# EMOTIONAL CONSCIOUSNESS:
# A NEURAL MODEL OF HOW COGNITIVE APPRAISAL AND SOMATIC PERCEPTION INTERACT TO PRODUCE QUALITATIVE EXPERIENCE

*Paul Thagard and Brandon Aubie*
*University of Waterloo*

Thagard, P., & Aubie, B. (forthcoming). Emotional consciousness: A neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. *Consciousness and Cognition*.

**ABSTRACT:** This paper proposes a theory of how conscious emotional experience is produced by the brain as the result of many interacting brain areas coordinated in working memory. These brain areas integrate perceptions of bodily states of an organism with cognitive appraisals of its current situation. Emotions are neural processes that represent the overall cognitive and somatic state of the organism. Conscious experience arises when neural representations achieve high activation as part of working memory. This theory explains numerous phenomena concerning emotional consciousness, including differentiation, integration, intensity, valence, and change.

## 1. INTRODUCTION

Everyone has experienced emotions such as happiness, sadness, fear, anger, pride, embarrassment, and envy. Dramatic progress has been made in understanding the neural mechanisms that underlie emotions, including the contribution of brain areas such as the amygdala and insula. Although many psychologists, neuroscientists, and philosophers have observed that conscious experience is an important aspect of emotion, no one has proposed a detailed, general theory of emotional consciousness. This paper provides an account of how conscious emotional experience emerges in the brain as the result of many interacting brain areas coordinated through working memory. It sketches a model of how emotions arise from a combination of neural representation, somatic perception, cognitive appraisal, and working memory.

First we review the range of phenomena that a theory of emotional consciousness needs to be able to explain. These include the broad range of different emotions, the varying intensity of emotions, the positive/negative character of emotions, and the beginnings and ends of emotional experience. We then summarize the crucial cognitive and physiological components needed to construct a theory of emotional consciousness, including representation, sensory processes, cognitive appraisal, and working memory. The best hope of integrating these diverse elements is by a neurocomputational account that shows how populations of neurons organized into identifiable brain areas with sensory inputs can generate high-level representations in working memory that constitute different emotional experiences. Building on recent models of decision making and parallel constraint satisfaction, we outline an integrated model of emotion in the brain that includes an account of working memory. We then show how the model explains a wide range of crucial phenomena of emotional consciousness. In order to elucidate the mechanism of emotional appraisal that we propose as an important part of emotional consciousness, we describe two new computational models: one shows how emotional appraisal can be construed as a kind of coherence computed by parallel constraint satisfaction, and the other shows how such computations can be carried out in a neurologically realistic fashion. Finally, we discuss the relevance of the theory and model for philosophical issues about the relation of mind and body.

Many discussions of the neuroscience of consciousness set themselves the task of discovering the “neural correlates” of conscious experience (Metzinger, 2000), but our aim is more ambitious. We will attempt to identify neural mechanisms that *cause* conscious experience, and will describe experimental manipulations that begin to justify such causal claims.

## 2. PHENOMENA TO BE EXPLAINED

The key phenomena that a theory of emotional consciousness should explain include differentiation, integration, intensity, valence, and change. Each of these aspects provides a set of explanation targets in the form of questions that a theory should answer. Answers should take the form of hypotheses concerning mechanisms that could produce the observed features of consciousness. A mechanism is a structure performing a function in virtue of the operations, interactions and organization of its component parts (Bechtel and Abrahamsen, 2005; see also Machamer, Darden, and Craver, 2000, and Thagard, 2006). Candidates for explaining emotional phenomena include: neural mechanisms in which the parts are neurons and the operations are electrical excitation and inhibition; biochemical mechanisms in which the parts are molecules and the operations are chemical reactions organized into functional pathways; and social mechanisms in which the parts are people and the operations are social interactions.

By *differentiation* we mean that people experience and distinguish a wide variety of emotions. The English language has hundreds of words for different emotions, ranging from the commonplace “happy” and “sad” to the more esoteric and extreme “euphoric” and “dejected” (Wordnet, 2005). Some emotions, such as happiness, sadness, fear, anger and disgust, seem to be universal across human cultures (Ekman, 2003), while others may vary with different languages and cultures (Wierzbicka, 1999). Some emotions such as fear and anger appear to be shared by non-human animals, whereas others such as shame, guilt and pride seem to depend on human social representations. A theory of emotional consciousness should be able to explain how each of these different experiences is generated by neural operations.

By *integration* we mean that emotions occur in interaction with other mental processes, including perception, memory, judgment, and inference. Many emotions are invoked by perceptual inputs, for example seeing a scary monster or smelling a favorite food. Perceptions stored in memory can also have strong emotional associations, for example the mental image of a sadistic third-grade teacher. Hence a theory of emotional consciousness needs to explain how perception and memory can produce emotional responses. Although there are diffuse, unfocussed moods such as contentment and anxiety, most emotions are directed toward objects or situations, as when you are happy that you got a raise or enjoy lasagna. A theory of emotional consciousness must therefore explain how we combine our awareness of an object with an associated emotion. Finally, a theory of emotional consciousness must account for how different interpretations of a situation can lead to very different emotional reactions to it, as when a tap on the shoulder is construed as an affectionate pat or an aggressive gesture.

A theory of emotional consciousness need not fully explain what it is like to feel happy or sad; as the concluding philosophical section discusses, this question is only partially answerable. But the theory should be able to explain ubiquitous aspects of conscious experience such as intensity and valence. The *intensity* of an emotional experience is its degree of arousal, which varies among different emotions. For example exuberance and elation involve much more arousal than plain happiness or even less intense contentment. Similarly, terror is more aroused than fear or anxiety. A theory of emotional consciousness should provide a mechanism for explaining such differences in intensity. It should also provide a mechanism for *valence*, the positive or negative character of emotions. Positive emotions like happiness and pride have very different qualitative feel from negative ones like fear, anger, and disgust. We need to identify the neural underpinnings of experiences with these different valences.

The last set of emotional phenomena that a theory of emotional consciousness should be able to explain concern *change*. Emotions are not constant: you can be feeling frustrated that your writing is going slowly, then shift to happiness when you hear on the radio that your favorite sports team has won. Emotional changes include shifts from one emotion to another as the result of shifts in attention to different objects or situations, but can also stem from a reinterpretation of a single object or situation, as when a person goes from feeling positive about a delicious food to feeling negative when its caloric consequences are appreciated. Emotional changes can also be more diffuse, as when a generally positive mood shifts to a more negative one as a frustrating day unfolds. Emotional changes can occur over long stretches of time, for example when people gradually change their attitude toward an object or state of affairs, or when therapy and medication help a depressed person to assume a more positive view of life. How might emotional change, valence, intensity, integration, and differentiation be explained?

## 3. ASPECTS OF A THEORY

Producing a theory of emotional consciousness is a daunting task, because it requires integrating controversial aspects of both emotions and consciousness. Putting critical discussion aside for the moment, here are some of the crucial ingredients. William James (1884) and others have claimed that emotions should be understood as a kind of perception of bodily states (see also Griffiths, 1997; Niedenthal et al., 2005;

Prinz, 2004) According to Damasio (1999), consciousness is an "inner sense" that is involved with wakefulness, attention, and emotion. He distinguishes between core consciousness and extended consciousness, which involves a sense of self. Core consciousness requires only an image, which is a mental pattern in any of the sensory modalities, and an object such as a person or other entity. He hypothesizes: “Core consciousness occurs when the brain’s representation devices generate an imaged, nonverbal account of how the organism’s own state is affected by the organism’s processing of an object, and when this process enhances the image of the causative object, thus placing it saliently in a spatial and temporal context.” (Damasio, 1999, p. 169). Extended consciousness requires memory that makes possible an autobiographical self. Feeling an emotion “consists of having mental images arising from the neural patterns which represent the changes in body and brain that make up the emotion” (p. 280).

Other emotion theorists have emphasized the cognitive rather than the somatic side of emotions. They contend that emotions are more like judgments than perceptions and arise from appraisal of a person’s general state (e.g. Clore and Ortony, 2000; Nussbaum, 2001; Oatley, 1992; Ortony, Clore, and Collins, 1988; Scherer, Schorr, and Johnstone, 2001). Our view is that the somatic and cognitive theories of emotion are in fact compatible, with each being part of the generation of emotions and hence of emotional consciousness. Rolls (2005, pp. 26-29) reviews several kinds of evidence against the view that emotions are just perceptions of bodily states. Barrett (2006) summarizes a large body of research that finds only weak correlations of emotions such as anger and fear with physiological properties such as facial behavior and autonomic arousal. Hence emotions cannot be understood merely as somatic perception. The neurocomputational theory sketched below shows how bodily perceptions and cognitive appraisals can be integrated.

Philosophers such as Lycan (1996) and Carruthers (2000) have argued that consciousness involves representations, but differ in whether the representations are like perceptions or like thoughts about mental states. The neurocomputational theory of consciousness sketched below shows how emotional representations can combine perceptions and judgments. It also shows how emotions can involve a representation of value, which is required for a theory of emotional consciousness (Seager, 2002).

Many cognitive psychologists have linked consciousness with working memory, which involves both short-term storage of different kinds of information and executive processes for manipulating the information. LeDoux (1996, p. 296) argues that "you *can't* have a conscious emotional feeling of being afraid without aspects of the emotional experience being represented in working memory." Neurocomputational models of working memory have been proposed using a variety of mechanisms such as recurrent excitation (Durstewitz, Seamons, and Sejnowkski, 2000). A neurocomputational theory of emotional consciousness should therefore have at least the following components: representation, somatic perception, cognitive appraisal, and working memory,

## 4. NEUROCOMPUTATIONAL THEORY: COMPONENTS

### Representation

We need a theory of neural representation sufficient to explain how the brain can represent the world, bodily states, and its own representations. A good start is the rich account developed by Eliasmith and Anderson (2003; see also Eliasmith, 2003, 2005).

They describe how a neural population (group of interconnected neurons) can represent features of the world by encoding them, that is by firing in patterns that are tuned to objects in the world in the sense that there are causal statistical dependencies between when the neurons fire and when our senses respond to the objects. Without going into the technical details, this kind of representation is sketched in figure 1(a), which shows the world having a causal effect on sensors such as eyes and ears, which produce neural signals that generate patterns of firing in neural populations. The neural population represents aspects of the world by virtue of the causal correlation between its firing patterns and what occurs in the world. See Appendix B for technical details.

