From Human Emotions to Robot Emotions

Jean-Marc Fellous

The Salk Institute for Neurobiological Studies
10010 N. Torrey Pines Road, la Jolla, CA 92037
fellous@salk.edu

Abstract[1]

The main difficulties that researchers face in understanding emotions are difficulties only because of the narrow-mindedness of our views on emotions. We are not able to free ourselves from the notion that emotions are necessarily *human* emotions. I will argue that if animals *have* emotions, then so can robots. Studies in neuroscience have shown that animal models, though having limitations, have significantly contributed to our understanding of the functional and mechanistic aspects of emotions. I will suggest that one of the main functions of emotions is to achieve the multi-level communication of simplified but high impact information. The way this function is achieved in the brain depends on the species, and on the specific emotion considered. The classical view that emotions are 'computed' by specialized brain centers, such as the 'limbic system', is criticized. I will suggest that an ensemble of well-known neurobiological phenomena, together referred to as *neuromodulation*, provide a useful framework for understanding how emotions arise, are maintained, and interact with other aspects of behavior and cognitive processing. This framework suggests new ways in which robot emotions can be implemented and fulfill their function.

There are many inherent aspects of emotions that are extremely difficult to study and to account for. I will start by listing a few of them, and then suggest that it may be more fruitful to focus on the functions of emotions rather than on what emotions are. I will then suggest that animals do in fact have emotions, at least functionally, even though we might not be able to empathize with them. This will lead to the conclusion that, functionally, robots could have emotions as well. I will then briefly open a new window on the neural bases of emotions that may offer new ways of thinking about implementing robot-emotions.

Why are emotions so difficult to study?

A difficulty in studying human emotions is that here are significant individual differences, based on experiential as well as genetic factors (Rolls, 1998; Ortony, 2002; Davidson, 2003a, b; Ortony et al., 2004). My fear at the sight of a bear may be very different from the fear experienced by a park-ranger who has a better sense for bear-danger and knows how to react. My fear might also be different from that of another individual who has had about the same amount of exposure to bears, but who is more prone to risk-taking behaviors. For emotions such as perceptual fear, the commonalities in expression, experience and underlying neurophysiological mechanisms are large enough between individuals, and between species, so that it can be studied in animal models (LeDoux, 1996). However, the issue of individual differences may be overwhelming for emotions such as love or depression.

Naturally occurring human emotions often arise in sequences, or in time varying intensity, and often outlast the stimuli that elicited them. After a Halloween prank, fear may yield to anger at the perpetrators followed by shame at oneself for being angry with neighborhood kids on Halloween. The fear of falling at the start of a roller coaster ride becomes less intense towards the end of the ride, even though the curves and speeds might be very similar. It may therefore be an oversimplification to speak of an emotional 'state', because emotions may be intrinsically dynamical phenomena of widely different time constants (from a few seconds for perceptual fear, to hours or days for moods, to months or years for depression or love). This makes the study of emotion more difficult since the emotional (and cognitive) contexts have often to be accounted for. Laboratory studies of emotions try to limit and control these factors. However, one must question the extent to which emotions in such controlled settings resemble the naturally occurring emotions. This problem is even more acute if, as I will suggest below, one of the primary roles of emotion is to

---

[1] New address: Duke University, Biomedical Engineering Department and Center for Cognitive Neuroscience, P.O. Box 90281, Durham, NC 27708-0281. fellous@duke.edu.

The author wishes to thank Dr. Michael Arbib for many constructive comments on previous versions of this manuscript.

Paper presented at the AAAI Spring 2004 symposium on Architectures for Modeling Emotion: Cross-Disciplinary Foundations. E. Hudlicka and L. Canamero.

modulate the flow of behaviors, in which case studying emotion in a carefully controlled setting might defeat its purpose.

Should we always express our emotions? In some circumstances, it may be better to inhibit our emotions rather than letting them all out. Are all emotions useful? Some emotions may be useless or counterproductive. For example, you may be caught in unexpected evening traffic on a freeway on your way back home. Even if you made no particular plans and are not late for anything, anger may easily set in. This emotion has apparently no purpose and no actual eliciting object (no one is individually responsible for this traffic), and you would be better off using this time to listen to a new radio station, or to reflect on a problem that needs to be addressed. Why do we have this emotion?

A difficulty, from the neuroscience point of view, is that there does not seem to be any emotional center in the brain. The idea that a 'left brain', a 'limbic system' or an 'amygdala' is there to provide the brain with emotions has been scientifically and convincingly disproved. However, there are specific brain areas that are specifically involved in specific emotions. For example, a long line of work by LeDoux and others has shown how specific nuclei of the amygdala process specific aspects of fear (LeDoux, 1996, 2002). This work provides useful information and understanding about fear conditioning, and possibly other kinds of fear. But no claims are made about the amygdala being the 'fear center' of the brain. On the contrary, the more researchers know about the role of the amygdala in fear, the more they learn about the involvement of other structures such as the hypothalamus, the prefrontal cortex or the brain stem neuromodulatory centers. Of course, none of these areas can be labeled as 'fear centers' either, because of their known involvement in non-emotional behaviors. Said differently, there is no set of brain areas you could lesion to selectively eliminate fear (without affecting other (non-emotional) brain functions). Unfortunately, because it has been such a convenient way of invoking the brain, the idea of "emotional centers" still pervades the thinking of many in Artificial Intelligence (AI), psychology, and philosophy. There is no emotional homunculus in the brain. The neural substrate of emotion is far more puzzling than commonly thought, and it may be fundamentally different from the one used for perceiving an object, or memorizing an item. An understanding of its true nature might uncover fundamental basic principles on multi-functional use of computational resources, parallelism, control of behavior and information flow and so on. These principles can in turn inspire the design of revolutionary new software and robotic systems.

