Bridging the Gap between HRI and Neuroscience in Emotion Research: Robots as Models

Lola Cañamero
Embodied Emotion, Cognition and (Inter-)Action Lab
School of Computer Science
University of Hertfordshire
College Lane, Hatfield
Herts, AL10 9AB
United Kingdom
L.Canamero@herts.ac.uk

ABSTRACT

Emotions are a fundamental aspect of social cognition and interaction, and their importance has been acknowledged by both the HRI and neuroscience communities. 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. In this paper, I argue that the use of robots as both stimuli (tools) and models is meaningful in both disciplines as far as emotion research is concerned. I focus on the view of robots as models, illustrate it with representative examples from both disciplines and their potential for cross-fertilization, and discuss ways in which the use of robots as interaction partners is central towards bridging the interdisciplinary gap.

1. INTRODUCTION

“This suggestion is admittedly vague and 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.” [20, (Jean-Marc Fellous, 1999)]

The workshop website suggests the use of robots as models as being the perspective of robotics, their use as a stimulus as the perspective of neuroscience, and their use as an interaction partner as the result of integrating the two perspectives. I will however argue that both the “stimulus” and “model” roles can be found in both disciplines as far as emotion research is concerned although, for the sake of space, in this paper I focus on the robot a model.

Emotions are a fundamental aspect of social (as well as non-social) cognition and interaction. The HRI community has acknowledged their importance and considerable work is being done to endow social robots with different of emotional capabilities tailored to their interaction with humans. Inspiration is often taken from psychological models and, to a lesser extent, from neuroscience, to build operational models of specific aspects of emotions in robots and study empirically the effects that endowing robots with different emotion elements and capabilities have in humans, sometimes using the robot as a stimulus rather than as a model. The predominant interest in doing so is to increase the believability and human likeness of the robots to make them better understood and accepted by humans. The interest in a more conceptual understanding of emotions is not as yet as present in this community as in other areas of robotics. At the same time, over the last decade, the study of emotion has become a core topic in neuroscientific research, and the use of robots as tools (or stimuli) for this research is starting to attract interest and provide very interesting results. Emotions thus 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.

Modeling different aspects of emotions in robots and computers can be considered the realm of the roboticist, and work in this area has flourished over the last two decades (see e.g., [42, 33] for overviews of computational and robotic models of emotions stemming from specific neuroscientific or psychological models). This is not necessarily synonymous with the view of robots as models of (different aspects of) emotions. Robotics often take inspiration from psychological or neuroscientific models as heuristics, purely or primarily with the purpose of building robots can function efficiently—in this case endowing them with some particular mechanism that achieves some of the adaptive functions that emotions have in social interactions in humans. I have called this approach “engineering emotions” [9] and, although it can be very useful for the purpose of building robots that efficiently function in their environment and engage humans in (specific types of) interaction, it usually does not provide much feedback to the neuroscientific or psychological model that inspired the design of the behavior of the robot.
since the robotic implementations do not shed additional light on the rationale or the mechanisms underlying such behavior/function.

Using the robot as a model involves the additional step of building an operational model with the purpose of contributing to the (ideally inter- or trans-disciplinary) understanding of the phenomenon under scrutiny [11, 12]. From a robotics perspective, such understanding can be of a theoretical nature, e.g., understanding the scope and implications of a model, or a more empirical one, e.g., understanding its use in HRI by making and testing specific predictions about how specific aspects of the model affect some dimension in the human-robot interaction loop. Although less common, conceiving of robots as models (beyond their use as tools) has also attracted interest from affective neuroscientists and psychologists. This interest in modeling can have different flavors. In the remainder of the paper, I will examine representative examples from both disciplines and their potential for cross-fertilization, and discuss ways in which the use of robots as interaction partners is central towards bridging the interdisciplinary gap.

2. ROBOT AS HYPOTHETICAL MODELS

2.1 Philosophical Issues

A theoretical speculation regarding e.g., whether a robot could have emotions [1] or how to build one that could have them [39], could 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. As neuroscientist Ralph Adolphs [1, page 9] puts it (italics are mine):

“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 that I have highlighted in italics and that we will see reappear in the other types of models discussed in the remainder of the paper. I kindly ask the reader to keep those highlighted terms in mind and think about them when reading about those other models. I will simply give a word of caveat and offer a general remark at this point.

