How Cognition Meets Emotion:
Beliefs, Desires, and Feelings as Neural Activity

Paul Thagard
University of Waterloo
pthagard@uwaterloo.ca

Thagard, P. (forthcoming). How cognition meets emotion: Beliefs, desires, and feelings as neural activity. In G. Brun, U. Doguoglu & D. Kuenzle (Eds.), *Epistemology and emotions*. Aldershot: Ashgate.

Deep appreciation of the relevance of emotion to epistemology requires a rich account of how emotional mental states such as happiness, sadness and desire interact with cognitive states such as belief and doubt. Analytic philosophy since Gottlob Frege and Bertrand Russell has assumed that such mental states are propositional attitudes, which are relations between a self and a proposition, an abstract entity constituting the meaning of a sentence. This chapter shows the explanatory defects of the doctrine of propositional attitudes, and proposes instead that beliefs, desires, and emotions should be construed naturalistically using current understanding of brain mechanisms. Mental states are patterns of neural activity, not relations between dubious entities such as selves and propositions. From this perspective, it becomes easy to see how cognition and emotion are intertwined, and hence how emotions can be integral to epistemology.

I begin by reviewing some of the ways in which emotions are relevant to epistemology: as frequent contributors to the growth of knowledge, as sometime impediments to knowledge acquisition, and as components of knowledge about persons and morality. I then argue that propositional attitudes do not exist, because the selves and the propositions that they purportedly relate do not exist. Thus the doctrine of propositional attitudes is as useless for epistemology as it is for explaining human action. I argue for an alternative construal of mental states as patterns of neural activity, and

describe how it is possible to give a theoretically rich and empirically supported account of the neurophysiological interconnections of cognition and emotion. Finally, I discuss the epistemological significance of this naturalistic, materialist reconstrual of cognitions and emotions.

CONNECTIONS BETWEEN EPISTEMOLOGY AND EMOTIONS

There are many ways in which emotions and epistemology are mutually relevant, but I will merely review some connections that I have explored at much greater length elsewhere (Thagard, 2006-b). Traditional analytic epistemology has largely ignored the role of emotions in knowledge, but from a broader perspective they are clearly relevant. In particular, understanding the development of scientific knowledge requires noticing a positive contribution of emotional thinking. Emotions such as interest, curiosity, wonder, and surprise are inextricable from the cognitive processes of scientific investigation, guiding researchers to generate important questions and to try to produce acceptable answers to them. The search for empirical and theoretical success can be accompanied by episodes of hope and happiness, but inevitable impediments can also lead to negative emotions such as worry, fear, frustration, anger, and disappointment. Positive emotions provide the motivational fuel to conduct the difficult work that is crucial to any scientific investigation. Both positive and negative emotions provide signposts about current progress and directions to pursue. Important breakthroughs are marked by elation, whereas setbacks can prompt disappointment and even despair. Philosophical thinking is also often driven by emotions, ranging from wonder at the marvels of the universe to angst and despair about the human condition.

In addition to their crucial role in helping to direct investigation, emotions are also involved in the acceptance and rejection of beliefs. Belief revision is a matter of explanatory coherence: you accept or reject beliefs on the basis of how well they fit with your full set of observations and explanations (Harman, 1986; Thagard, 2000). The process of assessing explanatory coherence operates unconsciously and in parallel, so we have no conscious access to it. All that comes to consciousness is a feeling that a belief seems to fit, which is part of a positive emotion that things make sense. On the other hand, if a belief does not fit, we get a negative feeling about it or about the overall state of our belief system. Coherence and acceptance feel good, whereas incoherence and doubt are irritating. Hence emotions are relevant to the development of knowledge with respect to acceptance as well as investigation and discovery. The emotional character of belief acceptance and rejection is especially evident in ethics, where utterance of an appealing principle, such as that it is moral to help the poor, generates positive emotion. In contrast, utterance of a dubious principle, such as that it is moral to torture suspected criminals, may be met with negative emotions, perhaps even anger or disgust.

