EMOTIONAL CONSCIOUSNESS: A NEUROCOMPUTATIONAL THEORY

Paul Thagard
University of Waterloo
pthagard@uwaterloo.ca

DRAFT 3, AUGUST 17, 2006. COMMENTS WELCOME.

ABSTRACT: This paper outlines 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. 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 neurocomputational model of how emotions arise from a combination of neural representation, somatic signals, cognitive appraisal, and working memory.

First I 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. I 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 neurocomputational models of decision making and parallel constraint satisfaction, I outline an integrated model of emotion in the brain that includes an account of working memory. I then show how the model explains a wide range of crucial phenomena about emotional consciousness. Finally, I 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 my aim is more ambitious. I 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

phenomena 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 phenomena. 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). 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* I 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 (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 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* I 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.

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 one. Emotional changes include shifts of 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 change their attitude toward an object or state of affairs. Another kind of emotional change occurs when therapy, medication, or both help a depressed person to assume a more positive view of life.

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 such as sight and sound, 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 the 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). My own 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. 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 integrate perceptions and judgments.

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). On this account, 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.

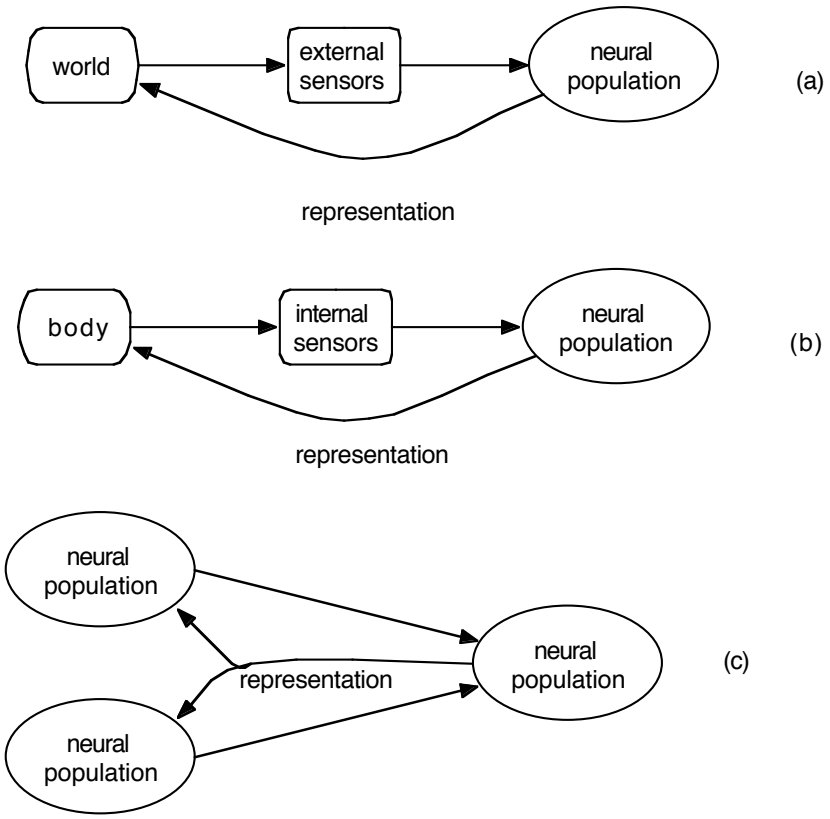

Figure 1. Representation by neural populations of (a) aspects of the world, (b) aspects of the body, and (c) other neural populations.

Similarly, neural populations can represent bodily states and events as shown in figure 1(b). Just as our bodies have externals 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 the representation 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 populations 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. My starting point is a recent theory of emotional decision making proposed by Litt, Eliasmith, and Thagard (2006, forthcoming): Neural Affective Decision Theory. According to this theory, all decision making has an emotional component that involves the interaction of

at least seven major brain areas that contribute to valuation 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 structure of ANDREA is sketched in figure 2. The role of each of the indicated areas in emotion is well 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 is important for processing negative emotions such as fear and also for modulating the intensity of positive emotions. 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. Litt, Eliasmith, and Thagard (forthcoming) show that 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 about conscious experience, but I shall describe natural extensions that provide the additional ingredients needed to account for consciousness.