The word of caveat concerns the fact that, while using sound criteria backed-up and grounded in a theory is very important to ascertain what emotions (of elements thereof) can be, we should be careful to avoid the circular argument that defines emotions to be X, Y and Z, and then call 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 “meta-analyses” issued from interdisciplinary effort [29, 16].

As a general remark, those highlighted issues are on the one hand important centuries-old philosophical (and scientific) questions, and on the other hand a potentially powerful tool that, used in an open-minded critical theoretical reflection can make the emotion theorist (e.g., psychologist or neuroscientist) think in different ways about what the key concepts of their theory could “really” mean—what they entail, what alternative types of mechanisms could underly them, what they could have been given different evolutionary / developmental / socio-cultural histories, etc.

2.2 Design Blueprints

Between philosophical and “operational” models lie what we could term “design blueprints”. The most relevant example here is perhaps Valentino Braitenberg’s “Vehicles” [6], a set of thought experiments in robot and robot-environment interaction. A neuroscientist, Braitenberg took inspiration from the evolution and dynamics of the brain to design very simple robots in terms of simple sensors and motors directly connected. From such simple robots, very complex behavior can emerge as a function of their interactions with the environment, teaching us the lesson that complex behavior can emerge from very simple “machinery”. Starting with the simplest case of a robot “getting around”, Braitenberg proceeds to incrementally vary connections between sensors and motors and add other simple elements to give rise to (in an emergent fashion) different behaviors, affective phenomena and cognitive capabilities. Interestingly, the biological foundations of the Vehicles are provided at the end of the book. Perhaps less well known in HRI, this “blueprint” that was conceived as a series of thought experiments by a neuroscientist became a very popular and fruitful design philosophy and source for many actual robotic models and studies (e.g., [22]). Such models offer potential for interdisciplinary research in a very similar way as the model discussed in detail in Section 4.2 and for the sake of space, I refer the reader to the discussion in that section.

3. ROBOTS AS MODELS OF SPECIFIC EMOTIONAL PHENOMENA

Robots have also interested neuroscientists as potential testbeds for specific emotional phenomena. An interesting example is offered by Joseph LeDoux, whose studies of fear processing in the amygdala have been very influential in the robotics community. As argued in [21], the modeling approach he advocates 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 such as conditioning learning, memory, perception of emotional meaning, and even feelings and consciousness [21, pages 80, 82]:

“The approach presented here is a straightforward experimental approach to emotion, which
avoided vague concepts such as “affect”, “hedonic tone”, and “emotional feelings.” It is important that the mistakes of the past be not made again, and that we expand from this foundation into broader aspects of mind and behavior. [...] We shall speak of the processing approach to emotions as the approach we propose here, which grounds emotion in possibly unconscious processes. [...] An understanding of these more fundamental processes is what the processing approach is all about.

The processing approach allows us to study unconscious emotional functions similarly in humans and other animals.”

... and robots, as he draws on the experience gained through his experimental approach in the neuroscientific investigation of fear to offer advice to computational modeling and roboticists [21, 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 [...] These basic principles might serve as a starting point in the design of computational models of emotions.”

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 Emotional memories differ from other memories in that they last longer and are more vivid.

P4 There are two parallel routes of emotional processing of a stimulus, one fast, the other slower and modulates the fast route.

P5 There are two separate inputs to an emotional evaluation system: one for simple stimuli such as a tone, the other for complex stimuli such as context.

P6 Emotional expressions are triggered by a central signal but the specifics of expressions are determined locally in the circuit according to the state of the animal.

A number of computational and robotic models have been inspired by such (and similar) principles. 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. However, the advantages from the perspective of neuroscience are, at present, more limited. The more useful models go beyond a restrictive “biomimetic” approach that tries to replicate (by design) the structural features of emotional circuits found in biological systems (what I have termed “structural models” [9]), to try to model principles that are key to the understanding of how emotions function at different levels and how emotional information and behavior relate to anatomy and physiology [2, 4, 45, 32, 31, 35].

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 robotic models would target at “closing the loop” by making some unique contribution arising from the use of robots. Such type of contributions requires close collaboration across disciplines.