The contributions of emotions to epistemology are often positive, as in their encouragement of investigation and coherence, but there are also ways in which emotions can skew beliefs. Thagard (forthcoming-d) discusses the "affective afflictions", which are systematic ways in which belief fixation is biased by emotions. These include wishful thinking and motivated inference, in which people's inferences are shaped by their personal goals rather than consideration of all the evidence. Another affective affliction is self-deception, which also involves people being unable to see why they are

making the inferences that they do. Conflicts of interest in government and business often distort inferences through motivated inference and self-deception.

A third way in which emotions are relevant to epistemology concerns knowledge about the emotional states of ourselves and others. Such knowledge is crucial to the functioning of human organizations from families to governments, all of which depend on emotional intelligence (Goleman, 1995). Some of this knowledge is constituted by explicit beliefs, for example when I infer that a friend is angry about something. But knowledge about the emotions of others can also be implicit, as when I use empathy to appreciate nonverbally my friend's emotional state by perceiving the physical manifestations of distress (Thagard, forthcoming-c). An account of the interrelations of cognition and emotion should illuminate the nature of cognitions about emotions, as well as the contributions of emotions to scientific thinking and belief acceptance. I will now argue that traditional analytic epistemology based on propositional attitudes is unsuited for understanding the epistemological significance of emotions.

AGAINST PROPOSITIONAL ATTIDUDES

I will not attempt to review the vast philosophical literature on propositional attitudes (see e.g. Fitch, 2005; Gale, 1967; Iacona, 2003; King, 2006; Richard, 1983; Salmon and Soames, 1988). There are different theories about what propositions are, but the most common view is that they must be invoked as the meanings of sentences. When two sentences are synonymous, it is because they express the same proposition, which is an abstract entity independent of any uttered sentence. Mental states such as belief and desire are described by sentences with *that* clauses: I believe that today is Monday, and I desire that I will have a productive week. Hence my belief is a propositional attitude, a

relation between me and the proposition that today is Monday. When we say that someone believes or desires that P, a proposition is the reference of the *that* clause. Emotions are also propositional attitudes, as when I am happy that I have written several pages today, but frustrated that I have not had time to organize my course for next term. My actions are explained by my propositional attitudes such as beliefs, desires, and emotions. For example, when I believe that an action is the best way to satisfy a strong desire, I perform that action.

Part of the appeal of the propositional attitude doctrine is that it fits well with our ordinary way of talking: we do utter sentences about beliefs, desires, and emotions that contain *that* clauses. But there is no reason to assume that this way of talking is scientifically or metaphysically justified. People commonly talk as if the sun rises, as if what goes up must come down, and as if whatever happens is God's will. What evidence is there that mental states are propositional attitudes?

From a scientific perspective, there is no evidence and there cannot be, because abstract entities cannot contribute to explanations. My argument has the following structure:

1. Theoretical entities are justified only by inference to the best explanation.
2. Explanation requires a causal link between entities and what they explain.
3. Abstract entities have no causal links with observable phenomena.
4. Therefore, belief in abstract entities is not justified.

Each of these steps requires amplification and defense.

In science, we have good grounds for believing in the existence of theoretical entities such as atoms, electrons, forces, viruses, and genes. Justification does not come

from direct observation, the way we are justified in believing in the existence of dogs and tables, because theoretical entities are not directly observable. Fortunately, human cognition has the power and flexibility to go beyond our limited perceptual apparatus and find good reasons for postulating things that we cannot perceive. The form of this reasoning is *inference to the best explanation:* we infer that an entity exists because the hypothesis that it exists is part of the best explanation of what we know through observation and experiments (see Harman, 1986; Lipton, 2004; Thagard, 1988, 1992). There is ample historical evidence that inference to the best explanation is ubiquitous in science and everyday life. This kind of inference is highly fallible, but still provides us with an often reliable way of going beyond the information given to our senses (Thagard, forthcoming-a).

But what is explanation? If explanation is a logical relation, as maintained by the hypothetico-deductive model of Hempel (1965), then there might be hope for the doctrine of propositional attitudes. We might be able to generate deductive explanations something like the following:

1. Whenever a person P has a desire that D and a belief B that an action A is the best way to accomplish D, then P does A.

