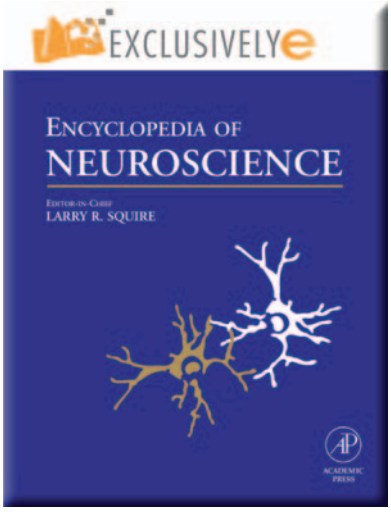

Fellous J -M (2009) Emotion: Computational Modeling. In: Squire LR (ed.) Encyclopedia of Neuroscience, volume 3, pp. 909-913. Oxford: Academic Press.

Emotion: Computational Modeling

J-M Fellous, University of Arizona, Tucson, AZ, USA

Introduction

Understanding the neural basis of human emotions is difficult because emotions are, to a large extent, subjective and nondeterministic. The same stimulus may create different emotions in different individuals, and the same individual may express different emotions in response to the same stimulus, at different times. However, in spite of this variability, it is assumed that there are basic principles, perhaps even basic neural mechanisms, that make a particular event 'emotional.' To find these principles and their underlying mechanisms, neuroscientists typically study specific emotions, using specific tasks. As is appropriate, they use a combination of animal and human preparations, yielding various types of data, from single neuron firing patterns, to activation levels of a whole brain area. The approach, while rigorous, is slow and yields an increasingly complex body of often conflicting data. An integrative approach is needed. Computational models of emotion have emerged as a promising tool for integration. Because they require that all assumptions be made explicit, they offer a new language in which to express and test hypotheses, as well as explain and predict neural mechanisms.

It is paradoxically the advances in cognitive neuroscience that have allowed for the emergence of emotion research. The paradox is in the fact that usually, the 'cognitive' is opposed to the 'affective.' We now know from experimental and modeling work that both cognition and emotion develop and work hand in hand. Fortunately, as a result of the cognitive neuroscience push to understand emotion, the techniques used in that field have carried over to emotion research. Models once used to understand cognitive functions, such as perception or memory, are now extended and sometimes fundamentally modified, to address issues pertaining to emotion.

The past 10 years has seen an explosion of artificial intelligence (AI) models that have directly addressed issues on the role of emotion. Though generally not neurally motivated, these models have emphasized several important functional aspects of emotions, such as their role in communication (human–machine interface), in decision making, in general body homeostasis, and in environmental adaptation (in autonomous robots). Several books have been published, one of the first being that of Rosalind Picard, *Affective Computing*. Artificial intelligence focuses on basic principles, without necessarily the need to find neural mechanisms that support them (e.g., chess playing). Neural modeling, on the other hand, focuses on mechanisms and hopes to emerge basic computational principles. The elucidation of the underlying mechanisms is powerful in that it explains how a system works and fails, and predicts the behavior of the system under conditions not yet tested experimentally. This aspect of modeling is particularly interesting for the understanding and treatment of emotional disorders; the focus here is on such neural modeling approaches, leaving AI approaches for further reading.

Building neurally motivated models of emotions or specific aspects of emotion processing can be achieved at many levels of abstractions. At the one end of the spectrum are models that implement a specific neural pathway, involving specific neural populations. These biophysical models typically describe neurons at the level of individual action potentials. Their explanatory and predictive power comes from the emergence of interesting dynamical network behaviors (e.g., oscillation, competition between neuronal pools) on millisecond timescales. At the other end are models that are mainly concerned with neural processing principles, where emphasis is brought on the connectivity and level of activation of typical neurons, or groups of neurons. They are primarily designed on the basis of anatomical and imaging data. These connectionist models can effectively give a larger view of the computations achieved by a set of brain areas. The vast majority of connectionist models implement some form of learning, whereby interesting network behavior emerges from the reconfiguration of the synaptic connection patterns. They can often explain and predict neural activity on behavioral time scales. There is of course a wide range of hybrid models in between those two extremes, as will be seen in the following discussion.