Another issue that this type of models need to resolve 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? Whereas the brain area or circuit in question will be involved in many known and unknown functions through diverse known and unknown mechanisms, the computational construct only models part of the structure and its involvement in function. Are the modeled properties sufficient to justify the use of the biological term? How useful can that practice be? How misleading? Attempting to solve such fundamental issues involves, again, a deep interdisciplinary effort.

4. ROBOTS AS 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.

Once more, this type of “functional modeling” is not the exclusive realm of roboticists and has also attracted interest from psychologists and neuroscientists. In this section I will consider one example from each field and illustrate the counterpart robotics models with examples of my own work.

4.1 Regulation of Agent-Environment Interactions

Psychologist Nico Frijda adopts a functionalist view 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 [23], Frijda proposed guidelines to implement a functional model of emotions in a robot. From a functional point of view, we need to identify the properties of the structure of humans (and other animals) and their environment that are relevant to the study of emotions and are shared by a structurally different system (where system = robot + environment or niche), and we need to model them and model them so as to give rise to the same functions or roles. 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 constitute relevant models of emotions. According to Frijda, these properties are as follows.

Properties of humans relevant for the understanding of emotion: (a) they are autonomous; (b) they have multiple ultimate goals or concerns; (c) they possess the capacity to emit and respond to (social) signals; (d) they possess limited energy- and information-processing resources and a single set of effectors.

Features of the human environment relevant to emotions: (a) it has limited resources for the satisfaction of concerns; (b) it contains signals that can be used to recognize that opportunities for the satisfaction of goals/concerns, and occurrences of threats might be present; (c) it is uncertain; (d) it is in part social.

From the above characteristics (relevant to the understanding of emotion) of the human system and environment, Frijda posits the following functions of emotion:

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 real-locating resources.
F4 To do this in parallel for all concerns, simultaneously working toward whatever goal the system is actively pursuing at a given moment, and even in the absence of any actually pursued goal.

Such functions are valid across structurally different architectures and environments. A system possessing mechanisms that fulfill (some of) these roles (if they satisfy the above criteria, cf. Section 2), can thus be said to (partially) have emotions from a functional point of view.

Functional requirements leave more freedom concerning the particular elements to be used in order to achieve these functionalities. In other words, this model remains underspecified as far as the underlying design and implementation mechanisms are concerned. Such underspecification raises many conceptual and design issues [9, 10, 11] such as the choice of the level and complexity of the model, the design approach (e.g., engineered, emergent or evolved emotions), the analysis and design of the environment, or the connection with cognition, motivation and behavior, to name a few. However, while they pose hard problems, such issues provide wonderful challenges and opportunities for theoretical and empirical exploration and interdisciplinary collaboration.

For example, despite the fact that the initial computational model designed by Frijda and his collaborators [25, 24] 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 robotic models. They were for example used in my early work [8] to model basic emotions fulfilling the above functions as part of an autonomous agent designed from the opposite perspective of embodied AI [7, 34] and that integrated elements from ethological and neuroscientific models of behavior, motivation and emotion (taking particular inspiration from the models of [15, 27, 40]), to implement their underlying mechanisms and interaction dynamics. 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 and looking at problems from a vantage point that a single-model perspective. It also permits 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. There are however important questions that such type of model cannot answer due to the selection and explicit design of specific emotional subsystems with specific predefined functions. For example, this model cannot explain how some traits of emotions might emerge from or be the side effect of others. It cannot answer the question of what would be the minimal set of mechanisms that would allow to generate behavior that could be qualified (according to some criteria) of “emotion-like” or characteristic of specific emotions. Due to its complexity, it also makes it very 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, as we will see in the next section.

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 [20]:

“...one of the main functions of emotions is to achieve the multi-level communication of simplified but high impact information. By "simplified" I mean using as little "technical" resources as possible, by "impact" I mean the ability to be understood and interpreted in ways that significantly change the behavior of a receiver. Doing so is a tough job for two reasons. First, simplifying a message decreases its chances to be decoded properly [...] Second, increasing the impact of a message often complexifies it [...] Emotions, somehow, have evolved to achieve the tradeoff.”