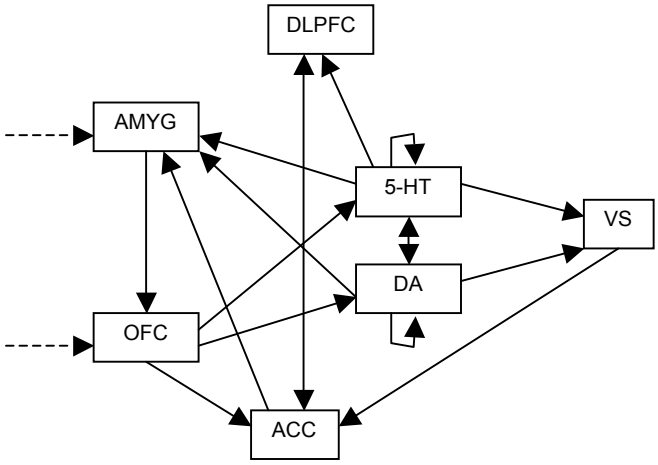

Figure 2. The ANDREA model of decision evaluation, from Litt, Eliasmith, and Thagard (2006). Dotted arrows represent external inputs to the model. Abbreviations: 5-HT, dorsal raphe serotonergic neurons; ACC, anterior cingulate cortex; AMYG, amygdala; DA, midbrain dopaminergic neurons; DLPFC, dorsolateral 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 on the Iowa gambling task and the behavior of people in the famous experiments of Schacter and Singer (1962). Thus the VMPFC and hippocampus need to 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). There therefore seem to be at least 11 interacting brain areas that need to be included in a full theory of emotional consciousness.

Inference and Appraisal

Still missing from my 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. 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.

Aubie and Thagard (forthcoming-a) show how parallel constraint satisfaction can be performed by distributed representations using large numbers of neurons in accord with the neural engineering framework of Eliasmith and Anderson (2003). Their model, called NECO for Neural-Engineering-Coherence, uses populations of thousands of neurons to perform complex computations, but does not introduce any new brain areas beyond the prefrontal cortex and amygdala already discussed. They are currently extending their distributed model to perform appraisal of emotional situations (Aubie and Thagard, forthcoming-b), including both cognitive appraisal with respect to goals and somatic information provided via the amygdala and the insula.

Working Memory

One last 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. Aubie and Thagard (forthcoming-a, b)

use 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 Morris (2002) about sensory inputs. I conjecture that emotional experience is the result of interactions among *all* the components shown in figure 3. My collaborators and I have not yet programmed such a large and complicated simulation, which exceeds our current computational resources, but I 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 has room for complex inferences in the dorsolateral prefrontal cortex 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.