There has been a flurry of experimental work producing data on various aspects of the neuroscience of emotions in the past few decades. These data have given rise to many anatomical models, and have been reviewed by Tim Dalgleish. However, computational models of emotions have been possible because of the emergence of complementary mechanistic theories of emotions. Four of those conceptual models (as seen in the work of Joseph LeDoux, Edmund Rolls, Antonio Damasio, and Klaus Sherer) are at the basis of several neurally inspired models of emotions. Negative emotions are approached through the modeling of various aspects of fear processing in the amygdala. Positive

emotions are addressed through the study the neural substrate of reward processing and decision making.

Fear and the Amygdala

According to LeDoux's conceptual model of auditory fear conditioning, the amygdala neurons receive fast and broadly tuned inputs from the thalamus, and slower and more refined inputs from the cortex. Armony and co-workers built a computational model of auditory fear processing whereby synaptic connections were modified so that each amygdala neuron 'learned' to represent a restricted range of tone frequencies; together, the neurons cover the audible spectrum. Interestingly, when the learning was conducted under conditions in which fear was associated with certain tones but not with others, a subset of neurons retuned and shifted their preference to those frequencies that were paired with fear elicitation. This fear-driven remapping of neural tone preference was later confirmed experimentally by recording from the amygdala of animals trained to associate tones and fearful electrical footshocks. Surprisingly, the model also showed that the removal of cortical inputs to the amygdala cells did not affect their stimulus preference, as would be expected from computational units that have access to less incoming information. This result, again confirmed experimentally by lesion studies, predicted that noncortical inputs (thalamic in origin) may have more influence than previously expected. The model, constrained by experimental data, was therefore able to make experimentally testable, and later confirmed, predictions about fear processing, and lent support to the notion of dual routes of fearful stimulus processing.

Other more abstract models using genetic algorithms and artificial neural networks showed how these fast and slow routes of processing could emerge from the necessity for survival when discrimination between predator and prey (i.e., positive and negative 'emotions') was made difficult by environmental circumstances.

Beyond its involvement in fear, the amygdala has also been modeled as a component of the general motivational system. One of Stephen Grossberg's models was designed to explore the cognitive–emotional interactions at play during various learning paradigms. This model included a sensory (visual) perceptual system that consisted of the thalamus, primary visual cortices, and temporal lobes, a motivational system that represented the activity of the hypothalamus and amygdala, and a generic motor system that included overall cortical and cerebellar motor outputs. The amygdala acted as a motivational bias to the sensory stream, and through its interaction with the prefrontal cortex influenced action selection. The activity of the amygdala relied on the dynamical interplay between opponent emotional representations such as fear and relief. With the proper tuning, arousal-like inputs to these representations could lead to a characteristic inverted U-shaped behavioral response, whereby too little or too much arousal decreased motor output. Mechanistically, the model predicted that a large arousal input may yield a depression of amygdala activity that in turn caused a decrease in prefrontal cortex activation (hypofrontality). This chain of event lead to the emergence of negative symptoms characteristic of schizophrenia, such as flat affect (modeled as low amygdala activation), decrease of motivation (modeled as low prefrontal activity), and attentional deficits (modeled as a lack of prefrontal cortex control on sensorimotor information flow).

The idea of emotion as a component of the motivational system, and of emotional behavior as a result of the competition between different motivational systems, is also at the basis of a connectionist model of anxiety in rats. In the Salum model, spatial exploration was the result of a dynamical conflict between exploratory behavior (curiosity) and avoidance for open/high places (anxiety generating). While based on experimental behavioral data obtained in rats, the model did not, however, explicitly address the role of the amygdala, but rather, as in Grossberg's model, illustrated the basic principle of dynamical competition between conflicting motivational biases.