What are emotions for?

Rather than seek to define 'emotion', I will focus on the functional roles of emotions. I will however define feelings as a subclass of emotions that may involve some form of consciousness, and/or perceived bodily feedback. The problem of understanding feelings is a special case of the mind-body problem. Some emotions such as guilt or shame require a sense of Self that includes the ability to evaluate one's own states. LeDoux and others have shown that it is not necessary to understand feelings to understand emotional behaviors. But eventually, researchers will have to elucidate the neural correlates of consciousness (Rees et al., 2002; Crick and Koch, 2003) and of the Self (Jeannerod, 2004) and the possibility of making a robot (self)conscious (Dennett, 1997).

I will concentrate here on delineating some of the basic functions of emotions with the hope that understanding what emotions *are for* will inform an understanding of what emotions *are*. However, some caveats are in order. In general, finding a function to a phenomenon is often a matter of being clever. Care must be taken to find a functional formulation that gives insights, rather than give explanations for everything. Looking for the possible functions of emotions is not to say that all individual emotions are functional at all times. Rather, it is to find some evolutionary (in the Darwinian sense), system-wide functions that help understand how emotions interact with the other functions the brain performs. For the purpose of this paper, I will not go further into the evolution of emotion, noting that we still know very little of evolution in general, so that a hypothesis on how emotions evolved, as insightful as it may be, has to rely on other hypotheses about how other system evolved. It is however clear that emotions have co-evolved with perceptual, cognitive and motor abilities. But what of 'useless emotions' such as rage, mentioned above? The modern world provides humans with new ways of eliciting emotions – whether the frustrations of the highway or the availability of drugs – that were not part of the environment in which the underlying mechanisms evolved. Furthermore, the emotional system comes as a package, and there may be fundamental (yet still unknown, but see below) reasons why some emotions sometimes are expressed and experienced in seemingly counter-productive manner, if the useful ones are to be so important and reliable. Rage is a small evolutionary price to pay for expressing fear and taking appropriate actions in a dangerous situation. The way the brain maximizes the benefits of emotion while minimizing its occasional inappropriateness is however still an open question.

Many researchers have explicitly proposed functions for emotions. For Rolls, emotions have ten major functions: 1) change in autonomic and endocrine system, 2) flexibility of behavioral response to reinforcing stimuli, 3) triggering

motivated behaviors, 4) communication, 5) social bonding, 6) improving survival, 7) affecting cognitive processing (mood-congruence) and facilitating its continuity, 8) facilitating memory storage, 9) allowing the persistence of motivated behaviors, and 10) facilitating the recall of memories (Rolls, 1998). For Levenson, emotions are used to coordinate behavioral response systems, shift behavioral hierarchies, recruit physiological support, short-cut cognitive processing, communicate and control, and establish our position in relation to other people, ideas and objects (Levenson, 1994). For Averill, there are different kinds of functions depending on the scope of their consequences. These consequences can be divided into four categories: intended vs. unintended, short-term vs. long-term, targeted to the individual vs. to the society or species and singular vs. predictable (Averill, 1994). How many of these functions really require "emotions" and distinguish them from other modes of processing and how many of these functions are necessary for a robot, are still open questions. Rather than attempting to list all the possible functions of emotions (or those that have been attributed to emotions by many researchers), I will list a few function of emotions that have natural robotics counterparts, and that all seem to share a common theme.

In general, purposeful communication requires that both emitter and receiver share common 'technical' conventions on how to communicate. English syntax is used for language-type communication between individuals; synaptic transmission (neurotransmitters and receptors) is used when cortical area A needs to communicate with area B. Communication also requires common models of each other's interpretations of the messages. Two humans need to understand the meaning of a given word the same way (similar 'semantic nets'); For A to excite B, A has to somehow mobilize the right subset of glutamatergic (i.e. excitatory) neurons. These models may be partly learned and partly genetically designed.

I will argue that one of the main functions of emotion is **to achieve a 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: saying one word out of three may introduce ambiguities, activating only a few glutamatergic neurons in A might have no effects, or unintended effects on B. Second, increasing the impact of a message often complexifies it: telling a story well requires that a lot of background and context be made explicit to the reader, for A to drive B for a long period of time, A must recruit many of its neurons in specific ways (e.g. synchronously, with inhibition and excitation properly timed) that are computationally 'expensive' to achieve. Emotions, somehow, have evolved to achieve the tradeoff.
An emotion is initiated at or close to the level that commits the organism to an overall course of action; this message is then broadcast to mobilize other levels that can support the course of action or modify the style of computation of certain subsystems appropriately. Thus the computational challenge is two-fold: (a) to find a relatively compact vocabulary for such messages (noting the debate over whether this vocabulary simply combines a set of "basic emotions" or has some more complex structure) and (b) to design subsystems that can best make use of such messages in ensuring the success (by appropriate criteria) of the organism or robot. Processes (a) and (b) cannot be separated – and thus systems with different behavioral repertoires, task-sets and success criteria may come up with very different solutions, thus ensuring that the "emotion vocabulary" in (a) may differ widely between different agents, and be very "un-human" in many robotic systems.