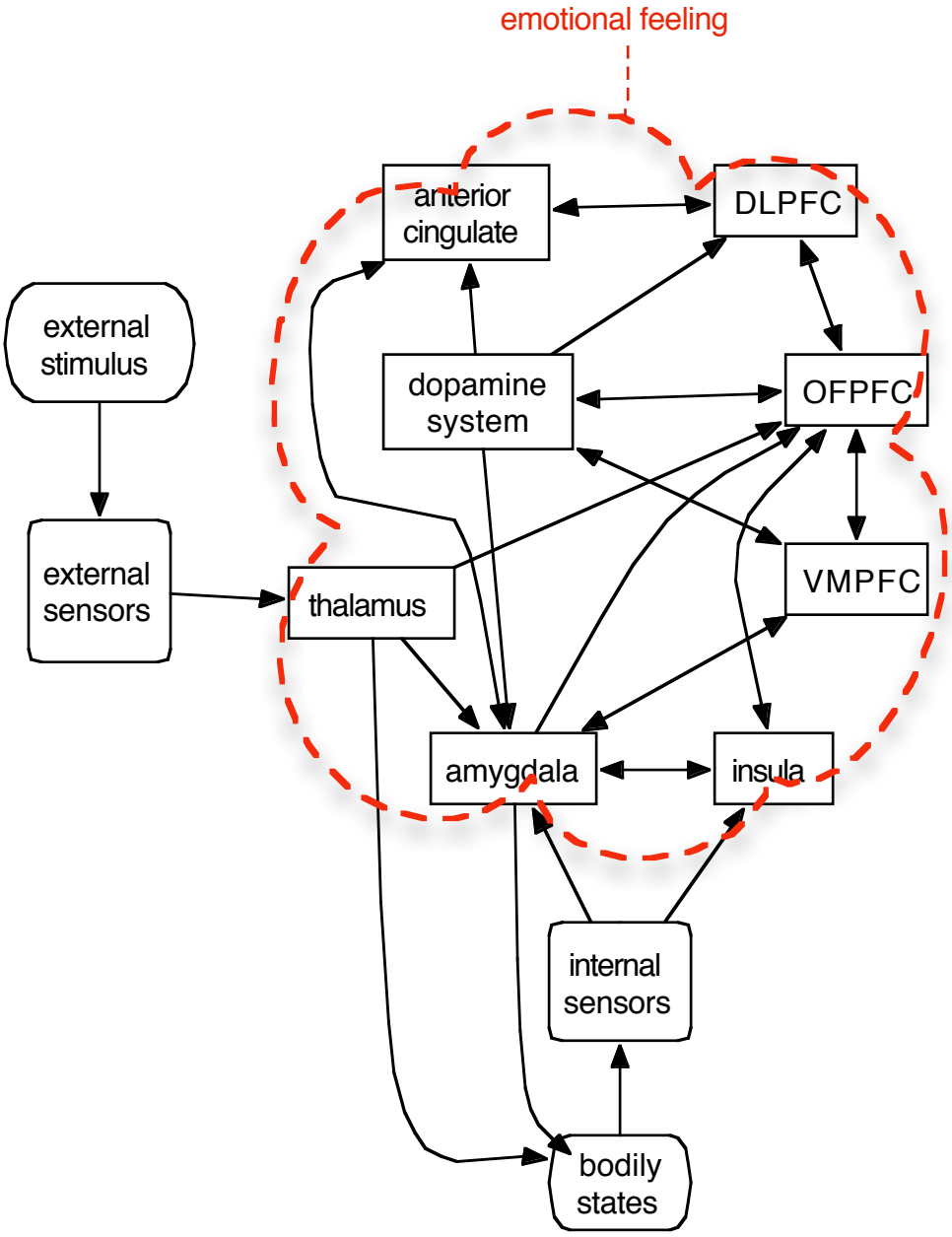

Figure 3. The EMOCON model of emotional consciousness, incorporating aspects of Litt, Eliasmith, and Thagard (2006), Wagar and Thagard (2004), Aubie and Thagard (forthcoming-a, b), and Morris

(2002). Abbreviations are PFC for prefrontal cortex, DL for dorsolateral, OF for orbitofrontal, and VM for ventromedial,

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. Note the presence in the diagram of numerous feedback loops, for example between the amygdala, bodily states, and internal sensors. It is important that 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. 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. I will now show how EMOCON explains valence, intensity, change, and differentiation.

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.

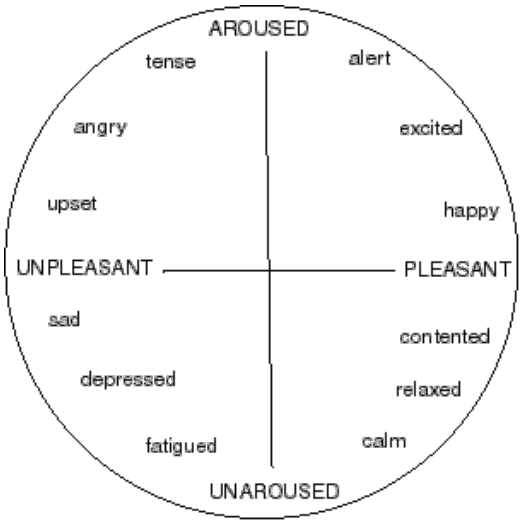

Figure 4. The structure of emotions with respect to pleasantness and intensity. Reprinted from Thagard (2005), p. 165.

From the EMOCON model in figure 3, it is easy to see 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. 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 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, would be 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 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. Unfortunately, I do not know of any experiments that show directly that increasing or decreasing brain activity will increase or decrease emotional intensity.