Other connectionist-like models by Taylor and Fragopanagos have attempted to clarify the link between emotion and attention. Their models relied on a fast, preattentive, processing route inspired by LeDoux's conceptual model. The amygdala evaluated the negative nature of incoming stimuli, and exerted a gain enhancing, excitatory effect on sensory processing which moved the focus of attention on the emotional stimulus (bottom-up attentional modulation). Alternatively, and perhaps complementarily, the focus of attention could also be modified by an amygdala-triggered interaction between ventromedial and dorsolateral prefrontal cortices modules that eventually changed the active cognitive goal (top-down attentional modulation). This model was used to explain various emotion–attention experiments and confirmed the view that the amygdala may have a fast, subattentive modulatory role in attention through both the production of saliency and the cognitive control of behavior. This model, as with that of Grossberg, points to a crucial interaction between the amygdala and prefrontal cortices. This interaction

has been further studied in the context of reward processing.

Reward Processing and the Prefrontal Cortex

The involvement of the orbitofrontal cortex (OFC) in emotion relies on a large body of data on reinforcement learning. For Rolls, this type of learning is at the basis of the processing of emotions in general, and the orbitofrontal cortex is in charge of decoding the reinforcement value of the stimuli and the rapid associations and reassociations between the stimuli and reinforcers. His experiments have shown that normal OFC function mediated one-trial reversal of stimulus–reward association. This phenomenon is mechanistically problematic because its time course is incompatible with learning mechanisms such as typically proposed by connectionist models, whereby synaptic modifications occur after multiple stimulus–response associations (i.e., training). Deco and Rolls' model needed to account for fast time scales, and had therefore to rely on spiking (biophysical) neurons. The model was based on 'biased competition' between pools of explicitly simulated neurons. A particular stimulus–reward association was represented by a pool of actively spiking neurons, and, interestingly, a change in association was simply a change in which pool of neuron fired (and not a change in synaptic strength). The change was dynamically achieved by a small input bias coming from a specialized OFC group of neurons called 'rule-specific' neurons. These neurons were context dependent and allowed for the same object to be perceived as rewarding or punishing, depending on the task contingency (context). They responded to punishment or the nondelivery of an expected reward by delivering an inhibitory bias to the currently active neuronal pool (a particular stimulus–reward association). When inactivated, the pool became momentarily refractory, and a new pool (a new stimulus–reward association) emerged rapidly, in one trial. This model made a number of testable predictions, one of which was the existence of specialized pools of OFC neurons that fired when a particular object–value combination (e.g., rewarded square, but not punished square) was presented. Such neurons may be used for general emotional reactions to a variety of cognitive contexts. Some experimental evidence indeed points to the existence of such neurons.

The idea that emotional signals bias the activity of competing active circuits is also at the basis of Damasio's somatic marker hypothesis. According to this theory, the perception of ones' own action in response to a particular stimulus has the potential to activate emotional memories that were previously associated with that stimulus. These memories ('somatic markers') in turn bias further cognitive processing, decision making, and action selection mechanisms. Several brain areas are known to be involved in this type of emotional bias, and special attention has been drawn to the ventromedial prefrontal cortex (VMPFC). The VMPFC is part of the executive system of the brain whereby actions are selected and plans are made. Because of its connection to the amygdala, the VMPFC also contains associations (the 'somatic markers') between previous actions and their expected consequences ('emotional' in nature). In the Wagar and Thagard model, when a stimulus is perceived, several potential actions are activated in the VMPFC. Some of these actions are part of somatic markers, and activate amygdala-dependent representations that 'reenact' the bodily state predictive of the action's consequences. These unconscious representations (representing emotional valence) in turn bias decision-making processes. The model further proposes that this bias depends on the context (hippocampus activation) in which the stimulus is perceived, and that it is gated by the nucleus accumbens.