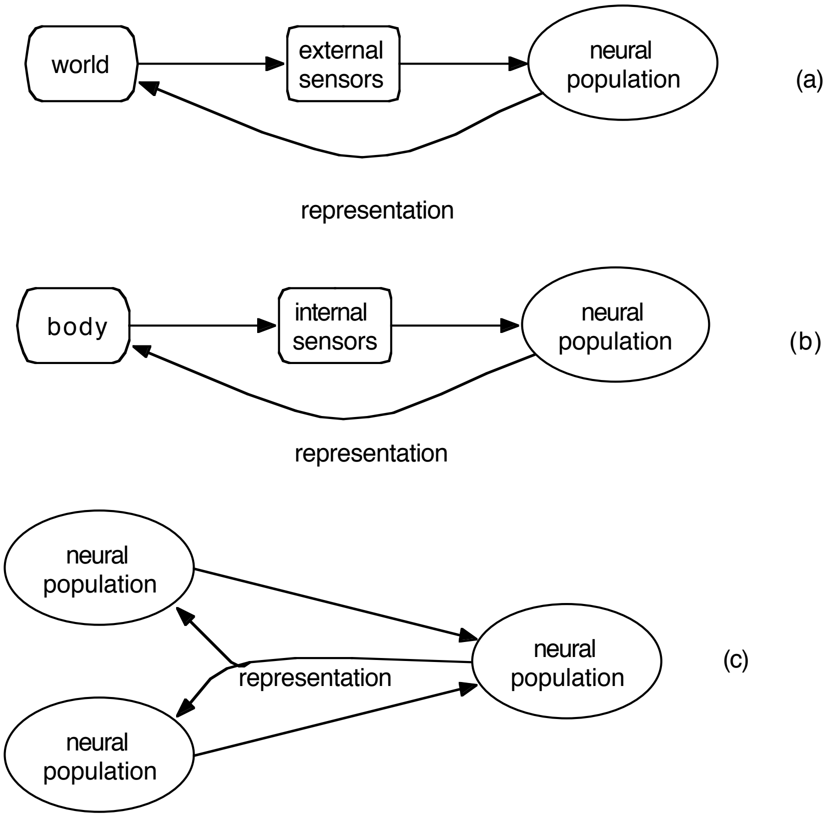

**Figure 1.** Representation by neural populations of (a) aspects of the world, (b) aspects of the body, and (c) other neural populations.

Similarly, neural populations can represent bodily states and events as shown in figure 1(b). Just as our bodies have external sensors such as eyes to detect what goes on the world, they have many internal sensors to detect their own states, including what is going on with crucial organs such as the heart and lungs, as well as concentrations of hormones and glucose in the bloodstream. Neural populations represent such states in the same way that they represent states of the world, by means of firing patterns that are tuned to particular occurrences via causal correlations. For example, there are neural populations in the hypothalamus that respond to blood glucose levels.

A brain that only encoded sensed aspects of the world and its own bodily states would be very limited in its range of inference and action. For more complex representations, neural populations need to respond to other neural populations, not just input from sensors, as in figure 1(c). The neural population on the right encodes aspects of the firing activity of the neural populations on the left by being tuned via statistical dependencies to their firing activities. Eliasmith and Anderson (2003) describe how neural populations can not only encode the representations of neural populations that affect them, but also transform these representations in complex ways. The result is that circuits of neural populations can produce representations of representations, making possible the kinds of higher-order thought that many philosophers have taken as an important aspect of consciousness.

When one neural population represents others, as in figure 1(c), it is not only because the firing of neurons in the input populations causes firing in the output population. The brain is full of feedback connections by which neural groups send signals back to groups that have sent them signals. Hence the correlation that develops between a neural population and other populations that it represents can be the result of causal influences that run in both directions.

**Emotional Decision Making**

Many brain areas contribute to human emotions, and an account of what they do and how they interact is crucial for a theory of emotional consciousness. Our starting point is a recent theory of emotional decision making proposed by Litt, Eliasmith, and Thagard (2006). According to this theory, human decision making has an emotional component that involves the interaction of at least seven major brain areas that contribute to evaluation of potential actions: the amygdala, orbitofrontal cortex, anterior cingulate cortex, dorsolateral prefrontal cortex, the ventral striatum, midbrain dopaminergic neurons, and serotonergic neurons centered in the dorsal raphe nucleus of the brainstem. How these regions interact has been modeled computationally by a system called ANDREA, for Affective Neuroscience of Decision through Reward-based Evaluation of Alternatives. ANDREA uses the neural engineering techniques developed by Eliasmith and Anderson (2003), and thus includes the representational capacities of neural populations described in the last section.

The connectivity structure of ANDREA is sketched in figure 2. The role of each of the indicated areas in emotion is known from a wide range of experimental studies in humans and other animals (see e.g. LeDoux, 1996; Panksepp, 1998; Rolls, 2005). The amygdala receives inputs from both external and internal sensors and modulates the intensity of positive and negative emotions, especially fear. The orbitofrontal cortex plays a central role in assessing the positive and negative valence of stimuli, and cooperates with the midbrain dopamine system and serotonergic system to compute the potential gains and losses of potential actions. The anterior cingulate cortex is involved in the detection of conflicts between current behavior and desired results. The dorsolateral prefrontal cortex contributes to the representation, planning, and selection of goal-related behaviors.

Evidence that ANDREA is a useful model of the neural mechanisms that underlie human decision making comes from its success in simulating two important classes of psychological experiments that previously had been accounted for by behavioral-level theories. ANDREA provides a detailed, quantitative model of decision phenomena described by the prospect theory of Kahneman and Tversky (2000) and the decision affect theory of Mellers, Schwartz, and Ritov (1999). Neural Affective Decision Theory and ANDREA by themselves say nothing explicitly about conscious experience, but we will describe natural extensions that provide the additional ingredients needed to account for consciousness.

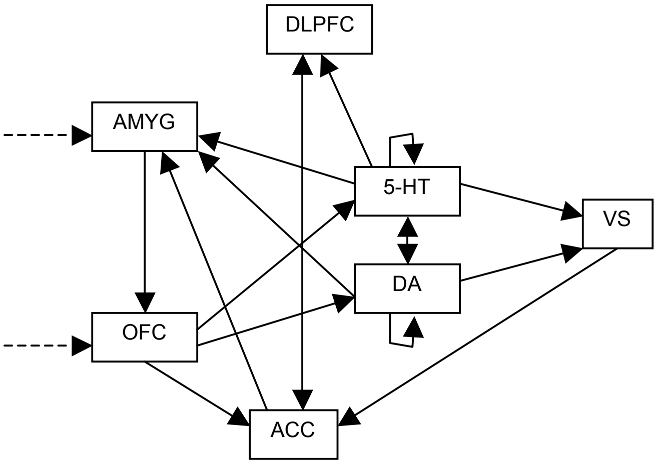

**Figure 2.** The ANDREA model of decision evaluation, from Litt, Eliasmith, and Thagard (2006). Dotted arrows represent external inputs to the model. Abbreviations: 5-HT, dorsal raphe serotonergic neurons; ACC, anterior cingulate cortex; AMYG, amygdala; DA, midbrain

dopaminergic neurons; DLPFC, dorsolateral prefrontal cortex; OFC, orbitofrontal cortex; VS, ventral striatum.

Additional brain areas relevant to decision making are part of the GAGE model of decision developed earlier by Wagar and Thagard (2004) to explain the historical case of Phineas Gage as well as the behavior of modern patients with damage to the ventromedial prefrontal cortex (VMPFC). This area is contiguous with the orbitofrontal cortex and is important for providing connections between the cortex and the amygdala. The GAGE model also includes the hippocampus, which is important for modeling the effects of memory and context on decision making. Wagar and Thagard (2004) used GAGE to simulate the behavior of people in the Iowa gambling task of Bechara et al. (1997), as well in the famous experiments of Schacter and Singer (1962). Thus the VMPFC and hippocampus should be added to the seven areas included in the ANDREA model as part of a fuller account of human emotion.

At least two other brain areas, the thalamus and the insula, appear important for emotional consciousness because of their connections to external and internal sensors (Morris, 2002). The thalamus receives many kinds of external sensory inputs and sends signals to the amygdala and the cortex. The insula cortex receives somatic information and passes it along to other cortical areas (Damasio, 1999). Thus there are numerous interacting brain areas that need to be included in a full theory of emotional consciousness.

## Inference and Appraisal

Still missing from our account of emotional experience is an explanation of how cognitive appraisal is performed. How and where does the brain assess its overall current state, making use of perceptual and somatic information as well as its accumulated knowledge? Such appraisal requires more complex inference than feedforward representation of sensory inputs and their representation. Nerb (forthcoming) presents a computational model of emotions that shows how appraisal can be construed as a kind of parallel constraint satisfaction accomplished by artificial neural networks using localist representations, that is with emotions and goal-relevant elements represented by single artificial neurons. We describe our own model of appraisal in section 7. For greater biological plausibility, it would be better if cognitive appraisal were performed by distributed representations in which single elements are represented by activity in many neurons and in which individual neurons participate in many representations. Section 8 describes how this might work.

## Working Memory

Yet another component is needed for a broad, plausible account of emotional consciousness. Like other kinds of consciousness, emotional experience has a serial character very different from the neural activities that go on simultaneously and asynchronously in many brain areas. Cognitive psychologists such as Smith and Jonides (1999) have described how working memory involves both short-term storage of different kinds of information in different brain areas and executive processes of selective attention and task management that involve the anterior cingulate and dorsolateral prefrontal cortex. Eliasmith and Anderson (2003) describe how working memory can be modeled by transformations of neural representations, and there are other possible neurocomputational models of working memory. Section 8 describes a model that uses working memory to provide a binding between cognitive and affective representations.

## 5. THE EMOCON MODEL

The task now is to combine the many neural components and mechanisms discussed in the last section into an integrated mechanism capable of explaining a broad range of phenomena of emotional consciousness. Figure 3 sketches a model of the integrated mechanism, EMOCON, that incorporates ideas from the ANDREA, GAGE, and NECO models, along with the observations of Damasio (1999) and Morris (2002) about sensory inputs. We conjecture that emotional experience is the result of interactions among *all* the components shown in figure 3. We have not yet programmed such a large and complicated simulation, but we will extrapolate from the parts that are functioning in simpler models to offer explanations of emotional experience.

Notice how the EMOCON model in figure 3 combines all the aspects of emotion and consciousness specified earlier. It includes neural representations of the world, of the body, and of other neural representations. It has the most important brain areas known to be involved in positive and negative bodily responses to stimuli, and also includes the the dorsolateral prefrontal cortex which is capable of complex inferences about the social significance of a wide range of information. Not shown for reasons of complexity are the hippocampus which is part of the GAGE model and the serotonergic system for negative rewards which is part of the ANDREA model. The ventral striatum from the ANDREA model (including the nucleus accumbens from the GAGE model) is included as part of the dopamine system. Many interconnections between brain areas are not shown. Working memory is largely associated with activity in the dorsolateral prefrontal cortex and the anterior cingulate.