Critizicing models that posit specialized emotion brain centers, he advocates a view of emotions as dynamical patterns of neuromodulations rather than patterns of neural activity [19]. 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 extents depending on the complexity and properties of the circuitry. Evolutionarily older systems (e.g., underlying reflexes) are more resilient to neuromodulation and hence are less affected by emotions. Evolutionarily newer systems (e.g., underlying motivation, attention, memory, learning,
etc) have an increasingly larger potential for neuromodulation and therefore for emotional modulation. This theory has some important consequences regarding the conceptualization of emotions [20]:

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 Basic emotions are better described as attractor states. While emotions are continuous, attractor states are discrete.

C4 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 [20]:

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 are not combinations of some pre-specified basic emotions. Such an implementation would implicitly assume that such basic emotions are independent from each other.

G4 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.

However, not being a roboticist, Fellous calls for collaborations between neuroscientists and roboticists to carry the more specific modeling and implementation of these principles in robots, and particularly a mechanisms that could correspond to neuromodulation in robots.

Since the paper was written, roboticists have started providing some answers to the latter question. Models of neuromodulation in robots span different aspects such as: increasing the plasticity and flexibility of the robot controller [43, 14], improving its evolvability [41, 18], improving its adaptation and survival by affecting its action selection [22, 26, 37], producing different emotion-like behaviors (different functionalities) from the same underlying controller [3, 13] or affecting its emotional development [5, 30].

Let us have a closer look at the models of emotional modulation [3, 13] that model emotions as side-effect of the interactions among the different elements of the architecture of robot and the environment, in other words as “emergent functionality” [44], 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 the dynamics of selected examples of emotional modulation of the underlying “nervous system” (and hence of “cognition-action”) in a more abstract way. In [3], an autonomous robot endowed with internal survival-related homeostatic variables that constitute drives motivating it to look for and consume the appropriate resource out of the two (light and dark paper patterns located on the floor and that the robot could perceive using infrared sensors) available in the environment (a walled area enclosing the resources or “arena”) to satisfy the most urgent need in a timely fashion in order to survive. The most urgent need changes dynamically as a function of the robot’s interaction with and perception of the environment. This is called in the adaptive behavior literature a “two-resource problem” (TRP), considered to be the simplest action selection problem. While a solitary robot could easily survive in the initial environment, the inclusion in the same environment of a second identical robot trying to solve its own TRP made the problem harder for each robot, turning it into a “competitive” TRP (C-TRP). This was due to the fact that, although the robots could not communicate with each other and were not even “aware” that another robot was present, they limited resource availability for the other robot when they were consuming or blocking access to them. In such C-TRP environment, robots endowed with the same internal architecture as in the TRP “died” easily, and the main causes identified were “goal obstructions”—access to resources blocked by the other robot—and “overopportunism”—excessive opportunistic consumption of the less needed resource when the most needed one is blocked, leading to such high increase of the error of neglected homeostatic variable that it could not be corrected in time. Such problems arose because the robots didn’t have the capability to interact with 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.

1The 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.
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 the perception element (in this case a parameter influencing what we could describe as the “salience” of the external stimuli) of the TRP architecture produced “flee-fight” behavior providing some basic “competition skills” that permitted the robot to overcome the above-mentioned problems of the CTRP. 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 (carried 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 greatly improved the problem of overopportunism. Its effect on the contact sensor produced behavior akin to a “flee-fight” system, since when a robot bumped into the other when the latter was blocking the resource the former was trying to consume, if would “push it” out of the resource when it had enough “stamina” (low risk of death) due to the cancelation of the contact sensor, and would move away and “abandon the goal” when it didn’t (high risk of death). This greatly improved the “goal obstruction” problem. 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 got 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.

Following an incremental approach, we complexified 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, the 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 successfully solved the CTRP was again unsuccessful in this new C3RP environment, primarily due to the fact that, when they prey was attacked, it rarely managed to “get rid” of the predator to either reach the nest and heal, or attend to the other survival needs. Increasing the “escape” speed of the prey would have been an easy but not entirely satisfactory solution since the prey would still be repeatedly damaged and suffer frequent interruptions when attending to its other needs. We again opted for treating this as a problem on the perception side of the loop\(^3\), in this case, a time-related problem that could be solved if perception had an effect on the robot’s behavior before damage was caused. This time we explored the temporal properties of hormonal modulation and introduced another hormone that affected interception—namely the perception of the level of damage—and had a slower decay. Following [36], 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 attending to this need and working towards avoiding the predator, before it was too close, thus managing to survive in the C3RP much more easily. This system can be regarded as a simple model of phenomena underlying the anticipatory role of “anxiety”.