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 neurons and their synaptic connections, and activity is degree of neural firing. Thus 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 determined 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 there. However, working memory is also subject to decay, so that if there is no ongoing stimulation of its active contents from perception and memory, they will tend to drop out of working memory. Another likely mechanism in working memory is inhibition, in which the activation of some elements tends to suppress the activation of others.

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.

Differentiation

All emotions involve positive or negative valence and different degrees of intensity, but these two dimensions are not sufficient to distinguish 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 no one would 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) presents 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.

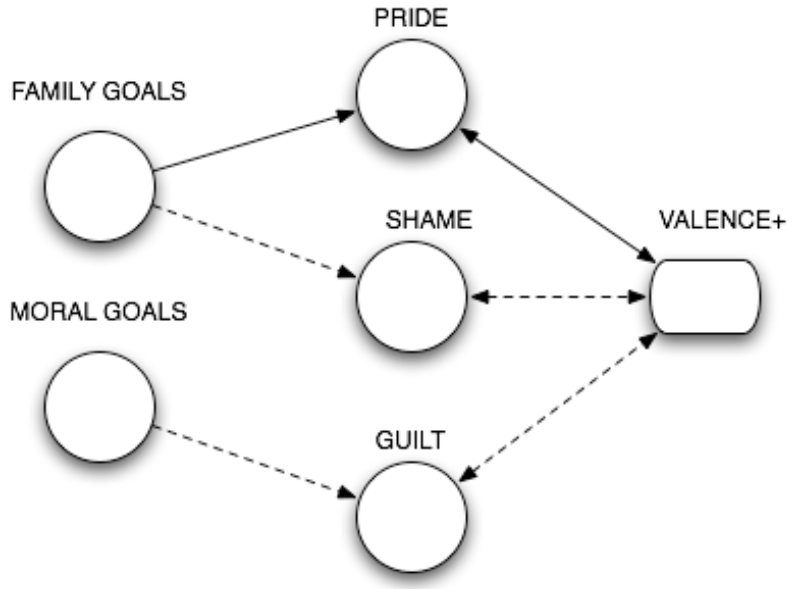

Figure 5. Drastically oversimplified constraint network for some social emotions. Solid lines are positive constraints, whereas negative lines are negative constraints.

The network shown in figure 5 is 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. Aubie and Thagard (forthcoming-b) describe how emotional appraisal can be modeled in a highly distributed fashion while still accomplishing parallel satisfaction of cognitive and affective constraints. 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. PHILOSOPHICAL ISSUES

My 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 emotion 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 sketch 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 I 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. My 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 note that my 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.

8. CONCLUSION

I 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 explanation of phenomena that involve both emotions and consciousness requires a wide range of neurological and physiological mechanisms, of which the EMOCON model shown in figure 3 is only a rough simulation. 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.

Despite its comprehensiveness, there are notable elements not included in the model because they do not seem to increase its explanatory power. Most notably, I 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 I postulated arcane quantum-mechanical processes that generate consciousness, because I 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 as an effect of neural processing rather than a causal factor. I 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, I am 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 Humphreys (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.

Acknowledgments. Development of these ideas has benefited from conversations with Brandon Aubie, Chris Eliasmith, and Abninder Litt, although they should not be held responsible for my wilder speculations. Thanks to the Natural Sciences and Engineering Research Council of Canada for funding.

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.

Aubie, B., & Thagard, P. (forthcoming-a). Coherence in the brain: A neurocomputational model of parallel constraint satisfaction.

Aubie, B., & Thagard, P. (forthcoming-b). Emotional appraisal: A distributed neurocomputational model.

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. S. (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.

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* (Vol. 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.

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.

Litt, A., Eliasmith, C., & Thagard, P. (forthcoming). Neural affective decision theory: Choices, brains, and emotions.

Lycan, W. (1988). *Judgement and justification*. Cambridge: Cambridge University 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.
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.
Smith, E. E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science, 283*, 1657-1661.
Thagard, P. (2005). *Mind: Introduction to cognitive science* (2nd ed.). Cambridge, MA: MIT Press.
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/