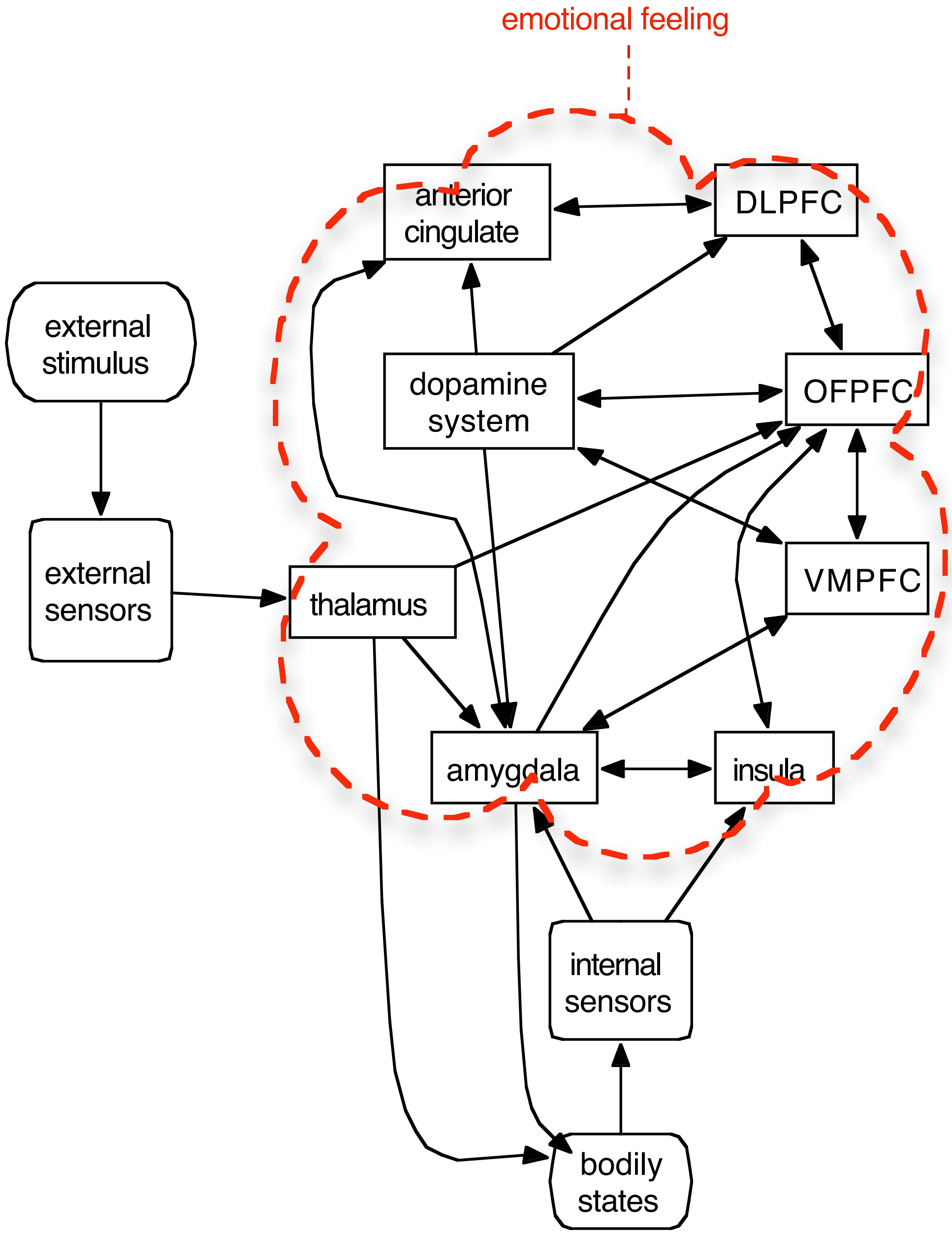

**Figure 3.** The EMOCON model of emotional consciousness, incorporating aspects of Litt, Eliasmith, and Thagard (2006), Wagar and Thagard (2004), and Morris (2002). Abbreviations are PFC for prefrontal cortex, DL for dorsolateral, OF for orbitofrontal, and VM for

ventromedial. See text for explanation of the dotted line representing emotional experience.

So what is an emotion? It is not just a perception of bodily states, nor is it just a cognitive appraisal of one's overall situation. Rather, an emotion is a pattern of neural activity in the *whole* system shown in figure 3, including inputs from bodily states and external senses. The EMOCON model shows how to combine somatic perception and cognitive appraisal into a single system that transcends the century-old conflict between physiological and cognitive theories of emotions.

Note the presence in the diagram of numerous feedback loops (also called recurrent or reentry connections), for example between the amygdala, bodily states, and internal sensors. Emotional consciousness is *not* represented as an output from any of the brain areas or their combination. Rather, the shadowy dotted line signifies that emotional consciousness *just is* the overall neural process that takes place in the interacting brain areas. The section on philosophical issues below will discuss the legitimacy of identifying emotional experience with neural activity.

## 6. EXPLANATIONS

As section 2 outlined, a theory of emotional consciousness should be able to explain many properties of emotional experience, including valence, intensity, change, differentiation, and integration.

### Valence

Emotion researchers such as Russell (2003) have recognized that emotions vary along two major dimensions: valence, which is the character of being positive/negative or pleasurable/unpleasurable, and intensity, which is the degree of arousal involved in the emotional experience. Emotions can be located along these two dimensions, as shown in figure 4.

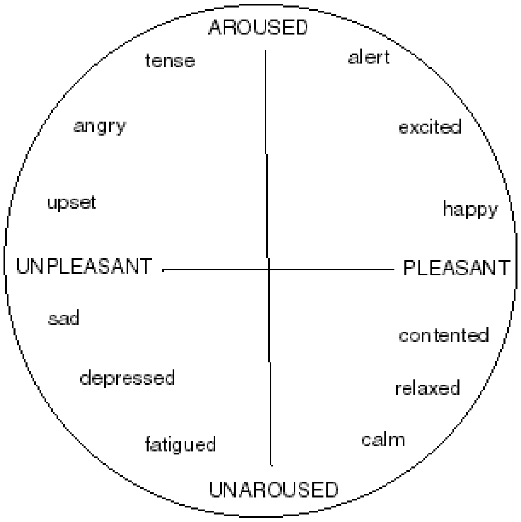

**Figure 4.** The structure of emotions with respect to pleasantness and intensity. Reprinted from Thagard (2005), p. 165.

EMOCON model in figure 3 explains how states can have positive or negative valence. Positive valence is known to be associated with a complex of neural states, including increased activation of the dopamine system and the left prefrontal cortex (see e.g. Damasio et al., 2000; Davidson, 2004; Dolcos, LeBar, and Cabeza, 2004; Prohovnik et al., 2004). Negative valence is associated with increased activation of the dorsal raphe serotonergic neurons and the right prefrontal cortex. The negatively-valenced emotion of disgust correlates with activity in the insula which has neural populations that represent visceral states. According to Hamann et al. (2002), the left amygdala and ventromedial prefrontal cortex were activated during positive emotion, whereas negative emotion was associated with bilateral amygdala activation.

The studies cited in the last paragraph establish neural correlates of positive and negative valence, but do not in themselves show that a causal mechanism has been identified. Such brain activity may correlate with emotions because emotions are the cause rather than the effect of the brain activity, as a dualist would maintain. Or perhaps there is some common cause of both emotion and brain activity. The standard way of distinguishing causation from correlation is intervention: to find out whether A causes B, manipulate A and examine the effect on B. Can we show that manipulating the brain changes emotional experience?

People perform such manipulations whenever they have a beer, cigarette, or line of cocaine. Alcohol, nicotine, and many recreational drugs such as ecstasy and cocaine increase dopamine levels, leading temporarily to increased positive valence attached to various representations. Depletion of dopamine through repeated use of such drugs leads to decreased positive valence. Depression can be treated by transcranial magnetic stimulation of the left prefrontal cortex: intense electromagnetic radiation outside the skull increases brain activity not only in left prefrontal cortex (which we saw is associated with positive valence) but also in dopamine areas (Gershon, Dannon, and Greenhaus, 2003). Deep brain stimulation can be used to treat severe depression by modulating activity in a region, the subgenual cingulate gyrus, known to be overactive in depressed people (Mayberg et al, 2005). Such experiments involving changes in valence justify the claim that brain activity causes emotional experience in addition to correlating with it.

**Intensity**

The most natural explanation of difference in intensity between emotional states with the same valence, for example being happy and being elated, is in terms of firing rates in the relevant neural populations. For extreme happiness, we would expect more rapid firing of more neurons in regions associated with positive valence such as the dopamine areas and the left prefrontal cortex than would occur with moderate happiness. However, it is difficult to test this prediction because of ethical limitations on research on humans using single-cell recordings, and because of limitations in the resolution of brain scanning techniques such as fMRI and PET.

Anderson et al. (2003) discuss the difficulty of disassociating intensity and valence in studies of brain activity. They ingeniously used odors to distinguish stimulus intensity from valence. Using fMRI, they found amygdala activation to be associated with intensity but not the valence of odor, whereas the orbitofrontal cortex is associated with valence independent of intensity. Dolcos, LaBar, and Cabeza (2004) found that dorsolateral prefrontal cortex activity is sensitive to emotional arousal. Hence there is some evidence that brain activity is correlated with emotional intensity. As with valence, the effects of drugs suggest that the relation between brain and emotion is more than correlational. For example, amphetamines increase neural firing in dopamine circuits and thereby increase the intensity of some positive emotional experiences.

**Change**

Unlike moods, which can last for hours or days, emotions are relatively short in duration. Part of the explanation of the beginning of an emotional experience is obviously new external stimuli such as the television image of a team winning and the email from a co-author about a paper acceptance. But many external stimuli do not produce new emotions, so what gets an emotional experience going?

The key to understanding the onset and cessation of emotions is working memory, which is the part of long-term memory that is currently most active (Fuster, 2003). In neural terms, long term memory consists of particular neurons and their synaptic connections, and activity is degree of neural firing. Working memory consists of those neural populations that have a high firing rate. The model sketched in figure 3 shows how neural populations in the main brain areas implicated in working memory, anterior cingulate and DLPFC, can become activated as the result of an external stimulus. But working memory can also be generated indirectly by activation of the contents of long term memory through cognitive processes such as association and inference.

In neurocomputational terms, working memory has two crucial aspects: recurrent activation and decay. Recurrent (also called reentry) connections are ones that enable neural populations to stimulate themselves, so that the contents of working memory tend to stay active. However, working memory is also subject to decay, so that if there is no ongoing stimulation of its active contents from perception and memory, those representations will tend to drop out of working memory. Another mechanism in working memory is inhibition, in which the activation of some elements tends to suppress the activation of others, so that stimuli compete for conscious representation (Kastner and Ungerleider, 2005).

Stimulation, recurrence, decay, and inhibition explain how emotions can enter and leave working memory. Suppose you hear that your favorite soccer team has won the World Cup, which activates your long-term representation of the team and the Cup and generates the feeling of happiness through the complex feedback process shown in figure 3. As long as the neurons that represent the belief that your team won have a high hiring rate, you continue to feel happy about your team, because the feedback process involving cognitive appraisal and bodily states continues. But when new external stimuli come in, or associative memory shifts your thinking to another topic, then the combination of activation of new information and decay of existing neural representations reduces to below threshold the activation of the complex of neural populations in working memory that represent the content of the emotion and their associated bodily states. Thus the mechanism depicted in figure 3, including working memory, can explain the onset and cessation of emotional experience.

## Integration and Differentiation

The EMOCON model in figure 3 can easily handle integration, as it incorporates and ties together brain areas for external perception such as the thalamus, areas for somatic perception such as the insula, areas for evaluation of rewards such as the orbitofrontal prefrontal cortex and the dopamine system, and areas for high-level cognition such as the dorsolateral prefrontal cortex. Neural processing provides a common mechanism for combining low-level perception of the world and bodily states with high-level inferences about the extent to which an agent is accomplishing its goals in a particular environment.

All emotions involve positive or negative valence and different degrees of intensity, but these two dimensions are not sufficient to differentiate consciousness of a full range of emotions. For example, sadness and anger are both negative and can have intensity ranging from moderate to extreme, but people do not confuse them, even though the bodily states associated with them are fairly similar; for an attempt to pin down some physiological correlates of emotions, see Rainville et al. (2005). Hence cognitive appraisals are needed for fine discrimination of emotions, but such appraisals can be rapidly performed in parallel by a process of constraint satisfaction. Nerb and Spada

(2001) present a computational model of how anger may arise because an observed stimulus is associated with  damage, human agency, and controllability, whereas sadness is associated with a person's higher goals.  Nerb (forthcoming) describes a constraint satisfaction model that covers more emotions, which are not simply perceptions of bodily states, but require inferences about how a person's overall situation is related to external stimuli and internal states.