Despite the fact that these robotic models did not explicitly intend to implement Fellous’ model, they meet the “conceptual consequences” and guidelines it provides, notably C1, C2, C4, G1, G2, G3 and G4. As for C3, although it is certainly possible to describe the above models using the vocabulary of dynamical systems in a “metaphorical” way, we did not carry a formal mathematical analysis in terms of dynamical systems theory, and therefore I will not make any claims in this point.

Despite (or perhaps due to) their simplicity, the value of this type of models goes well beyond the fact that they can implement and demonstrate in robots the principles and criteria of neuroscientific models. They provide an excellent framework to test neuroscientific hypotheses, make predictions and test them in a very incremental and highly controlled and methodic way. The also foster identification and isolation of a small number of key variables to be studied and high control of the parameters used; this permits to focus on detailed investigations of the interactions among different elements of the system (where system = robot + (physical and social) environment). The simplicity and transparency of the architecture of the robot, and of the links of its different elements with behavior and interaction also permits detailed quantitative (as well as qualitative) analysis of the system from “inside” and “outside” at different “levels” (e.g., “physiological”, behavioral, interactional), as well as the interactions among such levels, using methods and metrics from different disciplines.

Finally, the use of interacting robots (as opposed to robots interacting with humans) can be very valuable to isolate key aspects of the interaction that cannot be isolated when one of the interaction partners is a human, whether the model tested stems from neuroscience, psychology, or HRI.

\(^2\)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 of the robot’s physiological space described in [3].

\(^3\)Further studies in this direction [38, 37] investigated various trade-offs in the embodiments of prey and predator, particularly regarding proximal and distal perception.
5. BUILDING BRIDGES: INTERACTION

Interaction is the keystone in the bridge between HRI and neuroscience, and this in different ways. Clearly, interaction between (among) disciplines is crucial to make progress in bridging the gap.

Another important, and perhaps more unique, aspect in which “interaction” is crucial towards bridging the gap between HRI and neuroscience is what the workshop organizers suggested as the “third perspective” from which we can look at robots (in addition to “stimulus” and “model”) and the result of integrating the two other perspectives: robots as interaction partners. In a sense, the robot as interaction partner can be the result of integrating their use as “stimulus” and “model” to help us understand the many big unresolved questions about ourselves and our interactions with others from a perspective that encompasses (at a “meta-level”) and brings together (by combining models, methods, techniques and tools) that of the single disciplines. However, I think that in addition to a result, taking seriously the robot as an interaction partner is a prerequisite to bridging the gap. At present, there is a gap between “interaction” and “partner” in both HRI and neuroscience, aspects that are for the most part studied rather in isolation from each other, certainly as far as emotion is concerned. The present interest in the use of robots as stimuli to study emotion both in neuroscience and HRI carries the potential risk of limiting the use of robots as passive perceptual stimuli, as is currently the case in numerous perceptual studies. Such perspective is massively underexploring the great potential that robots have to interact as agents, rather than as mere 3D devices that can look more or less human-like and move. Similarly, models of emotion in neuroscience and HRI (and generally robotics) tend to be either individual-centered (the vast majority) or interaction-centered, but very rarely integrate both dimensions. My suggestion for overcoming such limitations involves starting by thinking of, analyzing, and modeling (synthesizing) the robot as a proper, active interaction partner, a “whole agent” or “complete creature” [7] in which perception and action and interaction with the environment are tightly coupled and intertwined, and where the interaction provides the context for such analysis and synthesis [28, 17]. Such perspective shift requires the involvement of many other disciplines too, from embodied cognitive science and robotics to philosophy to the social sciences dynamical systems theory [29], to give some examples.

Acknowledgments

The writing of this paper and part of the research reported here has received partial support from the European Commission under grants ALIZ-E (FP7-ICT–248116), FEELIX GROWING (FP6-IST–045169) and HUMAINE (FP6-IST–507422). The opinions expressed are solely the author’s.

6. REFERENCES