A fear reaction (scream, facial expression, adrenaline rush) is extremely poor in information (nothing can be inferred from the scream, facial expression, or level of adrenaline as to the cause for alarm), but the impact of this reaction is high (in others, and on one's own body). Note that the "emotion vocabularies" used for social communication with others, and for communication to the body may be very different from each other. Yet, in both cases, the two types of communications are low information and high impact. The diffuse brain release of norepinephrine (simple communication from the locus coeruleus to say the cortex) will drastically affect the excitability of neurons in A, and synaptic transmission between A and B, so that A will easily be able to drive B, without requiring much internal synchrony or excitation/inhibition balancing from A.
Emotions use communication channels (hormonal, facial, vocal expressions) that have evolved to carry such abstract/simplified representations of complex stimuli, situations, memories (Schwarz and Clore, 1983) or the perception and action biases that are linked with them. These representations are informative and may be used internally to trigger or modulate ongoing actions and thought processes (autonomic, hormonal), or can be communicated to others (facial, vocal expressions). When the information is 'decoded', its impact is significant on perception (e.g. focus of attention), information processing (e.g. trigger 'danger detectors', speed of neural communication) or action (e.g. start running instead of walking).

This aspect of emotion is apparent at many levels of granularity, from communication between humans to communication between neurons, and at many level of

information processing. Works by Aaron Sloman and Andrew Ortony offer separate perspectives on what these levels can be (see their chapters in (Fellous and Arbib, 2003)). From the point of view of robotics, this functional view can be translated in at least three important 'implementable' domains.

Communication. Whether they are vocal, postural or facial, emotional expressions are compact messages exchanged between individuals. Not paying attention to some of them may yield catastrophes (copulating with an angry male, or ignoring the fearful expression of a conspecific that has detected the presence of a predator). Beyond matters of life and death, expression increase the efficiency of communication, as noted between human and robots (Picard, 1997).

Resource mobilization, and conservation (Clark and Watson, 1994) and **prioritization of behaviors** (Simon, 1967). These operating system-like tasks rely on compact signals that have high-impact on the functioning of an autonomous agent. The underlying hypothesis is that emotions are one way to help animals cope with the complexity and unreliability of our environment (external or internal) and that have evolved partly because of the constraints that are imposed on our bodies and mind (time, physical limits, energy resources).

Decoupling stimulus and response (Scherer, 1994). Emotions allow for context dependent computations. Without boredom, we would be stuck processing the same stimulus in the same way. Without curiosity, we would never try anything new. 'Boredom' and 'curiosity' (if you accept them as emotions) change our perception of our world and the way we process it. Extending this notion further, emotions may themselves become stimuli, and push the organism to action (emotions as motivations (Arkin, 2004)).

These functions can easily be implemented. In fact, some computers and robots already have some of these functionalities embedded in their operating system, or as add-ons. I would however argue that because these functions still have little to do with one another (e.g. no 'expressions' due to resource mobilization, context-dependent computations do not depend on perceived expressions), they are more engineering hacks than attempts at implementing emotions. I would further claim that putting those functions together, and making them interact with one another in a way that optimizes their individual performance amounts in effect to starting the design of an emotional system.

The idea of emotion for adaptation, survival and success is familiar to roboticists. But what of species that are extremely successful (ants) and have no emotions we can empathize with. Ants and bees are successful because they rely on social constructs and cooperation, not on their individual ability to adapt. This sort of group behavior has been the focus of much AI interest in what is now called 'swarm intelligence' (Bonabeau et al., 1999). So, by extension, are there 'swarm emotions'? This issue is a particular example of emotions as emerging phenomena, rather than a phenomenon that has specifically evolved to perform some function. Similar arguments have been made by some who argue that emotions have evolved from a complex system of mutual regulation of behaviors and that the subjective meaning of an emotion is given by the observation of the response of other people to it (Brothers, 2001). Others suggest that emotions are a result of natural selection pressures and that they are not a qualitatively distinct set of subjective experience, but just a particular set of action tendencies (learned or innate) that happen to involve the body, and hence trigger 'feelings' (Dennett, 1991). Unfortunately, little is known of the neural substrate of social behavior (but see (Adolphs, 2004)) or of action tendencies. However, such approaches raise the issue of 'internal' versus 'external' emotions.

Some emotions are elicited by external stimuli (fear), or are directed at others (anger) and are usually thought of as 'external'. Other emotions are more 'internal' in essence (depression, happiness) and may not have a clear eliciting stimulus or target. As mentioned above, these two classes of emotions may differ significantly in the 'technical' ways they operate (facial expression Vs serotonin down-regulation). I would argue however that the more we know of the neurobiology of 'external' emotions, the blurrier this distinction becomes. For example, attachment behaviors (an 'external emotion') in prairies voles (a small rodent) depends crucially on the levels of neuromodulatory substances such as oxytocin and vasopressin (Insel, 1997; LeDoux, 2002), which many would argue, are involved in 'internal' emotions such as anxiety and stress (Carter et al., 2001; Neumann, 2003). Similarly, serotonin levels are well known determinants of 'internal' emotions such as depression, but are also involved in the regulation of social hierarchy. So-called 'internal' and 'external' emotions are different, but may be related in many interesting ways.

Who has emotions: The animal *vs.* the human

Do animals have emotions? The more an animal 'looks like' us (2 eyes, 2 ears, nose and mouth), the more we show empathy, and attribute emotions to them. It has been convenient to use human terminology to characterize animal emotions: your dog is happy, your cat is afraid. But is this anthropomorphism, or genuine detection of emotion? In the former case, it would say nothing of whether these animals actually *have* emotions. In the latter case, we could argue that if evolution has given us such a good '(human) emotion detector' (because it works so well in humans), and if this