A similar constraint-satisfaction analysis could be given for more complex social emotions such as shame, guilt, and pride.   Figure 5 shows  a constraint network that could be used to model what social emotions someone might feel in different circumstances.  Depending on the combination of positive or negative valence (deriving in part from internal representations of bodily states) and cognitive representations of the overall situation, the overall state of working memory will vary along lines that people call by familiar names such as shame, guilt, and pride.  Figure 5 shows a highly simplified localist neural  network that crudely differentiates pride, shame, and guilt based on the satisfaction of family and moral goals in interaction with valence.   For example, the experience of pride in the accomplishment of one's children – what in Hebrew is called *nachus* – arises from the interaction of bodily states and appraisal of the extent to which one's family goals are being accomplished.

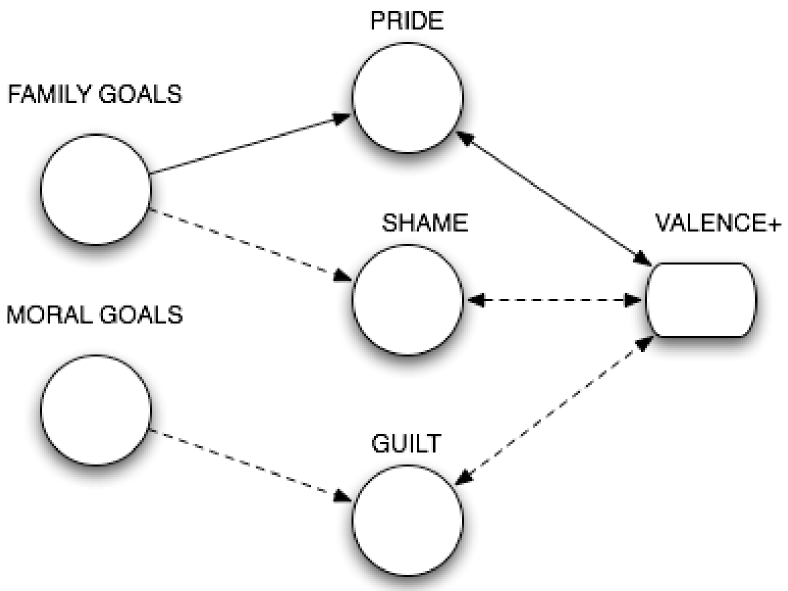

**Figure 5**. Simplified model of a constraint network for some social emotions. Solid lines are positive constraints, whereas negative lines are negative constraints. Positive valence is enhanced by pride arising from the satisfaction of family goals, whereas positive valence is suppressed by guilt arising from lack of satisfaction of moral goals.

Obviously, cognitive appraisal involves many more inputs and emotions than are shown in figure 5. In section 7 we describe a much more complete model of cognitive appraisal as constraint satisfaction. The network shown in figure 5 is also crude from a neurological perspective, in that it suggests that emotions can be represented by a single node rather than by activity in many neural populations across many brain areas as shown in figure 3. In section 8 we describe how parallel satisfaction of cognitive and affective constraints can be modeled in a highly distributed manner. Hence it is reasonable to think of the parallel constraint satisfaction processes for cognitive appraisal shown in figure 5 as being part of the overall computational process shown in figure 3. Thus the process of cognitive appraisal can go on in parallel with the process of representation of internal bodily states, producing differentiation of a wide range of emotional experiences.

## 7. A COHERENCE MODEL OF EMOTIONAL APPRAISAL

Although we lack the computational power required for a full implementation of the EMOCON model, we have been able to supplement the computationally realized GAGE and ANDREA models with new models that fill in some of the details of the overall process of emotion. First, to help justify our account of emotional appraisal as parallel constraint satisfaction, we describe a computational model that shows how appraisal involving many inputs and emotions can be understood in terms of coherence.

Emotional appraisal is the continual processes of updating and evaluating an array of internal states to both evoke and discriminate between emotions. Sander, Grandjean, and Scherer (2005) offer a psychologically plausible account of emotional appraisal that treats emotions as continuous temporal phenomena not based on discrete rules. Their Component Process Model uses a set of Sensory Evaluation Checks (SECS) to discriminate between emotions, as summarized in table 1. These SECS take into account both low-level and high-level cognitive representations used for distinguishing emotions.

Discrete rule based systems can model emotional appraisal from the SECs but they lack the natural ability of being implemented in neural network architectures. A rule of the form "IF *x* THEN *y*" can discretely recognize emotions but it fails to model any temporal aspects or succeed in partial recognition. Partial recognition could be modeled by a matching rule that returns a degree of matching, but this requires a template for comparison. Ideally a model of emotional appraisal should take advantage of natural neural network conditions, be sensitive to the temporal dimension of appraisal, and naturally allow for partial recognition.

EACO (Emotional Appraisal as Coherence) is a model of emotional appraisal that uses parallel constraint satisfaction in an artificial neural network to discriminate between the emotions presented in table 1. Our model is non-discrete in that emotions are given a real number activation level between 0 and 1 instead of a binary yes/no value. Furthermore, all SECs and emotion nodes interact either directly or indirectly via symmetric links (activation flows in both directions) over time.

EACO combines parallel constraint satisfaction models of coherence (Thagard, 1989, 2000) with the Component Process Model of emotional appraisal (Sander, Grandjean, and Scherer, 2005) to offer a computational model of emotional appraisal in the brain. To compute the appraisals shown in table 1, EACO requires 67 units and 262 links between them. For mathematical details, see Appendix A.

| Criterion | ENJ/HAP | ELA/JOY | DISP/DISG | CON/SCO | SAD/DEJ | DESPAIR | ANX/WOR |
|---|---|---|---|---|---|---|---|
| Relevance | | | | | | | |
| Novelty | | | | | | | |
| Suddenness | Low | High/med | Open | Open | Low | High | Low |
| Familiarity | Open | Open | Low | Open | Low | Very low | Open |
| Predictability | Medium | Low | Low | Open | Open | Low | Open |
| Intrinsic pleasantness | High | Open | Very Low | Open | Open | Open | Open |
| Goal/need relevance | Medium | High | Low | Low | High | High | Medium |
| Implication | | | | | | | |
| Cause: agent | Open | Open | Open | Other | Open | Other/nat | Other/nat |
| Cause: motive | Intent | Cha/int | Open | Intent | Cha/neg | Cha/neg | Open |
| Outcome probability | Very high | Very high | Very high | High | Very high | Very high | Medium |
| Discrepancy from expectation | Consonant | Open | Open | Open | Open | Dissonant | Open |
| Conduciveness | Conducive | Vcond | Open | Open | Obstruct | Obstruct | Obstruct |
| Urgency | Very low | Low | Medium | Low | Low | High | Medium |
| Coping potential | | | | | | | |
| Control | Open | Open | Open | High | Very low | Very low | Open |
| Power | Open | Open | Open | Low | Very low | Very low | Low |
| Adjustment | High | Medium | Open | High | Medium | Very low | Medium |
| Normative significance | | | | | | | |
| Internal Standards | Open | Open | Open | Very low | Open | Open | Open |
| External Standards | Open | Open | Open | Very low | Open | Open | Open |
| Criterion | FEAR | IRR/COA | RAG/HOA | BOR/IND | SHAME | GUILT | PRIDE |
| Relevance | | | | | | | |
| Novelty | | | | | | | |
| Suddenness | High | Low | High | Very low | Low | Open | Open |
| Familiarity | Low | Open | Low | High | Open | Open | Open |
| Predictability | Low | Medium | Low | Very high | Open | Open | Open |
| Intrinsic pleasantness | Low | Open | Open | Open | Open | Open | Open |
| Implication | | | | | | | |
| Cause: agent | Oth/nat | Open | Other | Open | Self | Self | Self |
| Cause: motive | Open | Int/neg | Intent | Open | Int/neg | Intent | Intent |
| Outcome probability | High | Very high | Very high | Very high | Very high | Very high | Very high |
| Discrepancy from expectation | Dissonant | Open | Dissonant | Consonant | Open | Open | Open |
| Conduciveness | Obstruct | Obstruct | Obstruct | Open | Open | High | High |
| Urgency | Very high | Medium | High | Low | High | Medium | Low |
| Coping potential | | | | | | | |
| Control | Open | High | High | Medium | Open | Open | Open |
| Power | Very low | Medium | High | Medium | Open | Open | Open |
| Adjustment | Low | High | High | High | Medium | Medium | High |
| Normative significance | | | | | | | |
| Internal Standards | Open | Open | Open | Open | Very low | Very low | Very high |
| External Standards | Open | Low | Low | Open | Open | Very low | High |

**Table 1.** Predicted appraisal patterns for some major emotions. ENJ/HAP, enjoyment/happiness; ELA/JOY, elation/joy; DISP/DISG, displeasure/disgust; CON/SCO, contempt/score; SAD/DEJ, sadness/dejection; IRR/COA, irritation/cold anger; RAG/HOA, rage/hot anger; BOR/IND, boredom/indifference; Other/nat, Other or Natural (e.g.weather); Cha/int, Chance or Intent; Cha/neg, Chance or Negative Intent. From Sander, Grandjean, and Scherer (2005), p. 326.

Our model of emotional appraisal has a network consisting of three layers of nodes. The first layer consists of a single *special* node that maintains an activation level of 1 at all times and is used to activate the particular *sensory evaluation checks* (SECs) used in picking out a particular emotion. The second layer consists of 52 nodes that represent the individual SECs and the different values they can assume. For example, a low familiarity of the situation and a high familiarity of the situation are each represented with a distinct node. A total of 16 SECs with various levels of gradation are represented. The third layer is composed of 14 emotional nodes that represent the activation level of 14 distinct emotions as characterized by Sander, Grandjean, and Scherer (2005). When the SECs corresponding to a particular emotion are activated, that emotion node is also activated by excitatory links between each layer. Each emotion node is connected to each other emotion node with an inhibitory link. This results in the strongest emotion gaining full activation and thereby suppressing other emotions. When two emotions have similar SECs, they may become co-activated such as is the case with happiness and pride, as in figure 6. Hence EACO can model the occurrence of mixed emotions, as well as all the individual emotions listed in table 1.

From a neural perspective, the major limitation of EACO is that it uses a localist representation, with an emotion or input represented by a single node. We lack the computational resources to produce a fully distributed version of the whole EACO model, but we will now show on a smaller scale how coherence can operate in more neurologically realistic networks.

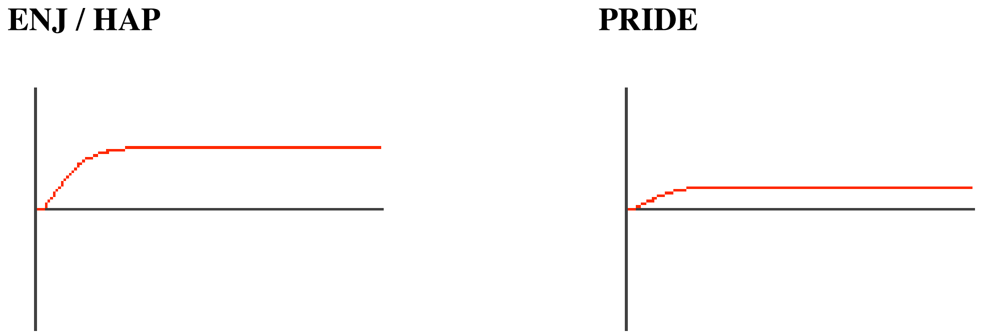

**Figure 6.** The activation levels of ENJ/HAP and PRIDE are shown for 150 time steps where the SECs corresponding to ENJ/HAP were activated. Even though the ENJ/HAP node is sending inhibitory activation levels to the PRIDE node, the SECs in common are enough to overcome this inhibition and activate PRIDE a small amount.