The nucleus accumbens receives inputs from the VMPFC, the amygdala, and the hippocampus, and is thought to be key to high-level cognitive processes and action selection. However, because the nucleus accumbens is generally inhibited by dopamine (from the ventral tegmental area), these inputs are not sufficient to make it fire. The model showed that firing occurred only if the amygdaloid inputs (appraisal of a stimulus) arrived first, followed by a VMPFC input (a possible action to perform), and provided that the hippocampus is active throughout (the correct contextual information). In a simulated 'Iowa gambling' task, the model correctly showed that in normal conditions, choices are made according to long-term predicted outcomes (VMPFC activity), while in VMPFC lesions conditions (as in the case of the famous Phineas Gage), choices are made instead on the basis of immediate outcomes (amygdala activity). In the specific case of positive and negative stimulus evaluation taken as different 'emotional states,' this model offered insights on the role of the nucleus accumbens in the integration of cognitive, contextual, and emotional information flow.

One major issue in both the Deco and Rolls and the Wagar and Thagard model is that learning is assumed. The connectivity patterns among the various components of the model are configured 'manually' and emphasis is brought on the dynamical aspects of the network. Frank and Claus attempted to explicitly model the learning phase, and their model focused on the interactions between OFC and the

dopaminergic system of the basal ganglia. They showed that the probability of reward/punishment was likely to be encoded by the basal ganglia. In contrast, the variations in the magnitude of reward and punishment were both encoded by the amygdala and actively maintained in the medial and lateral portions of the OFC, respectively. The model was tested on the Iowa gambling task, reversal learning, and devaluation paradigms. They explained various aspects of decision-making deficits in OFC-damaged patients, Parkinson's disease patients, and in drug abusers, and offered a partial explanation for individual differences in risk-averse and risk-seeking normal individuals.

Modeling Emotions

Mainly because of the availability of experimental data, the focus of neural models of emotions has been on fear and reward processing; there have been relatively few attempts at using neural networks to model emotions in general. One step in this direction was taken by Scherer and co-workers, on the theoretical basis that emotions may be understood as the result of a complex sequence of appraisals. According to Scherer, emotions have five major components: a cognitive component (rational evaluation of objects and events), a peripheral efference component (homeostasis, body regulation), a motivational component (preparation for action), a motor component (expression of emotion, communication to others), and a subjective 'feeling' component (monitoring of emotional expressions). Formulated in this framework, an emotion is the dynamical coactivation of these components in response to the evaluation (by, e.g., the amygdala) of an internal or external stimulus that is particularly important to the subject. The architecture of the model is that of a feed-forward network, with lateral connections. The stimulus activates a set of interacting representational units that are specialized in the recognition of predefined aspects of the stimulus, such as its novelty, intrinsic pleasantness, or outcome probability. Together these units determine the significance of the stimulus, and activate a second layer of interacting units that prepare the organism for action. These units assess the relevance of the action, its appropriateness with respects to a set of standards, its potential consequences on the world, and the potential for coping with those consequences.

Emotion is therefore a pattern of activity, an emergent property of the interaction of these units, rather than the weighted activation of some set of basic and independent emotion 'nodes,' or modules. Unfortunately, no implementation of this connectionist model has yet been performed, so while conceptually useful, the model has not quantitatively explained or predicted emotional phenomena. In particular, such a model would be useful to understand the appraisal basis of emotional disorders, as well as the extent to which differences in appraisal processes explain individual differences in emotional processing in normal individuals.

For Mischel and Shoda, the emotional system is indeed at the core of what defines personality, individual differences, and the way we interact with others. In their connectionist model, abstract units representing cognitive and affective knowledge are randomly interconnected and interact dynamically. In some conditions, the response of the network to different situations (activation of a subset of the units) displays the same variability as that of human behaviors when individuals are repeatedly presented with a fixed set of situations. In other conditions, the network shows response consistencies across different situations, as would also be observed in humans. According to their model, individual differences and variability of behavior reside in the different connections between cognitive and affective units, rather than depending on the exact nature of the units.

Conclusion

Computational models of emotion, being biophysical or connectionist, point to several important conclusions regarding the neural bases of emotions. First, it has become clear that there are no 'emotional centers' in the brain. Models show that individual modules (being neurons, populations of neurons, or brain areas) do not work in a sequential manner. They are part of complex recurrent networks, so much so that their activity intimately depends on the activity of many other modules, being emotional or not. No region is specifically dedicated to the generation of emotions in general, or of a specific emotion in particular. Furthermore, models show that the dichotomy between cognitive and emotional processing is probably too simple, if not fundamentally incorrect. Brain areas such as the prefrontal cortex or amygdala are inherent parts of the 'computing' networks of stimulus perception and decision making, yet also mediate 'emotional' functions such as stimulus evaluations and emotional expressions.