detector is genuinely triggered by the behavior of your dog, then there is a good chance that your dog actually is experiencing this emotion. Otherwise, our 'emotion detectors' would be somewhat faulty. I would argue that the answer is probably a mixture of the two: animals do trigger our 'emotion detectors', but we are using the wrong vocabulary to interpret their outputs. We unfortunately cannot do otherwise: the 'proper' characterization of animal emotions is not important enough to force the evolution of a specific dog-and-cat-emotion vocabulary or concept in human language. For most practical purposes, and for those animals that are important to us (cats, dogs, horses or cows), using our garden variety human-emotion 'detectors', and human-emotion language is good enough. How do we know then that animals do indeed have emotions, and to what extent does their emotional experience differ from ours? I believe the answer lies partly with neuroscientists. The similarity in expression or behaviors (at least in certain animals (Darwin, 1872)), together with the similarities in brain structures indicate that there is no reason to think that animals do not in fact have emotions. The extent to which those emotions are similar to ours can be rigorously studied by looking at how their brain differs from ours. Some would argue that there may be quantitative differences in the complexity and depth of emotions (more than between 2 humans?), while others would argue that evolution may yield qualitative differences. Most would agree however that functionally at least, animals that are evolutionarily close to us, do indeed have emotions; not *human* emotions, but *animal* emotions. What about evolutionarily distant animals? Do amoebae, frogs, ants or snakes have emotions? Is there a transition on the evolutionary scale below which no animals have emotions and above which all the others do? If it where the case, we could find a species and its predecessor, one of which would have emotions, while the other one would not. An analysis of their genetic makeup, or possibly just an analysis of their nervous system, would therefore pinpoint to a singularity that would unequivocally be the *sine-qua-non* characteristic of emotion. Evolution does not seem to work that way. All known structures and functions of any living organism have evolved smoothly and gradually, and all is a matter of degree. It is certainly interesting to note jumps in animal behaviors, such as for prairie and mountain voles, two very similar species, the former showing social attachment behaviors while the latter does not (LeDoux, 2002). Using the proper terminology, one could speculate that prairie animals have voles-love while mountain ones do not. Perhaps a hasty conclusion, as it might simply be a difference of expression (the mountain vole might have other ways of showing vole-love), rather than a difference of emotional set. In any case, both species have fear and other common kinds of emotions, so voles are not the transition point. Do frogs have emotions? Certainly none we can empathize with, but given the nature of evolution, and the unlikely possibility that a species-transition point exists, it seems reasonable to conclude that, at least functionally, frogs do have frog-emotions. The structural (underlying mechanism) link between frog-fear and human-fear is certainly a complicated one, but we already know of some commonalities. For example, in both species, fear is correlated with the release of certain stress hormones, and other specific neuromodulatory substances. The function of fear in these two very different species is very similar (self-protection, escape and so on), although their expression and underlying mechanisms is significantly different and 'optimized' to the specifics of their bodies, internal organs and so on.

In sum, our ability to attribute (recognize, label) emotion to a species is constrained by the paucity of our emotion language, and we are forced to use human-emotion words. By doing so, we activate essentially human notions of emotions, and all their human connotations ('semantic net'). We are biased toward attributing emotions only to species we can empathize with to some sufficient degree. Our ability to attribute emotions is not a true reflection of whether animals have emotions. Structural and functional analyses of emotion, rooted in their evolution, suggest that animals have, at least functionally, emotions. So what of robots?

One might argue that robots and animals (including human) are fundamentally different. The former are silicon and mechanical devices, the latter are water-based, gene-regulated, neuron-controlled entities. They are of course entirely unrelated evolutionarily and therefore the argument I used to suggest that indeed all animals have emotions does not apply here. But as we understand the biological systems more and more, we are able to simulate them numerically better and better to the extent where the simulation actually predicts the biological system. This is particularly evident in the fields of computational neuroscience and biomedical engineering. With enough sophistication and computing power, this computational/mechanistic understanding allows for the extraction of the *basic principles* that underlies how a brain area or an organ works. Those basic principles are by definition implementation independent, and can be implemented in plastic and silicon (artificial heart, artificial retinas, artificial valves, cochlear implants, artificial retinas, contact lenses). Of course, simulating a heart is different from 'being' a heart, but for most (known) functional purposes, artificial hearts are just as good. The functions of a heart have been very well studied, qualitatively and quantitatively. There is of course still room for improvements, but critical mass has already been reached and artificial hearts save lives. Moreover, in the process of designing and building an artificial heart, scientists (cardiologist and engineers alike) have learned a tremendous lot about how the heart works. I would argue that so will be the case of emotions. An understanding of their function (qualitatively and quantitatively) in animals will lead to the

derivation of basic principles such as the ones discussed above that are independent of biological ancestry and human-like emotional expressions. These principles might then be instantiated for particular robotics architectures and task-sets, as appropriate. The sophistication of the instantiation and the careful adherence to the basic principles make the implementation an actual robot-emotion. Note that this is not to say that all robots need emotions, but that in principle, robot-emotions are indeed implementable, using human and animal emotions as a source for basic principles.

So can robots "have" emotions? If you ask a patient who has been implanted with a mechanical device that pump his blood in the center of his chest if he has a heart, his answer will most certainly be "Yes, I *have an artificial heart*!" Similarly, it will come a time when you will be able to ask your computer if it has emotions, and its answer will undoubtedly be "Yes, I *have computer-emotions*!" In the meantime, how do we even begin to think about how to implement emotions? Why not use the brain as a source of inspiration?

Emotions and the Brain

In order to scientifically study emotions, we need to be able to measure them, and if possible, to manipulate them. Because of the multi-level nature of emotion, from visceral to cognitive, many 'access points' are possible. While all levels bring useful insights, too high a level (e.g. psychoanalysis) leaves room for too many unconstrained hypotheses, and hence many chances for confabulating work. On the other hand, low levels of investigation (e.g. molecules and atoms) are too complex, do not allow for the separation of the emotional from the non-emotional, and no basic principles can be easily extracted. Of course, a multi-level approach is necessary, but have we identified all the possible useful levels of investigations? In this section, I suggest that the neural level, while extremely useful, does not give a handle on a very fundamental aspect of emotion that can potentially explain many evolutionary, functional and structural aspects of our emotions, and that can inspire the implementation of robot-emotions in novel ways. I will focus here on the brain, and will speculate next on the possible consequences for robotic implementation. The fundamental aspect missed by a neural-level analysis is the intimate relationship between emotion and neuromodulation.