2. Paul desires to go to the baseball game and believes that buying a ticket is the best way to go.

So 3. Paul buys a ticket.

Unfortunately, there are myriad reasons for doubting that deduction is necessary or sufficient for explanation (Salmon, 1989). It is not necessary, because there are good explanations in biology and other fields, for example employing Darwin's theory of

evolution by natural selection, that are not deductions. That deduction is not sufficient for explanation is shown by myriad examples where deduction is not explanatory, for example when one explains a man's non-pregnancy by deducing it from the facts that he took his wife's birth-control pills and that people taking birth control pills do not get pregnant.

An alternative account, more consistent with scientific practice, is that explanations are representations of causal mechanisms (Salmon, 1984; Bechtel and Abrahamsen, 2005). A mechanism is a system of objects related to each other in various ways including part-whole and spatial contiguity, such that the properties of the parts and the relations between them produce regular changes in the system. For example, modern medicine explains diseases in terms of the interactions of parts of the body such as organs and cells with external influences such as bacteria. The high fever and other symptoms of influenza are explained by causal interactions among viruses and body cells including the immune system. These interactions cause the symptoms of the disease, not in some abstract deductive sense, but in the physical sense that they make the symptoms happen (Woodward, 2004).

Now the problem with propositional attitudes becomes evident. Because propositions are abstract entities, they cannot be part of any physical mechanism, and so they cannot make anything happen. Hence hypotheses about propositions cannot be part of any explanation of observable facts, let alone part of the best explanation. Therefore, because inference to the best explanation is the only way to infer the existence of theoretical entities, there is no way that the existence of propositions can be justified.

We must accordingly conclude that propositions, construed as abstract entities constituting the meaning of sentences, do not exist.

It follows immediately that propositional *attitudes* do not exist, because there are no propositions for a self to have a relation to. Moreover, the notion of a self is metaphysically dubious, redolent as it is of the Cartesian soul. I have argued elsewhere that the hypothesis of the existence of a non-material soul is not part of the best explanation of human behavior (Thagard, 2000, ch. 4). Perhaps there is some other more plausible conception of the self that could stand in some relation to abstract propositions, but I do not know what it would be. A better course is to develop a more empirically supportable conception of the self as part of a neurophysiological account of cognition and emotion (Metzinger, 2003).

My argument would be undercut if there were some way to justify the existence of propositions besides inference to the best explanation, perhaps by direct acquaintance. A proponent of direct acquaintance would have to explain, however, how it is that people manage to have this kind of direct Platonic connection with abstract entities. Intuition should be psychologically explainable, not utterly mysterious. An alternative attempt to save propositional attitudes would be to have a much looser notion of explanation, perhaps some vague intuitive notion of making sense. But this would run counter to our best examples of successful explanations, those found in science. Even if it could be argued that propositions and propositional attitudes furnish some sort of explanation of our mental life and semantic capabilities, it would remain to be shown that they are part of the best explanation, and I will try to provide a provide a better account below.

Another way of attempting to undercut my explanation-based argument against propositional attitudes would be to invoke the view that explanations are just answers to questions (van Fraassen, 1980). It could then be argued that hypotheses about propositions and propositional attitudes answer many questions about meaning and mental states, so we should count them as explanatory. My first response to this argument is that for scientific and philosophical purposes it is crucial to spell out what counts as an answer, and there is ample evidence that in contemporary biology and psychology the appropriate kinds of answers are provided by causal mechanisms (Bechtel and Abrahamsen, 2005). My second response is that the invocation of van Fraassen's pragmatic approach to explanation should give scant support to proponents of the existence of propositions and propositional attitudes, because the point of that approach is to avoid making commitments to the existence of non-observable entities. From van Fraassen's perspective, the most one could say is that theories about propositions are empirically adequate. I will argue below that there are alternative theories of mental states that are much more successful at answering empirical questions.