## 8. COHERENCE IN THE BRAIN

We now show how parallel constraint satisfaction can be accomplished in more realistic neural systems. We employ the Neural Engineering Framework of Eliasmith and Anderson (2003) to develop a model of parallel constraint satisfaction that reproduces the behavior of localist models such as the explanatory coherence account of Thagard (1989, 2000). Moreover, we show how biologically realistic neural networks can compute a kind of emotional coherence through interactions among multiple brain areas, including the prefrontal cortex and the amygdala. The result is a psychologically plausible and neurologically detailed account of how brains achieve coherence, showing how the coherence model of emotional appraisal presented in the last section can contribute to the overall picture of emotional consciousness given by the EMOCON model in figure 3.

**Explanatory Coherence**

The method we propose for neurologically realistic models of parallel constraint satisfaction is general, but we illustrate it with application to a particular kind of inference based on explanatory coherence. The problem of choosing among competing explanations arises in everyday life when people attribute causes to the behaviors of others, in law when detectives and jurors infer who is responsible for crimes, and in science when investigators evaluate theories that can explain experimental evidence. To take a simple example, suppose we have two pieces of evidence, E1 and E2, and two competing, incompatible hypotheses, H1 and H2, for explaining this evidence (figure 7). H1 is able to explain both pieces of evidence while H2 is only able to explain E1. In this case, H1 is superior to H2 since it can explain more than H2. Thagard (1989) showed how theory evaluation can be understood as a problem whose positive constraints are based on explanatory connections between hypotheses and evidence, and whose negative constraints are between competing hypotheses.

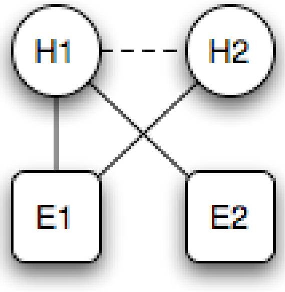

**Figure 7.** A simple explanatory coherence network where H1 is accepted over H2 because it explains more of the evidence. Solid lines are excitatory links and the dashed line is inhibitory.

Thagard's computational model of explanatory coherence, ECHO, works by calculating the activation levels between 1 and -1 of units representing propositions in a localist network. Evidence nodes are connected to a special unit that always sends an activation level of 1 to them. At each time step, the activation level of a node is updated as a function of the nodes connected to it via either excitatory or inhibitory links. After numerous updates, the network tends to stabilize with nodes being either positively activated or negatively activated. This stabilized network generally has the highest overall level of coherence, and hence has approximately solved the parallel constraint problem.

Table 2 summarizes how ECHO and similar models accomplish parallel constraint satisfaction: units stand for propositions; activation stands for their acceptance or rejection; excitatory and inhibitory links implement constraints; and spreading activation makes units interact until the network has stabilized and the constraints are maximally satisfied. The third column of table 2 sketches how parallel constraint satisfaction can be managed in a more biologically realistic distributed fashion, which we now describe.

## Explanatory Coherence by Neural Engineering

The Neural Engineering Framework of Eliasmith and Anderson (2003) provides a set of tools for building systems of neural populations that perform complex functions, and we have used it to produce a distributed version of ECHO. We call the resulting model NECO (for Neural-Engineering-COherence). Whereas ECHO uses a single unit to represent a proposition, NECO uses a neural population, currently with 1000 spiking neurons, to represent the acceptability levels of all propositions. These neurons are more biologically realistic than ECHO's units, which have a real number associated with them representing the acceptability of the unit, analogous to the spiking rate of a neuron construed as a proportion of the maximum rate a neuron can spike. In contrast, like neurons in the brain, NECO's artificial neurons have actual spiking (firing) behavior that must be decoded externally to be construed as acceptability levels. Maximum spiking rates are between 50 Hz (times per second) and 60 Hz, in line with observed spiking rates in the prefrontal cortex (Kawasaki et al., 2001).

| *Constraint satisfaction* | *Localist Model* | *Distributed Model* |
|---|---|---|
| Proposition or other representation | Unit: single artificial neuron | Population of artificial neurons |
| Positive and negative constraints | Excitatory and inhibitory links between units | Neurons exciting and inhibiting other neurons. |
| Acceptance/rejection | Activation of unit | Pattern of spiking in neural population |
| Constraint satisfaction | Spreading activation among units until network settles | Adjustment of spiking patterns until stable patterns reached |
| Valence | Valence of unit | Pattern of spiking in associated neural population |

**Table 2.** How constraint satisfaction can be performed by localist and distributed models. See below for discussion of valence.

In NECO, the neural population for one proposition overlaps with the neural populations of other propositions. The spiking behavior of members of a neural population together represent the acceptability of propositions between 1 (accept) and –1 (reject). This relation uses the neural population to stand for a real number vector space of *n* dimensions where *n* propositions are being represented. Each dimension represents the acceptability of one proposition. For example, in a 3-dimensional vector space, any point can be written as $(x,y,z)$ where $x$ is the value of the first dimension, $y$ is the value of the second dimension and $z$ is the value of the third dimension. NECO uses a single neural population to store multiple real numbers by having each neuron tuned such that its spiking rate correlates with a direction and point in that vector space. Our model randomly tunes each neuron to a direction and point at the start of each run, whereas neurons in the human prefrontal cortex are tuned through years of learning. For one dimension, each NECO neuron is tuned to begin spiking at a number between 1 and –1, as shown in figure 8. For example, curve A shows the tuning curve of a neuron that does not spike for acceptability 0.5, and that spikes approximately 57 times per second for acceptability -1. In a multidimensional space, some neurons may be tuned in between multiple dimensions. A neuron could be tuned to spike rapidly whenever the second and third dimensions are at 1, for example. That neuron would be tuned to a direction in between the second and third dimensions and hence contribute to more than one numeric representation.

Together, the spiking of the neurons in the entire population represent the acceptability of each proposition, whose acceptability would be represented by single unit activations in a localist network. Since some neurons are tuned to directions in between dimensions, a single neuron is often excited by multiple propositions and, hence, the acceptability values of propositions overlap and become truly distributed throughout the neural population. Positive and negative acceptability is generally represented with different neurons because acceptance and rejection are completely opposite directions in the vector space (think positive and negative on a 1-dimensional number line). Hence when a proposition is accepted, positively tuned neurons will fire more and negatively tuned neurons will fire less, and the opposite is true when a proposition is rejected.

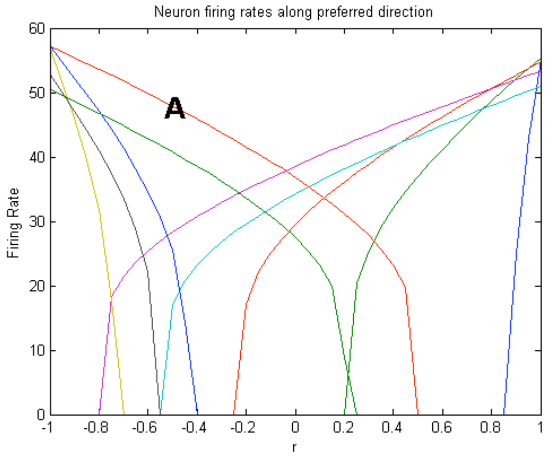

**Figure 8.** Tuning curves of selected neurons, showing how they spike while representing numbers between –1 (reject) and 1 (accept). The firing range is from 0 to 60 spikes per second.

The next step is to represent positive and negative constraints. NECO accomplishes this by generating a set of recurrent synaptic connections between the neurons in the neural populations representing different propositions. If two propositions cohere, the neurons tuned to the same acceptability (high or low) of each dimension representing those propositions will excite each other, and neurons tuned to opposite acceptability values will inhibit each other. If two propositions contradict, the opposite is true. The amount by which one neuron can affect another neuron is controlled by changing the synaptic weight between those neurons. Such wiring is considerably more complex than the simple symmetric link between units used by localist models. But it can accomplish a similar purpose, in that the interactions among all the neurons maximize constraint satisfaction. In parallel, neurons spike and cause other neurons to spike until the spiking patterns of all the neurons in all the population have stabilized. Neurons representing the acceptance of evidence propositions are excited by an external signal to simulate one's basic acceptance of observed evidence. Appendix B provides mathematical details of the NECO model and describes how the spiking pattern of the neurons in a population can be decoded to discern the acceptability of the proposition represented by the population.

Each neuron becomes either more active as acceptability for a proposition increases or as acceptability decreases depending on whether the neuron was tuned in the positive or the negative direction of that proposition. The spiking rates of each neuron are limited by the refractory period (minimum time between spikes) and the level of excitation directed into that neuron. For our model, we used a biologically plausible refractory period of 2ms (Eliasmith & Anderson, 2003). This upper limit of spiking rate translates to an upper limit of numeric representation since the numeric decoding is performed by multiplying windowed spiking rates by precalculated decoding weight constants. This limit allows the network to stabilize and is psychologically plausible, as there is a maximum to how firmly a person can accept a proposition.

Although we decode spiking rates to numerical values, we are not claiming that the actual number is represented in the brain when performing parallel constraint satisfaction. Instead, “1” stands for “very high” and “-1” stands for “very low”. The decoding processes is merely an appropriate external representation of the spiking patterns in the neurons. Simply by multiplying the decoding weights by a constant, we can change the range of values possible and decode different numbers while having the same neural activity. Decoding the neural activity to a number simply makes for easy analysis and representation of results. As seen later in figure 10, the decoded values are not precise due to various elements of noise found in the system.

Several examples from Thagard (1989) were recreated in the Neural Engineering Framework. One, shown in figure 9, involves not only determining which hypothesis is most correct, but also rejecting evidence because it does not fit with the best theory. Figure 10 graphs the acceptability of the different hypotheses and evidence, decoded from the spiking patterns of the neural populations representing activations of the hypotheses and evidence. Notice that E5 starts off as acceptable, but ends up rejected as constraint satisfaction is computed by the neural populations. Thus NECO duplicates the behavior of ECHO in a more neurologically realistic manner, and the same method could be used to solve a wide variety of parallel constraint satisfaction problems.

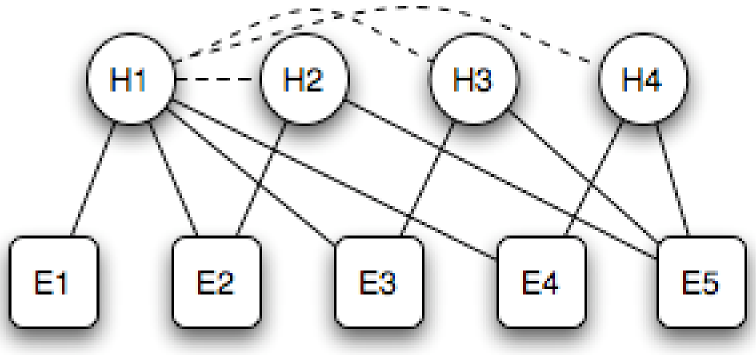

**Figure 9.** Explanatory coherence network in which evidence E5 is rejected because it is only explained by the inferior hypotheses H2, H3,

and H4. As in figure 7, solid lines are excitatory links (representing positive constraints) and dotted lines are inhibitory.