Neuromodulation refers to the action on nerve cells of a large family of endogenous substances called neuromodulators that include dopamine, norepinephrine and serotonin. These substances may be released by a few specialized nuclei that have somewhat diffuse projections throughout the brain and that receive inputs from brain areas that are involved at all levels of behaviors (from reflexes to cognition). They can also be released locally within a brain area by neurons that manufacture these substances, away from these centers. Some other neuromodulatory substances are released as a function of the activity of the target neurons (e.g. nitric oxide (Boehning and Snyder, 2003) or cannabinoids (Iversen, 2003; Sjostrom et al., 2003)), and are therefore qualified as 'activity dependent', or considered to be 'retrograde signals'. Each of these neuromodulators typically activates specific families of receptors that are inserted in neuronal membranes. Each receptor has very specific and synergistic effects on the neurons at various time scales (from few milliseconds to minutes and hours). Most of these effects can be described by electrophysiological parameters such as average membrane potential, excitability or synaptic strength (Kaczmarek and Levitan, 1987; Hasselmo, 1995). Each neuron has its own mix of receptors, depending on where it is located in the brain. Interestingly, even two neuron of the same anatomical structure might have significantly different receptor mixes for reasons that are still unclear.

Neural activity is related to Action Potentials (APs, the main event that can be detected by other neurons, through synaptic connections) generation, and can be quantified as a rate (number of APs per seconds), or as some measure of their timing. Similarly, neuromodulatory activity (or neuromodulation) can be defined as the state of activation of a subset (or possibly all) of its receptors: how many dopamine receptors are activated, how many serotonin receptors and so on. Unlike neural activity, neuromodulatory activity is a multi-dimensional quantity (each dimension is a receptor type). Patterns of neural activity are spatio-temporal description of neural activity. Similarly, patterns of neuromodulations are defined as spatio-temporal patterns of neuromodulatory activity. Neuromodulation affects neural activity by changing the way the neuron 'computes' (e.g. its excitability, its synaptic integrative properties) and hence the way it generates action potentials. The activation of a specific receptor subtype results in a *coordinated* modification of a specific set of biophysical parameters (input resistance, average membrane potential, synaptic strength and so on), so that the 'neuromodulatory space' is of much lower dimensionality than the 'biophysical parameter space'. Neuromodulatory substances are released by the activity of specialized neurons (either locally or from neuromodulatory centers) that are themselves controlled by 'regular' neurons (or neural activity). Neural and neuromodulatory activities are certainly related, but there is no known isomorphism between them; in other words, because you know one does not mean that you know anything of the other (even if you add glutamate and GABA in your definition of 'neuromodulator'). They are just two different ways of quantifying brain activity.

I argue here that it may be crucial to understand emotions as **dynamical patterns of neuromodulations**, rather than patterns of neural activity, as is currently done.
Many experimental facts point to the involvement of specific neuromodulatory substances in the initiation, maintenance or termination of emotional states (Fellous, 1999). Neuromodulation may be a very fruitful level of description of the mechanism underlying emotions, as is the case for depression and serotonin regulation for example. Note that in this case, the neuromodulatory centers (the raphe nuclei) do not seem to be defective per se; the problem resides at the level of the receptors. See also the examples of the prairie voles and the involvement of oxytocin and vasopressin (two neuromodulators) in social attachment behaviors (Insel, 1997), or the involvement of stress hormones in stress. In the latter case, the level of description is so adequate that the name of the modulator itself reflect its involvement in an emotion! These findings, and many others, serve to define the particular subset of neuromodulations that is of interest for a particular emotion. Whether or not there exists a set of neuromodulations that characterize all emotions is an open question, which at this point can only be the subject of wild speculations.

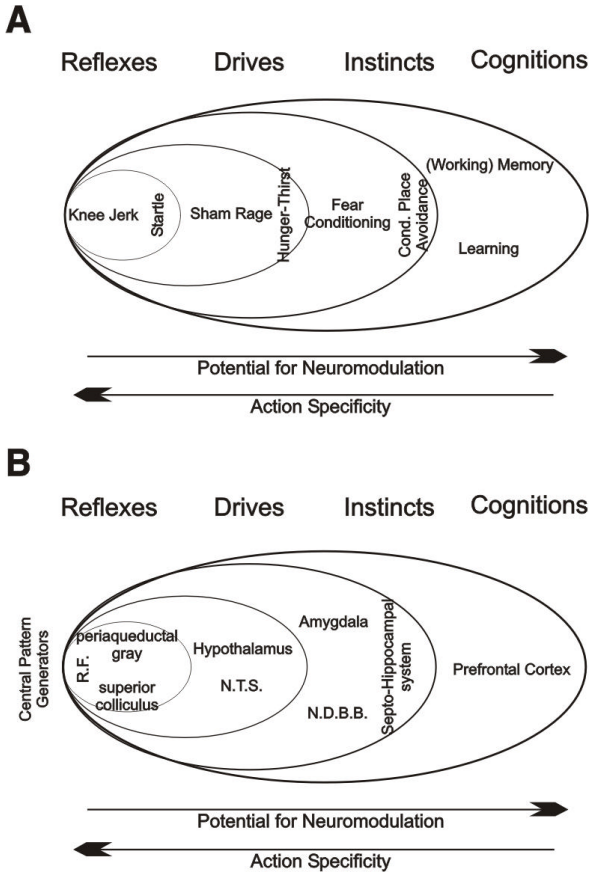

Fig 1: ***A****: Organization of behavior with respect to potential for neuromodulation and action specificity. Reflexes are fixed motor patterns, the neural substrate of which undergoes few neuromodulations, while 'cognitions' are unspecific with respect to sensory stimuli and are heavily susceptible to neuromodulation. Ellipses represent zone of direct influence and neural recruitment during emotional expression and experience.* ***B****: Mapping of brain structures to Reflexes, Drives, Instincts, and Cognitions. Abbreviations: NDBB (nucleus of the diagonal band of Broca), RF (reticular formation), NTS (nucleus of the solitary tract). Details can be found in (Fellous 1999).*

