. Petters. *Designing agents to understand infants.* Ph.D. thesis, School of Computer Science, The University of Birmingham.

[81] R. Picard. *Affective Computing*. MIT Press, Cambridge, MA, 1997.

[82] D. Purves, E.M. Brannon, R. Cabeza, S.A. Huettel, K.S. LaBar, M.L. Platt, and M.G. Woldorff. *Principles of Cognitive Neuroscience*. Sunderland, MA, Sinauer Associates, 2008.

[83] R. Reisenzein, E. Hudlicka, M. Dastani, J. Gratch, K. Hindriks, E. Lorini, J.-J. Meyer. Computational Modeling of Emotion: Towards Improving the Inter- and Intradisciplinary Exchange. *IEEE Transactions on Affective Computing*, 2013, vol. 4 (3), pp. 246–266.

[84] L.-F. Rodríguez and F. Ramos. Computational models of emotions for autonomous agents: major challenges. *Artificial Intelligence Review* 43, doi: 10.1007/s10462-012-9380-9.

[85] E.T. Rolls. *Emotion and Decision-Making Explained*. OUP, 2014.

[86] M. Sarter, W.J. Gehring, and R. Kozak. More attention must be paid: The neurobiology of attentional effort. *Brain Research. Brain Research Reviews* 51(2), 2006: 145–160.

[87] A. Seth. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, Nov. 2013, Vol. 17, No. 11, pp. 565–573.

[88] H.A. Simon. *The Sciences of the Artificial*. Cambridge, MA, The MIT Press, 1981 (2nd. ed.).

[89] O. Sporns and W. H. Alexander. Neuromodulation and plasticity in an autonomous robot. *Neural Networks*, 15:761–774, 2002.

[90] L. Steels. Towards a theory of emergent functionality. In J.-A. Meyer and S. W. Wilson, editors, *Proc. 1st Intl. Conf. Simulation of Adaptive Behavior: From Animals to Animats (SAB'91)*, pages 451–461, Cambridge, MA, 1991. The MIT Press.

[91] J. Storbeck and G.L. Clore. Affective Arousal as Information: How Affective Arousal Influences Judgments, Learning, and Memory. *Soc Personal Psychol Compass.* 2008, Sep 1; 2(5): 1824–1843.

[92] M.H. van Ijzendoorn, K.A. Bard, M.J. Bakermans-Kranenburg, and K. Ivan. Enhancement of attachment and cognitive development of young nursery-reared chimpanzees in responsive versus standard care. *Developmental Psychobiology* 51, 2009, 173–185.

[93] P. Vuilleumier, J.L. Armony , J. Driver, R.J. Dolan. Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nat. Neurosci.*, 2003, 6: 624–631

[94] W. Zhou and R. Coggins. Computational models of the amygdala and the orbitofrontal cortex: A hierarchical reinforcement learning system for robotic control. In B. McKay and J. Slaney, editors, *Advances in Artificial Intelligence*, pages 419–430, Berlin, Heidelberg, 2002. LNCS 2557, Springer-Verlag.

Lola Cañamero is Reader in Adaptive Systems and Head of the Embodied Emotion, Cognition and (Inter-)Action Lab in the School of Computer Science at the University of Hertfordshire in the UK, which she joined as faculty in 2001. She holds an undergraduate degree (*Licenciatura*) in Philosophy from the Complutense University of Madrid and a PhD in Computer Science (Artificial Intelligence) from the University of Paris-XI, France. She turned to Embodied AI and robotics as a postdoctoral fellow in the groups of Rodney Brooks at MIT (USA) and of Luc Steels at the VUB (Belgium). Since 1995, her research has investigated the interactions between motivation, emotion and embodied cognition and action from the perspectives of adaptation, development and evolution, using autonomous and social robots and artificial life simulations. Some of this research has been carried out as part of interdisciplinary projects where she has played Principal Investigator and coordinating roles, such as the EU-funded HUMAINE (on emotion-oriented information technology), FEELIX-GROWING (investigating emotion development in humans, non-human primates and robots), and ALIZ-E (development of social companions for children with diabetes), or currently the UH-funded Autonomous Robots as Embodied Models of Mental Disorders. She has played a pioneering role in nurturing the emotion modeling community. She is author or co-author of over 150 peer-reviewed publications in the above topics. Website: www.emotion-modeling.info.