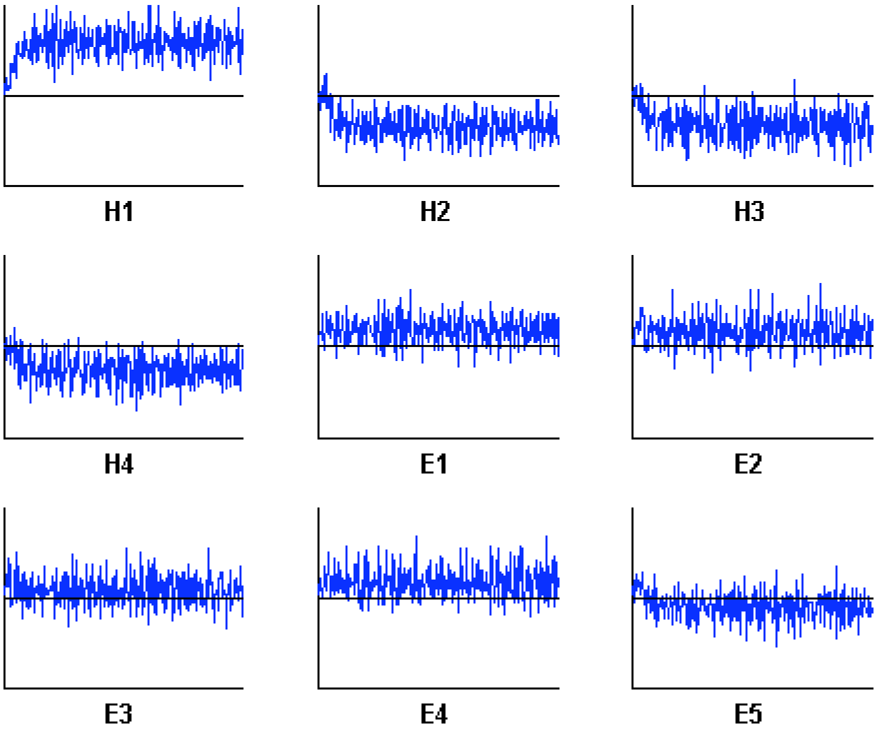

**Figure 10.** Behavior of neural populations over 300 simulated milliseconds representing the hypotheses and evidence in figure 9. The graphs show the degree of acceptability of each proposition represented by the firing pattern of the neurons. Firing above the line in the middle of each graph indicates acceptance.

## Emotional Coherence

Thagard (2000, 2003) showed how interactions between cognition and emotion can be understood in terms of parallel constraint satisfaction, if mental representations are assumed to have an emotional value, called a valence, as well as a degree of acceptability. Interactions were modeled by a program, called HOTCO for "Hot

Coherence", that is like ECHO except that units have a numerical valence as well as an activation. HOTCO has been used to model cases of motivated inference, in which people's beliefs are affected not only by the evidence for them but by their personal goals and other emotional attachments (Kunda, 1990).

In order to make NECO more neurologically realistic, we have modeled valence through interactions among multiple brain areas rather than as a function computed by the same neural populations that compute acceptability. NECO assumes that propositions carry an emotional memory with them that can be used to trigger emotional responses when these propositions are brought into working memory. NECO organizes neural populations into six interacting areas shown in figure 11. To model working memory, an external input signal for mental focus is used to activate neural populations. The lateral prefrontal cortex (Lateral PFC) carries out the cognitive coherence operations described in the last section with a single neural population computing the acceptabilities of all propositions. Also in the prefrontal cortex are three other populations called "PF Val Mem", "PF Val In" and "PF Val Switch" (where "val" stands for valence), each of which is made up of several sub-populations representing individual propositions. To speed up computation, we model values in distinct neural populations instead of the more distributed, overlapping manner used for acceptability.

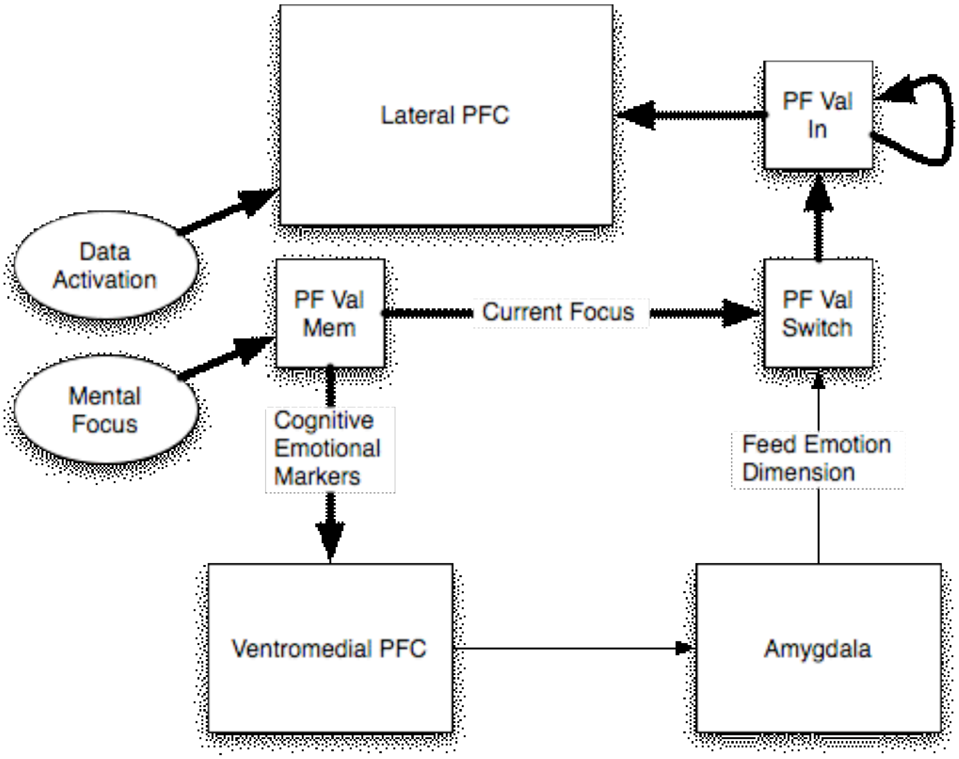

**Figure 11.** NECO model of emotional coherence. Lateral PFC is lateral prefrontal cortex. Ventromedial PFC is ventromedial prefrontal cortex. Mental focus is an external signal corresponding to working memory.

The emotional value (positive or negative) of each proposition is stored as long term memory in its corresponding neural population in "PF Val Mem" which is then connected to the ventromedial prefrontal cortex (VM). Bechara, et al. (2003) showed that the VM plays a crucial role in connecting PFC activity to a key emotional area of the brain, the amygdala. NECO uses the VM to consolidate the emotional valence input from all propositions and to send the sum of this input into the amygdala. The mental focus, modeled by an external signal, is directed at only one proposition at a time that causes the VM to represent the emotional value of only a single proposition at one time, but could be extended to model overlapping activations by combining emotional values.

The amygdala receives the emotional value from the VM and then feeds into the PF Val Switch population. This population is comprised of sub-populations, one for each proposition, that encode both the focus value of a proposition and the emotional value signal from the amygdala. The product of the encoding is then sent into PF Val In which is also comprised of sub-populations for each proposition. Since PF Val Switch outputs the *product* of the focus value and the amygdala's emotional state, only the proposition in focus will actually output a value (all others have a focus value of zero). This gives it a role similar to a light switch in that it completes the circuit. Once a value has been received by the PF Val In population, a recurrent connection is used to remember that value over time, retaining it in working memory. PF Val In, connected directly into the Lateral PFC population, "pulls" the acceptability levels of propositions in the direction of their emotional markers by adding either a positive or negative value to their current acceptability levels. In this way, a positive emotional value will encourage the acceptance of a proposition while a negative emotional value will encourage its rejection.

Consider again the network of hypotheses and evidence in figure 9, which produces the rejection of E5. Suppose that some scientists have a strong emotional attachment to this piece of evidence, perhaps because they themselves performed the experiments. In this case, E5 has a strong positive emotional valence, and motivated inference will militate against its rejection. Figure 12 shows the behavior of the neural populations from figure 7 modified to allow the influence of emotional valence in accord with the mechanism in figure 11. The positive valence connected to E5 sways the network into accepting E5 instead of rejecting it.

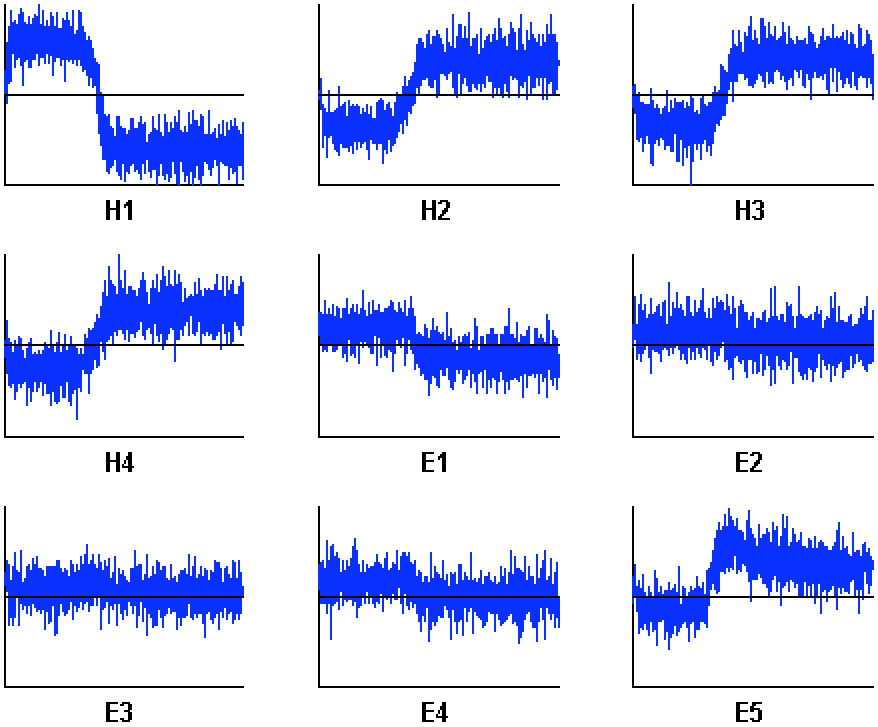

**Figure 12.** Effect of emotional valence over 1000 simulated milliseconds. When the Lateral PFC receives a strong positive emotional valence for E5 at 300 ms by virtue of connections with the amygdala, E5 becomes more acceptable and affects the rest of the network.

Figure 12 was produced by lowering the weights between populations in the Lateral PFC to downplay the effect of non-emotional explanatory coherence, allowing emotion to influence the simulation even after it had settled into a stable state. If higher weights are used in the Lateral PFC then there is only a small change in the activation of E5 that is not enough to tip the overall balance in the network. Thus figure 12 might be taken to represent the biased reasoning of someone with strong emotional commitments and weak logical ones.

The NECO model of emotional coherence falls far short of implementing the much more general EMOCON model. But NECO is important for the theory of emotional consciousness because it shows how emotional appraisal construed as parallel constraint satisfaction can be translated into a mechanism that has some neural plausibility, using distributed representations and interconnections with emotionally important subcortical regions such as the amygdala. Hence it fills in some of the gaps in the description of the neural mechanisms that constitute our account of emotional consciousness.