In this formalism, emotions are patterns of neuromodulations that affect brain areas involved at all levels of functions, from low-level motor control to planning and high-level cognition. The extent to which each of these functions is affected by the emotional dynamics depends therefore on the amount and nature of the neuromodulation its underlying neural substrate is capable of. Structures involved in reflexes have intrinsically less potential for neuromodulation than cortical structures that are involved in say planning; reflexes are therefore expected to be less affected by emotions than planning.

The 'potential for neuromodulation' can be used as an organizing principle for neural structures. The behaviors these structures mediate can be organized accordingly and can suggest a new 'neuromodulatory-based' behavioral order, from low-level reflexes, to drives, to instincts and to cognition (Fig 1, see also (Lane, 2000)). Note that this order is very similar to the ones suggested by others (Ortony or Sloman) but has been *derived* from a quantitative observations of some aspect of brain structures ('potential for neuromodulation'), rather than proposed *de facto*. Each of these levels may in turn activate neuromodulatory centers depending on whether it projects to them. Because emotional states are dynamical states (time varying patterns of neuromodulation) there is no initiation of emotional state per se. Instead, each level can bias the neuromodulatory pattern toward or away from an attractor state, a pattern of neuromodulation that is temporally and/or spatially stable. Labeled emotions such as fear are attractor states that bear structural similarities (same kinds of receptors, in the same kinds of areas) between individuals so that common patterns of expressions can be consistently noticed, and therefore labeled. For example, the fear-attractor state is such that the amygdala appears, across individuals, as a major player in the neuronal circuitry processing sensory information. Note that this does not necessarily mean that the amygdala itself is modulated, but that the overall effect of this peculiar stable neuromodulatory pattern is to configure the various areas it controls so as to make the amygdala a key player. In between these neuromodulatory attractor states, the brain may not in any labeled (according to human standards) emotional state, or may be in some transient state that bares some resemblance to several attractor states ('the feeling of being a bit angry and depressed…').
Neuromodulatory patterns are biased at many levels of processing, from reflexes to cognition, depending on the

nature of the projections from these brain centers to neuromodulatory centers. It is unlikely, for example, that a reflex would bias an emotion, because the reflex circuitry has no access to neuromodulatory centers. Conversely, it is possible to have a purely cognitive process bias the ongoing pattern of neuromodulation and elicit a strong felt emotion because the cortex projects heavily to many neuromodulatory centers. This view is compatible with others who sees multilevel generation of emotion as a central requirement for an emotional system (Picard, 2002), with the exception that here emotions are 'biased' rather than 'initiated'.

Patterns of neuromodulations, and therefore labeled emotions, can be externally influenced, and scientifically studied by pharmacological challenges (e.g. substance of abuse, Prozac), neurochemical intervention (intra-cerebral blockade of specific receptors in specific areas) or by neurophysiological means (stimulation or lesion of neuromodulatory centers). Unfortunately, such manipulations are still too global (injection of a substance in the blood stream), or too imprecise (lesion of a brain structure) to yield a selective manipulation of the emotional dynamics. Techniques such as neuro-imaging of receptor activation and reversible inactivation of neuro-anatomically restricted families of receptors would be required. These techniques are around the corner.

This theory has five consequences:

1) Emotion is not the product of neural computations *per se*. The fact that some structures are more involved in emotions than others results from the fact that they are more susceptible to neuromodulation and that they are anatomically in a position to mediate the proper emotional expression. For example, the amygdala may be viewed (possibly among other things) as a species-specific 'danger-detector' partly hard-wired and partly established through experience and learning (see the notion of automatic appraiser in (Ekman, 1994)); a 'pattern matcher' that constantly monitors the many modalities for sensory patterns that would probabilistically indicate danger. The amygdala projects to various centers that are required to express fearful behavior, to neuromodulatory centers as well as to structures that project to neuromodulatory centers (Aggleton, 1992). A possible scenario is that a fearful stimulus activates this detector (and possibly others), neuromodulatory signals are generated, and bias the ongoing neuromodulatory pattern towards a 'fear-attractor pattern'. This neuromodulatory pattern affects many other brain areas (and through hypothalamic influences possibly the body), and sets up their processing characteristics (decrease the excitability of planning areas, increase synaptic release probabilities along the thalamo-amygdala pathway and so on).

2) The coupling between the emotional state (e.g. anxiety, an attractor state) and information processing (e.g. memory of a car crash) does not presuppose that either one has a predominant nor causal role. Emotion and cognition are integrated 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 measured by say action potential rate). Neuromodulation depends on neural activity, and neural activity depends on neuromodulation. An emotion (neuromodulation pattern) tunes the physiological parameters of neurons, and the information these neurons process has the potential to affect how this tuning is achieved (some neurons project to neuromodulatory centers). Because of this tight two-way interaction cognitions and emotions are not, in general, separable entities. In this view, emotions are indeed more motivational (action tendencies, biases, resource allocations (Simon, 1967)) than behavioral (trigger specific actions).