Second, emotions are not simple 'states,' or patterns of activations of a few brain areas. While previous stimulus–emotion–action associations are a large part of what triggers an emotional behavior, it has become clear that emotions are dynamical modes of functioning. They temporally wax and wane as internal (e.g.,

memory driven) or external (e.g., perceptions) events occur. They develop over short timescales, as in fear, or over long timescales, as in moods or depression. As such, perhaps it is more useful to think about emotional flow, rather than emotional states.

Third, emotions have a functional role, and are not simple by-product of neural activity. The current computational models of emotions have so far focused on the notion of computing bias, whereby information flow is redirected or modulated in various ways. This view is in contrast with the notion of a parallel stream of 'emotional computations' that would simply be added as an additional information channel. The functional aspects of emotion are also clearly apparent in the AI literature.

In sum, the convergence of conceptual, experimental, and computational models seems to be key to the advancement of our understanding of the various aspects of emotional processing, emotional expression, and emotional experience.

See also: Amygdala: Contributions to Fear; Aversive Emotions: Molecular Basis of Unconditioned Fear; Emotion Systems and the Brain; Emotion: Neuroimaging; Emotional Learning in Humans; Emotional Control of the Autonomic Nervous System; Learning, Action, Inference and Neuromodulation; Pharmacology of Fear Extinction; Prefrontal Contributions to Reward Encoding; Reward Neurophysiology and Orbitofrontal Cortex.

Further Reading

Armony JL, Servan-Schreiber D, Cohen JD, et al. (1997) Computational modeling of emotion: Explorations through the anatomy and physiology of fear conditioning. *Trends in Cognitive Sciences* 1: 28–34.

Armony JL, Servan-Schreiber D, Romanski LM, et al. (1997) Stimulus generalization of fear responses: Effects of auditory cortex lesions in a computational model and in rats. *Cerebral Cortex* 7: 157–165.

Dalgleish T (2004) The emotional brain. *Nature Reviews Neuroscience* 5: 583–589.

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

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

Deco G and Rolls ET (2005) Synaptic and spiking dynamics underlying reward reversal in the orbitofrontal cortex. *Cerebral Cortex* 15: 15–30.

den Dulk P, Heerebout BT, and Phaf RH (2003) A computational study into the evolution of dual-route dynamics for affective processing. *Journal of Cognitive Neuroscience* 15: 194–208.

Fellous J-M and Arbib MA (eds.) (2005) *Who Needs Emotions? The Brain Meets the Robot*. New York: Oxford University Press.

Frank MJ and Claus ED (2006) Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review* 113: 300–326.

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

Mischel W and Shoda Y (1998) Reconciling processing dynamics and personality dispositions. *Annual Review of Psychology* 49: 229–258.

Picard RW (1997) *Affective Computing*. Cambridge, MA: MIT Press.

Rolls ET (2005) *Emotion Explained*. New York: Oxford University Press.

Salum C, Morato S, and Roque-da-Silva AC (2000) Anxiety-like behavior in rats: A computational model. *Neural Networks* 13: 21–29.

Sander D, Grandjean D, and Scherer KR (2005) A systems approach to appraisal mechanisms in emotion. *Neural Networks* 18: 317–352.

Taylor JG, Scherer K, and Cowie R (2005) Emotion and brain: Understanding emotions and modelling their recognition – introduction to special issue. *Neural Networks* 18: 313–316.

Trappl R, Petta P, and Payr S (eds.) (2002) *Emotions in Humans and Artifacts*. Cambridge, MA: MIT Press.

Wagar BM and Thagard P (2004) Spiking Phineas Gage: A neurocomputational theory of cognitive–affective integration in decision making. *Psychological Review* 111: 67–79.