## 9. PHILOSOPHICAL ISSUES

Our neurocomputational account of emotional conscious may be challenged by several philosophical objections to all materialist accounts of consciousness. Descartes (1964) argued that he could imagine himself without a body, but not without a mind, so that he was essentially a thinking thing rather than a material body. The modern version of this argument is the thought experiment of Chalmers (1996) and others that we can imagine a being that is just like us physically but lacks consciousness, which is supposed to show that consciousness is not open to neurological explanation. For emotional consciousness, the point would be that we can imagine a being that has all the physiological mechanisms shown in figure 3 but which lacks emotional experiences such as happiness. Hence these mechanisms do not explain why people feel happy.

As Paul Churchland (1996) and Patricia Churchland (2002, 2005) have repeatedly pointed out, this is a very weak argument. That we can imagine a being with the EMOCON mechanisms that lacks emotional consciousness is simply irrelevant to assessment of the theory, which concerns how people in this world have emotions, not with emotions in all possible worlds. We can imagine that combustion is not rapid oxidation, that light is not electromagnetic radiation, and that heat is not the motion of molecules. But we have ample scientific evidence in the form of experimental results explained by the respective theories to convince us that combustion *is* rapid oxidation, that light *is* electromagnetic radiation, and that heat *is* the motion of molecules. All these theories took centuries of intellectual development. For example, Lavoisier's theory had to surpass Stahl's phlogiston theory of combustion, which had surpassed the ancient theory that fire is an element. By the end of the eighteenth century, there was ample evidence that things burn because of chemical reactions with oxygen. Perhaps soon there will be sufficient evidence to convince impartial reasoners that emotional consciousness really is neural activity of the sort sketched in the EMOCON model. Some progress in this direction has been made by the demonstration above that it is already possible to use known neurophysiological mechanisms to outline explanations for valence, intensity, onset, cessation, and discriminability of emotional experiences.

Figure 3 might be misread as suggesting that emotional feelings result from neural activity but do not cause them. But we maintain that emotional consciousness just *is* the neural activity distributed across multiple brain areas, not the epiphenomenal result of neural activity. Hence consciousness can have effects, such as influencing actions, that are the result of the kinds of neural activity shown in figure 3.

Another standard philosophical objection to identification of mental states with brain states is that we can imagine non-human entities such as robots and space aliens that have the same mental states but different neurophysiology. Our reply is that the theory of emotional consciousness proposed here is not supposed to apply to all possible thinking beings, but only to humans and other similar terrestrial animals. If we ever encounter robots or space aliens that seem to have emotional consciousness, we would expect their experiences to be rather different from ours, because of the different character of their external sensors, internal sensors, and neural organization. The EMOCON mechanisms sketched in figure 3 would still constitute the best explanation of *human* emotional consciousness. A less speculative issue concerns the similarities and differences between human consciousness and that of non-human animals. This is an empirical question that cannot adequately be addressed until more is known about the full set of mechanisms that support human emotional experience and their closeness to analogous mechanisms in other animals. Similarities in functional units such as the thalamus and amygdala must be weighed against differences such as the much greater capacity of the human prefrontal cortex. See Panksepp (2005) for further discussion.

The final philosophical objection that must be addressed is that the explanation of emotional consciousness in terms of mechanistic neural activity is incomplete because it does not tell us *what it is like* to have experiences such as being happy and sad (Nagel, 1979). On this view, our own personal experience tells us much about emotional consciousness that no scientific theory could ever address. But our theory of emotional consciousness has in fact suggested explanations of many aspects of what it is like be happy, for example that happiness has positive valence, varying intensity, onset and cessation, and that it is discriminable from other emotions. A dualist theory that sees consciousness as a special kind of entity independent of neural mechanisms cannot even begin to explain in a non-mysterious fashion these aspects of emotional experience. Hence the hypothesis that emotional consciousness is neural activity of the sort presented in the EMOCON model, interconnected to the world and to the body, survives as the best current explanation of what we know about emotional experience.

## 10. CONCLUSION

We have presented a unified theory of emotional consciousness that integrates many components, including somatic representation, cognitive appraisal, neural affective decision making, and working memory. It should not be surprising that explanations of phenomena that involve both emotions and consciousness require a wide range of neurological and physiological mechanisms, of which the EMOCON model shown in figure 3 is only a rough approximation. As Einstein said, everything should be as simple as possible but not simpler, and greater simplicity is not to be had in this domain. There is ample experimental evidence for the relevance of each of the neurological components assumed in the model. The EMOCON model is largely consistent with sophisticated frameworks for naturalizing consciousness proposed by Baars (2005), Edelman (2003), Koch (2004), Pankseep (2005) and Tononi (2004). But it goes beyond them in proposing a specific neural mechanism for the generation of one kind of conscious experience, emotions. Unlike Rolls (2005), our account does not require that emotional consciousness involves higher-order linguistic thoughts, although these may be a part of some human emotions requiring complex cognitive appraisal. Our model is broadly compatible with the account of emotion experience presented by Lambie and Marcel (2002), which also integrates evaluation and physiological changes, but without specifying neural mechanisms. We have neglected, however, the modulation of action, which they rightly mark as an important function of emotion.

Despite EMOCON's comprehensiveness, there are notable elements not included in the model because they do not seem to increase its explanatory power. We have not included any special non-material substance of the sort that theologians and other dualists have postulated as the basis of consciousness. Nor have we invoked quantum-mechanical processes that generate consciousness, because we do not believe that quantum theory is much relevant to explaining psychological processes (Litt et al., 2006). We have not assigned any special role to neural synchronization accomplished by a 40Hz (40 cycles per second) brain wave that various theorists have speculated might contribute to binding representations together (e.g. Engel et al., 1999). If there is such synchronization, it is likely an effect of neural processing rather than a causal factor. We have also not seen any need to postulate a special role for the claustrum, which Crick and Koch (2005) have deemed relevant to consciousness but which does not seem to play any special role in emotional processing. However, we are open to the suggestion that a deeper understanding of temporal coordination in the brain will be a major part of a more detailed model, in keeping with the suggestions of Churchland (2005), Davidson (2002), and Humphrey (2006) that time and its representation plays a greater role in consciousness than has been yet acknowledged.

If the EMOCON account of emotional consciousness is correct, it has implications for other interesting psychological phenomena including intuition and ethical judgment. Intuitions, or gut feelings, are conscious judgments that can be viewed as arising from the same interconnected processes described in figure 3. Similarly, ethical judgments are always emotional and conscious, but they can also have a cognitive-appraisal component that complements the somatic signaling that is also part of our account. Thus the identification of some of the neurophysiological mechanisms responsible for emotional consciousness should help to illuminate many other aspects of human thinking.

In sum, this paper has attempted to make two contributions to the understanding of emotional consciousness. First, it provides a new theoretical account of the neural mechanisms of emotion that synthesizes previously disjoint accounts based on somatic perception and cognitive appraisal. Second, it shows how these mechanisms can give rise to central aspects of emotional experience, including integration, differentiation, valence, intensity, and change.

**Acknowledgments.** Development of these ideas has benefited from conversations with Chris Eliasmith and Abninder Litt, although they should not be held responsible for our wilder speculations. Thanks to the Natural Sciences and Engineering Research Council of Canada for funding. We are grateful to Abninder Litt and two anonymous referees for comments on a previous draft.

# REFERENCES

Anderson, A. K., Christoff, K., Stappen, I., Panitz, D., Ghahremani, D. G., Glover, G., et al. (2003). Dissociated neural representations of intensity and valence in human olfaction. *Nature Neuroscience, 6*(2), 196-202.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in Brain Research, 150*, 45-53.

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on psychological science, 1*, 28-58.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*, 1293-1295.

Bechtel, W., & Abrahamsen, A. A. (2005). Explanation: A mechanistic alternative. *Studies in History and Philosophy of Biology and Biomedical Sciences, 36*, 421-441.

Carruthers, P. (2000). *Phenomenal consciousness: A naturalistic theory*. Cambridge: Cambridge University Press.

Chalmers, D. J. (1996). *The conscious mind.* Oxford: Oxford University Press.

Churchland, P. M. (1996). The rediscovery of light. *Journal of Philosophy, 93*, 211-222.

Churchland, P. S. (2002). *Brain-wise: Studies in neurophilosophy*. Cambride, MA: MIT Press.

Churchland, P. S. (2005). A neurophilosophical slant on consciousness research. *Progress in brain research, 149*, 285-293.

Clore, G. L., & Ortony, A. (2000). Cognitive neuroscience of emotion. In R. D. Lane & L. Nadel (Eds.), *Cognitive neuroscience of emotion* (pp. 3-61). Oxford: Oxford University Press.

Crick, F., & Koch, C. (2005). What is the function of the claustrum? *Philosophical Transactions of the Royal Society B, 360*, 1271-1279.

Damasio, A. R. (1999). *The feeling of what happens:  Body and emotion in the making of consciousness*. New York: Harcourt Brace.

Damasio, A. R., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L. L., Parvizi, J., et al. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience, 3*(10), 1049-1056.

Davidson, R. J. (2002). Anxiety and affective style: role of prefrontal cortex and amygdala. *Biol Psychiatry, 51*(1), 68-80.

Davidson, R. J. (2004). What does the prefrontal cortex "do" in affect:  Perspectives on frontal EEG asymmetry research. *Biological psychology, 67*, 219-233.

Descartes, R. (1964). *Philosophical writings* (E. Anscombe & P. T. Geach, Trans.). London: Nelson.

Dolcos, F., LaBar, K. S., & Cabeza, R. (2004). Dissociable effects of arousal and valence on prefrontal activity indexing emotional evaluation and subsequent memory:  An event-related fMRI study. *NeuroImage, 23*, 64-74.

Durstewitz, D., Seamans, J., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience, 3*(Supplement), 1184-1191.

Edelman, G. M. (2003). Naturalizing consciousness: A theoretical framework. *Proceedings of the National Academy of Sciences, 100*, 5520-5524.

Ekman, P. (2003). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. New York: Henry Holt.

Eliasmith, C. (2003). Moving beyond metaphors: Understanding the mind for what it is. *Journal of Philosophy, 100*, 493-520.

Eliasmith, C. (2005). Neurosemantics and categories. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 1035-1054). Amsterdam: Elsevier.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Engel, A. K., Fries, P., König, P., Brecht, M., & Singer, W. (1999). Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition, 8*, 128-151.

Fuster, J. M. (2002). *Cortex and mind: Unifying cognition*. Oxford: Oxford University Press.

Gershon, A. A., Dannon, P. N., & Grunhaus, L. (2003). Transcranial magnetic stimulation in the treatment of depression. *American Journal of Psychiatry, 160*(5), 835-845.

Griffiths, P. E. (1997). *What emotions really are: The problem of psychological categories*. Chicago: University of Chicago Press.

Hamann, S. B., Ely, T. D., Hoffman, J. M., & Kilts, C. D. (2002). Ecstasy and agony: Activation of the human amygdala in positive and negative emotion. *Psychological Science, 13*(2), 135-141.

Humphrey, N. (2006). *Seeing red: A study in consciousness*. Cambridge, MA: Harvard University Press.

James, W. (1884). What is an emotion? *Mind, 9*, 188-205.

Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge: Cambridge University Press.

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience, 23*, 315-341.