3) The notion of basic emotions is replaced by the notion of 'attractor states'. Because of the conservation of neuroanatomical structures, their general connectivity patterns, and their relation to neuromodulatory centers, there will be patterns of neuromodulations that are similar across species and individuals, and that are more stable than others. Those patterns (attractors) may be considered as 'basic emotions' (e.g. fear). Because the exact underlying structure of such attractors need not be unique, differences between attractors in different individuals are possible (and likely), and account, at least in part, for individual differences in the expression and experience of the same emotion. Emotions are therefore continuous (one always is in an emotional state, what Damasio calls 'background emotions' (Damasio, 1999)), but attractor states are discrete, and hence can be labeled ('fear', 'surprise' and so on). The co-existence and co-design of many stable attractors in any dynamical system may result in the occasional transient 'spurious' emergence of some of them. Some specific patterns of neuromodulation may transiently occur in response to stimuli that were not meant (genetically, or through learning) to elicit them, and may result in 'useless' emotions (e.g. rage elicited by a traffic jam). In this sense, I agree with Ekman and his statement that all (labeled) emotions are basic (Ekman, 1992; Ekman, 1994), because they are all attractor states.

4) This view explains the very direct link between pharmacological challenges (e.g. drug of abuses, Prozac) and emotional state. For example, Prozac modifies the ongoing pattern of neuromodulation by selectively increasing serotonin receptors activation, in brain areas where serotonin is released. Prozac has therefore the potential to shift the neuromodulatory pattern from a local attractor (depression) to another (neutral, or happy).

5) Neuromodulatory phenomena commonly occur on a large range of time scales (from milliseconds to hours and possibly days), while neural activity is restricted to the millisecond time scales. This difference accounts, at least in part, for the fact that emotions (neuromodulatory patterns)

may significantly outlast their eliciting conditions (stimulus related patterns of neural activity). Emotions may have a 'life of their own', although they can certainly be biased by stimuli.

I will conclude this brief summary by stating that this view of emotions as patterns of neuromodulations is not in contradiction with the more classical, neuron-centered views (Damasio, 1994; LeDoux, 1996; Rolls, 1998; Damasio, 1999; LeDoux, 2002). It simply offers a different way (admittedly more complex) to characterize the genesis and impact of emotions on behaviors and cognitive processes.

Concluding remarks: From neuromodulation to robot-emotions

I have argued that there is no one emotion center in the brain. At best, many brain areas interact and mediate different aspects of different emotions, with the caveat that none of these areas are actually specialized for emotion processing, and that most have known non-emotional ('cognitive') functions. Emotions arise in sequences, and in a dynamical fashion, depending on the stimuli presented, the previous memories formed, and the current or recently past emotions. Not all emotions thus elicited are useful. I proposed however that one of the main functions of emotion is to achieve the communication of simplified but high impact information. This communication is implemented at multiple levels of functioning, from communication between brain areas to communication between brain and body, to communication between individuals. Part of its function is to achieve resource mobilization and conservation, to prioritize behaviors, and to decouple responses from stimuli. I have argued that indeed all animals have emotions in their functional sense, even though we may not empathize with them. While structurally (mechanistically) possibly very different from one animal to another, the functions of emotions are well conserved, and constitute a safe and stable starting point for understanding the basic principles that underlie emotion in general and specific emotions in particular. On the basis of this analysis, I suggested that robots might indeed be endowed with features that can functionally be related to emotion, and that they can indeed have robot-emotions, in the same sense as animals have animal-emotions, even though animals and robots are evolutionarily unrelated.

I finally suggested that neuromodulation offered a new way to quantify and characterize emotional dynamics that had advantages over the more neuron-centered view. This view offers a natural way to account for 1) the lack of 'emotional center' in the brain, 2) the non-causal interdependence between 'emotion' and 'cognition', 3) the emergence (rather than pre-specification) of 'basic' emotions, 4) the intimate relationship between emotional dynamics and pharmacological challenges and 5) the wide difference of time scale between emotions and between emotions and cognitions.

What may be the consequences of this analysis for the design of robot-emotions?

Not being a roboticist, I can only speculate.

1) Emotions should not be implemented as a separate, specialized module in charge of computing an emotional value on some dimension. While useful as a first step, such implementations would not be capable of handling the complexity of the emotional repertoire, its wide time scales, and its interactions with newly acquired knowledge (items that are not pre-specified, but learned through experience).

2) Emotions should not simply be the result of cognitive evaluations. While it is clear that such evaluations may explain some emotions (Ortony et al., 1988), it is almost certain that not all emotions are generated cognitively (Arbib, 1992). Implementing emotions as production rules would be significantly limiting, and would fail to capture the 'true nature' and function of emotions in general.

3) Emotions are not linear (or non linear) combinations of some pre-specified basic emotions. Such an implementation would implicitly assume that such basic emotions are independent from each other (or that there exist such an independent emotional set, in the mathematical sense).

4) Emotions should be allowed to have their own temporal dynamics, and should be allowed to interact with one another. Implementing emotions as 'states' fails to capture the way emotions emerge, wax and wane, and subside. Those temporal characteristics are not simple 'transients' and may have functional consequences.

As I suggested above, for the nervous tissue, neuromodulation fulfills these constraints. But what would neuromodulation correspond to in robotics term? I would venture to suggest that some aspects of neuromodulation could be implemented at the operating system level (i.e. architecturally, not as a process) as system wide control of some of the parameters of the many ongoing and parallel processes that make up the robot behavior. Each of these processes would have a handle on the way this control is achieved, and could influence it in their own specialized way. 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 and operating system designs. These advances in turn could lead to new insights on the functions of emotions and would suggest new avenue for research on their neural bases.