My rejection of propositions and propositional attitudes does not involve denial of the existence of beliefs and desires. I have given reasons to doubt the propositional-attitude understanding of mental states, but do not deny that there are mental states that are something like beliefs and desires. The folk psychology of mental states may be approximately true, even if the philosophical understanding of folk psychology as involving propositional attitudes is false. I am far from the first to challenge the explanatory usefulness of construing propositions as abstract entities: see the nominalism of Ockham (Panaccio, 2004), the behaviorism of Quine (1960), and the eliminative

materialism of Churchland (1989). My main contribution here will be positive rather than negative, to describe a neurophysiological account that integrates cognition and emotion.

BELIEFS AS NEURAL ACTIVITY

In current cognitive psychology, the term *proposition* is used to refer, not to an abstract entity, but to a sentence-like mental representation (e.g. Anderson, 1983; Kintsch, 1998). Such representations are structured, consisting of simpler representations of objects and concepts. For example, the proposition that Jennifer loves Vince is a mental representation constructed from representations for Jennifer, Vince, and the relation *loves*. In the early days of cognitive psychology, it was tempting to think of mental representations as akin to those in natural language, part of a language of thought. Only recently has it become evident how mental representations are at bottom neural.

It is still fashionable to deride the notion of a grandmother neuron, a nerve cell that fires just when you are thinking of our grandmother. But Quiroga et al. (2005) report the finding in human brains of individual neurons tuned to fire when a person sees pictures of well known personalities such as Jennifer Aniston. Such neurons were found by recording the electrical activity of individual neurons in patients undergoing brain surgery for epilepsy. No one would claim, however, that any one neuron is *the* representation of Aniston, because other neurons must fire to correspond to aspects of her other than facial appearance, and even the particular neurons that fire when her face is presented occasionally fire when similar actors are presented.

Accordingly, neural representations are best thought of, not as individual neurons, but as groups of neurons, often called neural populations. The approximately 100 billion neurons in the brain are organized into large areas such as the prefrontal cortex, which are in turn organized into populations of neurons with many interconnections with each other. Representations in neural populations are highly distributed, in that the firing patterns in a population may represent numerous objects and concepts, and each object and concept may be represented by many neurons in the neural population. As an approximation, we can describe each neuron as having a firing rate characterized by how frequently it fires given a stimulus. Taking this rate as a number, we can represent the rates of firing of a population of 100 neurons by a list of 100 numbers, which is called a 100-dimensional vector. For example, the vector (.5, .8, .1, …) represents a population of neurons whose response to a stimulus involves firing .5 of the maximum rate, .8 of the maximum rate, and so on.

However, the encoding power of neurons is not limited to rates of firing, but can also involve patterns of firing and not firing, also known as *spike codes*. Consider the following simple patterns of neural firing: (1) fire, rest, fire, rest; (2) fire, fire, rest, rest. Both patterns involve firing half the time, but they are clearly different in that (1) alternatives firing and resting whereas (2) fires twice and then rests twice. There is neurological evidence that brains encode information using spike codes, which can also be shown to be more computationally powerful than only using rates of firing (Rieke et al., 1997; Maass and Bishop, 1999). It is therefore plausible to conclude that neural populations represent objects by patterns of firing that employ spike codes. A neural population can represent an object when exposure to stimuli adjusts the excitatory and

inhibitory connections between its neurons in such a way that the population is tuned to the object, in that its neurons exhibit a specific pattern of firing in response to the object (see Eliasmith and Anderson, 2003). The same population of neurons may also represent a similar object by having a different pattern of firing in response to it.

Concepts that represent classes of objects can be encoded in the same way. A neural population is tuned to a class of objects if it has a pattern of firing that occurs when presented with one of the objects in the class. Hence the neural representation for a concept is structurally the same as the neural representation for an object, although the pattern of firing has to be flexible enough to respond to an indefinite number of similar objects. Of course, the neural pattern for an object has to be flexible too, since the object is not always presented in the same way. For example, the neural population for Jennifer Aniston has to be able to respond to different pictures as well as verbal and auditory inputs. Tuning of a neural population to represent a concept or object is not simply a matter of tying sensory experience to neural firing. If you have a neural representation of Jennifer Anniston, it should have a pattern of firing that occurs when you hear about her or imagine her, not just when you see a picture. Hence the pattern of firing in a neural population that constitutes the representation of an object or concept should be responsive to a variety of connected populations, not just ones that involve sensory stimuli. A neural theory of representation does not have to be behaviorist.