Kawasaki, H., Kaufman, O., Damasio, H., Damasio, A. R., Granner, M., Bakken, H., et al. (2001). Single-neuron responses to emotional visual stimuli recorded in human ventral prefrontal cortex. *Nature Neuroscience, 4*, 15-16.

Koch, C. (2004). *The quest for consciousnes: A neurobiological approach*. Englewood, CO: Roberts and Company.

Kunda, Z. (1990). The case for motivated inference. *Psychological Bulletin, 108*, 480-498.

Lambie, J. A., & Marcel, A. J. (2002). Consciousness and the varieties of emotion experience: A theoretical framework. *Psychological Review, 109*, 219-259.

LeDoux, J. (1996). *The emotional brain*. New York: Simon and Schuster.

Litt, A., Eliasmith, C., Kroon, F. W., Weinstein, S., & Thagard, P. (2006). Is the brain a quantum computer? *Cognitive Science, 30*, 593-603.

Litt, A., Eliasmith, C., & Thagard, P. (2006). Why losses loom larger than gains: Modeling neural mechanisms of cognitive-affective interaction. In R. Sun & N. Miyake (Eds.), *Proceedings of the twenty-eighth annual meeting of the Cognitive Science Society* (pp. 495-500). Mahwah, NJ: Erlbaum.

Lycan, W. G. (1996). *Consciousness and experience*. Cambridge, MA: MIT Press.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1-25.

Mayberg, H. S., Lozano, A. M., Voon, V., McNeely, H. E., Seminowicz, D., Hamani, C., et al. (2005). Deep brain stimulation for treatment-resistant depression. *Neuron, 45*(5), 651-660.

Mellers, B., Schwartz, A., & Ritov, I. (1999). Emotion-based choice. *Journal of Experimental Psychology: General, 128*, 332-345.

Metzinger, T. (Ed.). (2000). *Neural correlates of consciousness*. Cambridge, MA: MIT Press.

Morris, J. S. (2002). How do you feel? *Trends in Cognitive Sciences, 6*, 317-319.

Nagel, T. (1979). *Mortal questions*. Cambridge: Cambridge University Press.

Nerb, J. (forthcoming). The dynamics of the appraisal-emotion relationship. *Cognition and Emotion*.

Nerb, J., & Spada, H. (2001). Evaluation of environmental problems: A coherence model of cognition and emotion. *Cognition and Emotion, 15*, 521-551.

Niedenthal, P. M., Barsalou, L. W., Ric, F., & Krauth-Gruber, S. (2005). Embodiment in the acquisition and use of emotion knowledge. In L. Barrett, P. M. Niedenthal & P. Winkielman (Eds.), *Emotion and consciousness* (pp. 2-50). New York: The Guilford Press.

Nussbaum, M. (2001). *Upheavals of thought*. Cambridge: Cambridge University Press.

Oatley, K. (1992). *Best laid schemes: The psychology of emotions*. Cambridge: Cambridge University Press.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge: Cambridge University Press.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford: Oxford University Press.

Panksepp, J. (2005). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition, 14*(1), 30-80.

Prinz, J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford: Oxford University Press.

Prohovnik, I., Skudlarski, P., Fulbright, R. K., Gore, J. C., & Wexler, B. E. (2004). Functional MRI changes before and after onset of reported emotions. *Psychiatry Research Neuroimaging, 132*(3), 239-250.

Rainville, P., Bechara, A., Naqvi, N., & Damasio, A. R. (2006). Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal Psychophysiology, 61*(1), 5-18.

Rolls, E. R. (2005). *Emotion explained*. Oxford: Oxford University Press.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*, 145-172.

Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks, 18*, 317-352.

Schacter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*, 379-399.

Scherer, K. R., Schorr, A., & Johnstone, T. (2001). *Appraisal processes in emotion*. New York: Oxford University Press.

Seager, W. (2002). Emotional introspection. *Consciousness and  Cognition, 11*, 666-687.

Smith, E. E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science, 283*, 1657-1661.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435-467.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thagard, P. (2003). Why wasn't O. J. convicted?  Emotional coherence in legal inference. *Cognition and Emotion, 17*, 361-383.

Thagard, P. (2005). *Mind:  Introduction to cognitive science* (2nd ed.). Cambridge, MA: MIT Press.

Thagard, P. (2006). *Hot thought:  Mechanisms and applications of emotional cognition*. Cambridge, MA: MIT Press.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*, 42.

Wagar, B. M., & Thagard, P. (2004). Spiking Phineas Gage:  A neurocomputational theory of cognitive-affective integration in decision making. *Psychological Review, 111*, 67-79.

Wierzbicka, A. (1999). *Emotions across languages and cultures:  Diversity and universals*. Cambridge: Cambridge University Press.

Wordnet. (2005). *Wordnet:  A lexical database for the English language*. Retrieved August 18, 2006, from http://wordnet.princeton.edu/

## APPENDIX A: MATHEMATICAL DETAILS OF EACO

Our model is programmed in PHP, a Web-based scripting language, and runs on any standard web server supporting PHP4 or PHP5. The data input format has been designed for flexibility in defining connection weights and sensory activations. The full source code is available upon request.

Computational models of parallel constraint satisfaction work by spreading unit activation throughout a network of units (nodes) via weighted connections. Each connection can have either a positive or a negative weight that corresponds to an excitatory or an inhibitory connection respectively. Each node has a maximum activation level of 1 and a minimum activation level of either -1 or 0. At each time step the activation of unit *j* is updated with the following equation:

$$a_j(t+1) = a_j(t)(1-d) + enet_j(max_j - a_j(t)) + inet_j(a_j(t)-min_j)$$

The parameters are:

| | |
|---|---|
| $a_j(t)$ | Activation of unit j at time t. |
| d | Decay factor set to 0.05 for this model. |
| $enet_j$ | $\sum_i w_{ij} a_i(t)$ for $w_{ij} > 0$ which is the weighted sum of connecting excitatory units. |
| $inet_j$ | $\sum_i w_{ij} a_i(t)$ for $w_{ij} < 0$ which is the weighted sum of connecting inhibitory units. |
| $max_j$ | The maximum activation level for unit j. |
| $min_j$ | The minimum activation level for unit j. |

In EACO, each emotion has up to 16 different SECs that distinguish it from other emotions. Not all SECs necessarily have a specific value for an emotion and may be classified as *open*. Hence many different results may be compatible with that particular emotion or that the check is irrelevant, resulting in an unbalanced connection matrix between the emotions. For example, DESPAIR has 13 defined SECs used to activate it while SHAME has only 7. If weighted equally, DESPAIR would have a much stronger activation level then SHAME would. To combat this unbalance, connection weights are normalized so that their sum is always equal to one. In the case of DESPAIR, the connection weights from each SEC to the emotion node are 1/13 (or 0.077) and in the case of SHAME the connection weights from each SEC to the emotion node are 1/7 (or 0.143).

## APPENDIX B: MATHEMATICAL DETAILS OF NECO

The Neural Engineering Framework (NEF) used in this project provides the ability to represent real vector spaces of any dimension using spiking, leaky-integrate-and-fire neurons (for an in-depth review of LIF neurons and an argument for why they are appropriate for modeling biologically plausible models, see Eliasmith & Anderson (2003) pp. 81-89). An *n*-dimensional real vector space can be thought of as all ordered sets of *n* real numbers. For example, (1, 2, 3) is a point in a 3-dimensional vector space and (7, -103, 4.23, 40, 0) is a point in a 5-dimensional vector space. In a population of neurons, each neuron is tuned to be most responsive to a randomly chosen direction in that vector space. Thus, as more neurons are used, the vector space becomes better represented. Input signals are encoded into spikes for each neuron where the behavior of any individual neuron depends on the direction to which it is tuned. Decoding is done by taking the spiking rates of neurons at any point in time and multiplying them by decoding weight vectors. These decoding weight vectors are found by minimizing the difference between the desired decoded signal and the estimated decoded signal. We can optimize these decoding vectors to decode not only the input signal but also a function of the input signal. For example, a one-dimensional population of neurons with an input signal of $x(t)$ (some function of time) could encode $x(t)$ with spikes and the decoding weight vectors could be found such that they decode the spikes to $[x(t)]^2$. This would result in the population of neurons literally computing the function $x^2$. Using higher dimensional vector space representations, more complicated functions such as $f(x,y,z) = (\sin(xy), zx, y-x^3+\log z)$ can readily be computed. This encoding and decoding strategy performs better as the number of neurons increases, which ensures that the vector space is well represented.

In higher dimensional spaces the tuning curves become multidimensional, but retain the same nonlinear shape (see Figure 2). A point on a tuning curve represents the firing rate of that neuron for a particular input signal. The shape of the tuning curve can be derived from the of leaky integrate-and-fire model of a neuron and works out to

$$a(x) = \frac{1}{\tau^{ref} - \tau^{RC} \ln\left(1 - \frac{J^{th}}{\alpha x + J^{bias}}\right)} \qquad (1)$$

where $J^{th}$=1 generally and the other parameters are randomly varied to get the different tuning curves seen in figure 2. The decoding equation for these neurons is

$$\hat{x}(t) = \sum_{i}^{N} a_i(x(t))\phi_i \qquad (2)$$

where $\hat{x}$ is the decoded signal vector found by summing over the spiking rate of each neuron, $a_i(x(t)) = G[\alpha x(t)+J^{bias}]$, multiplied by the decoding weight vector of that neuron, $\phi_i$. The $\alpha$ term normalizes the input signal to a predetermined range and the $J^{bias}$ term is used to adjust for a bias, or background firing rate. The $G$ function is arbitrary and can be replaced with a linear firing rate, an LIF firing rate (see eq 1), or any other firing rate curve. In our case, we use actual spikes and let G be a postsynaptic current-derived windowing function that finds the instantaneous spiking rate at any point in time and ends up approximating the tuning curve derived in equation 1 (see Eliasmith & Anderson, 2003, pp. 112-113). Thus our model ultimately comes down to a firing rate but the properties of that rate come out of the spiking parameters such as the refractory time between spikes and the flow of current that generates these spikes. Thus our model is more true to the brain than a typical rate coded model that would have to manually extract these effects. Extending this method to decode functions of the input involves finding decoding weights over a function space (Eliasmith & Anderson, 2003), but the basic idea is the same. Computing the decoding weights require finding scalar values, $\phi_i$, for each neuron such that the difference between the desired decoding and the actual decoding is minimal. When dealing with a simple communication channel, where the output is meant to be the same as the input, we must minimize the error (*E* below) and solve for each neuron's decoding weight, $\phi_i$ (the function case is, again, similar).

$$E = \frac{1}{2}\int_{-1}^{1}\left[\vec{x} - \sum_{i=1}^{N} a_i(\vec{x})\phi_i\right]^2 dx \qquad (3)$$

Notice that the decoding equation (2) is linear, whereas the tuning curves of LIF neurons are nonlinear. Linear decoding schemes have been found to be more biologically plausible and the information loss of linear decoding schemes from nonlinear decoding schemes is only 5% (Eliasmith & Anderson, 2003). This loss can account for some of the noise in the system, while the rest can be accounted for by noise in spike times. A neuron will never spike at exactly 50Hz over any extended period of time, but instead, will fluctuate slightly due to electrical leakage, neurotransmitter properties, or other biological elements. As a result, encodings are never perfect and hence the decoding methods have elements of noise to contend with causing the interpreted values to be fuzzy.  It is well established that the brain does not encode information exactly as it is received, so this slight information loss is well within tolerance.