References

Adolphs R (2004) Could a robot have emotions? Theoretical perspectives from social cognitive neuroscience. In: Who needs emotions? The robot meets the brain (Fellous J-M, Arbib MA, eds), p in press: Oxford University Press.

Aggleton JP, ed (1992) The Amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction. New York: Wiley-Liss.

Arbib MA (1992) Book Review: The Cognitive Structures of Emotions (A. Ortony, G.L. Clore and A. Collins). Artificial Intelligence 54:229-240.

Arkin RC (2004) Moving up the food chain: Motivation and Emotion in behavior-based robots. In: Who needs emotions? The brain meets the robots (Fellous J-M, Arbib MA, eds), p in press: Oxford University Press.

Averill JR (1994) Emotions are many splendored things. In: The nature of emotion: fundamental questions (Ekman P, Davidson R, eds), pp 99-102. Oxford: Oxford University Press.

Boehning D, Snyder SH (2003) Novel neural modulators. Annu Rev Neurosci 26:105-131.

Bonabeau E, Dorigo M, Theraulaz G (1999) Swarm Intelligence: From Natural to Artificial Systems. Oxford: Oxford university press.

Brothers L (2001) Friday's footprint: How society shapes the human mind: London, Oxford University Press.

Carter CS, Altemus M, Chrousos GP (2001) Neuroendocrine and emotional changes in the post-partum period. Prog Brain Res 133:241-249.

Clark LA, Watson D (1994) Distinguishing functional from dysfunctional affective responses. In: The nature of emotion: fundamental questions (Ekman P, Davidson RJ, eds), pp 131-136. Oxford: Oxford university press.

Crick F, Koch C (2003) A framework for consciousness. Nat Neurosci 6:119-126.

Damasio A (1999) The Feeling of What Happens : Body and Emotion in the Making of Consciousness: Harcourt Brace.

Damasio AR (1994) Descartes'error: Emotion, reason and the human brain: Putnam.

Darwin C (1872) The Expression of Emotions in Man and Animals. London: Julian Friedman.

Davidson RJ (2003a) Affective neuroscience and psychophysiology: Toward a synthesis. Psychophysiology 40:655-665.

Davidson RJ, ed (2003b) Handbook of affective sciences: London, Oxford University Press.

Dennett DC (1991) Consciousness explained. New York, NY: Penguin.

Dennett DC (1997) Consciousness in human and robot minds: London, Oxford University Press.

Ekman P (1992) Are there basic emotions? Psychol Rev 99:550-553.

Ekman P (1994) All emotions are basic. In: The nature of emotion: fundamental questions (Ekman P, Davidson R, eds), pp 15-19. Oxford: Oxford University Press.

Fellous J-M (1999) The neuromodulatory basis of emotion. The neuroscientist 5:283-294.

Fellous J-M, Arbib MA, eds (2003) Who Needs Emotions? The brain meets the robot: Oxford University Press.

Hasselmo ME (1995) Neuromodulation and cortical function: modeling the physiological basis of behavior. Behav Brain Res 67:1-27.

Insel TR (1997) A neurobiological basis of social attachment. Am J Psychiatry 154:726-735.

Iversen L (2003) Cannabis and the brain. Brain 126:1252-1270.

Jeannerod M (2004) How do we decifer other's mind. In: Who needs emotion? The brain meets the robot (Fellous J-M, Arbib MA, eds), p in press: Oxford University Press.

Kaczmarek LK, Levitan IB (1987) Neuromodulation: The biochemical control of neuronal excitability. New York: Oxford University Press.

Lane RD (2000) Neural correlates of conscious emotional experience. In: Cognitive neuroscience of emotion (Lane RD, Nadel L, eds), pp 345-370: London, Oxford University Press.

LeDoux J (1996) The Emotional Brain. New York: Simon & Schuster.

LeDoux J, E. (2002) Synaptic self: How our brains become who we are. Harmondsworth, Middlesex, England: Penguin Books.

Levenson RW (1994) Human Emotion: A functional view. In: The nature of emotion: fundamental questions (Ekman P, Davidson R, eds), pp 123-126. Oxford: Oxford University Press.

Neumann ID (2003) Brain mechanisms underlying emotional alterations in the peripartum period in rats. Depress Anxiety 17:111-121.

Ortony A (2002) On making believable emotional agents believable. In: Emotions in humans and artifacts (Trappl R, Petta P, Payr S, eds), p 189. Cambridge, MA: MIT Press.

Ortony A, Clore GL, Collins A (1988) The Cognitive Structure of Emotions. Cambridge: Cambridge University Press.

Ortony A, Norman DA, Revelle W (2004) Affect and proto-affect in effective functioning. In: Who needs emotions? The brain meets the robot (Fellous J-M, Arbib MA, eds), p in press: Oxford University Press.

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

Picard RW (2002) What does it mean for a computer to "have" emotions? In: Emotions in Humans and Artifacts (Trappl R, P. PP, Payr S, eds), pp 213-236. Cambridge, MA: MIT Press.

Rees G, Kreiman G, Koch C (2002) Neural correlates of consciousness in humans. Nat Rev Neurosci 3:261-270.

Rolls ET (1998) The Brain and Emotion: Oxford University Press.

Scherer KR (1994) Emotion serves to decouple stimulus and response. In: The nature of emotion: fundamental questions (Ekman P, Davidson R, eds), pp 127-130. Oxford: Oxford University Press.

Schwarz N, Clore GL (1983) Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. Journal of Personality and Social Psychology 45:513 - 523.

Simon HA (1967) Motivational and Emotional Controls of Cognition. Psychological Reviews 74:29-39.

Sjostrom PJ, Turrigiano GG, Nelson SB (2003) Neocortical LTD via coincident activation of presynaptic NMDA and cannabinoid receptors. Neuron 39:641-654.