If objects and concepts have neural representations, what about beliefs such as that Jennifer loves Vince? How can the brain combine patterns of firing in neural populations for Jennifer, loves, and Vince into something that represents the state of Jennifer loving Vince? Fodor and Pylyshyn (1988) criticized early connectionist

(artificial neural network) models for their representational inadequacy: it was not clear how a neural representation could capture the crucial difference between Jennifer loving Vince and Vince loving Jennifer. However, there are now a number of effective techniques for encoding complex relational information in neural networks, including the tensor products of Smolensky (1990) and the holographic reduced representations of Plate (2003); see also Eliasmith and Thagard (2001) and Smolensky and Legendre (2006). I will avoid the highly technical details, but describe the general approach.

Suppose that we have patterns of firing, not only for Jennifer, Vince, and *loves*, but also for the agent role and recipient role of the relation *loves*. We can then generate a new pattern of firing that is tuned to the combination of the pattern for *loves* with two other combinations: the pattern for Jennifer integrated with the pattern for love-agent, and the pattern for Vince integrated with the pattern for love-recipient. The resulting pattern of firing is then a representation that Jennifer loves Vince. It is still unknown exactly how the brain encodes objects and relations and combines them into neural patterns as complex as sentences, but tensor products and holographic reduced representations provide two examples of mathematical techniques that show how it is possible to build complex representations out of building blocks as simple as neurons. These techniques are adequate for capturing additional syntactic complexity, making possible the encoding of sentences such as "Because Jennifer loves Vince, Jennifer struck the photographer" and even "Jennifer believes that Vince loves her."

The feasibility of construction of sentences out of neural activity legitimizes the claim that beliefs are neural activity: my belief that Jennifer loves Vince is a pattern of firing in my neural populations. I can still have this belief even if I am not at the moment

thinking about them, because the synaptic connections between the neurons in the relevant populations are such that they will generate the relevant firing pattern when required, for example if I am asked whether Jennifer loves Vince. Thus occurrent beliefs are patterns of neural firing, but dispositional beliefs are structures consisting of neural connections that lead to the patterns of firing that constitute occurrent beliefs.

But what makes that pattern of firing the particular belief *that* Jennifer loves Vince? The answer is partly in the combinatorics by which the pattern of firing for this beliefs gets built out of the patterns of firing for the relevant objects and concepts. We need an account of how the neural population succeeds in representing Jennifer, which presupposes a theory of neurosemantics. Such theories have been offered by Eliasmith (2005) and Ryder (2004). I think that the core of the answer is that the neural activity that represents Jennifer has a complex kind of causal correlation with Jennifer herself, enabling the neural population to represent Jennifer. Defending this account of neurosemantics would distract this chapter from its main goal, to show how cognition relates to emotion. Hence I now turn to an account of desires.

DESIRES AS NEURAL ACTIVITY

To develop a neural account of desires, we can begin with the reward-based theory of desire proposed by Schroeder (2004, p. 131): "To have an intrinsic (positive) desire that P is to use the capacity to perceptually or cognitively represent that P to constitute P as a reward." Schroeder's view is superior to previous philosophical views that tie desire to behavioral dispositions or the experience of pleasure, but it is exceedingly linguistic (Thagard, 2006-a). He only discusses desires *that* something be the case, which ignores human and animal examples where the object of desire is a thing,

as when Jennifer desires a beer. Schroeder plausibly ties his theory to rapidly increasing understanding of the nature of reward in humans and motivation in humans and other animals, but restriction of desires to propositional contents is biologically implausible. It also has the difficulties associated with the notion of a proposition that I described in the above section on propositional attitudes.

Fortunately, there is an available alternative consistent with Schroeder's basic claims about the relation of desire to reward. Thagard (2006-a, p. 153) rephrased Schroeder while avoiding the *that* clause: "To have an intrinsic (positive) desire for Y is to use the capacity to perceptually or cognitively represent Y to constitute Y as a reward." Y can still be a state of affairs if the desire is best described as a desire that something be the case, but it can also be a thing. Now we can put to work the account of neural representation that was sketched in the last section. Things and states of affairs that concern them are represented occurrently by patterns of neural activity, i.e. patterns of firing in neural populations, and dispositionally by neural structures (neurons and their synaptic connections) that generate those patterns of neural activities. Thus a desire is a pattern of neural activity that ties a representation of a thing or state of affairs with a representation of reward. Understanding reward requires a neurocomputational theory of reward processing that describes how brain areas such as the nucleus accumbens and the amygdala attach positive and negative evaluations to representations. I will now sketch such a theory as part of a neural theory of emotions.

EMOTIONS AS NEURAL ACTIVITY

The neural theories of belief and desire presented here complement the neural theory of emotions that I have presented in detail elsewhere and will here only sketch

(see Thagard, forthcoming-a). Philosophical and psychological theories of emotion have fallen into two camps, cognitive and somatic. Cognitive theories view emotions as appraisals, for example when you are happy that you are going to a good restaurant because it will help satisfy your goal of eating well. The view that emotions are cognitive judgments about a person's situation has been advocated by philosophers such as Nussbaum (2001) and by psychologists such as Oatley (1992), and Scherer, Schorr, and Johnstone (2001). In contrast, somatic theories that understand emotions as perceptions of bodily states have been advocated by psychologists such as James (1884), by philosophers such as Prinz (2004), and by neuroscientists such as Damasio (1994). Proponents of cognitive and somatic theories have offered many competing arguments, but I will not review them here because I think that the theories are complementary.

How they complement each other is shown in the EMOCON model that I have proposed as part of a neurocomputational theory of emotional consciousness (Thagard, forthcoming-b). Figure 1 sketches interrelations among some of brain areas that are most important for emotional experience. These include:

Thalamus: processes sensory information for transmission to cortical areas.

Amygdala: is involved in fear and other emotional evaluations.

Insula: represents bodily states, especially ones linked to disgust and pain.

Dopamine system: includes the nucleus accumbens, the ventral tegmental area, and other circuits that process positive rewards.

Anterior cingulate: is part of the cortex with functions such as error detection and modulation of emotional responses.

Dorsolateral prefrontal cortex: is part of the cortex at the front/top/sides, involved in executive processes and working memory.

Orbitofrontal prefrontal cortex: is part of the cortex at the lower front, involved in processing rewards.

Ventromedial prefrontal cortex: part of the cortex at the bottom-middle, providing connections between the cortex and the amygdala.

This is by no means an exhaustive list of brain areas involved in emotion, but does include some of the main areas relevant to somatic and cognitive processing.

In figure 1, the arrows represent some of the main causal influences among different brain areas, indicating that neural activity in one area causes neural activity in another. For example, suppose you are presented with the external stimulus of a car skidding towards you. This stimulus affects your external sensors – eyes and ears – which send information to your thalamus. Neural activity in the thalamus affects, by means of synaptic connections, neural activity in the prefrontal cortex, but the thalamus also has direct effects on the amygdala and on bodily states such as heart rate. Internal sensors transfer information about somatic states to the amygdyla and insula, which also interact with the prefrontal cortex, which at the same time is conducting an appraisal of the significance of the stimulus, including assessment of possible rewards and threats.

Figure 1 should not be viewed as a serial flow chart, but rather as a causal map of how neural activity spreads in parallel through different brain areas. Your emotional response to the skidding car is the overall state of the brain that is produced by the parallel pathways of cognitive appraisal, involving the prefrontal cortex, anterior

cingulate, and dopamine system, and somatic perception, involving the amygdala and insula.

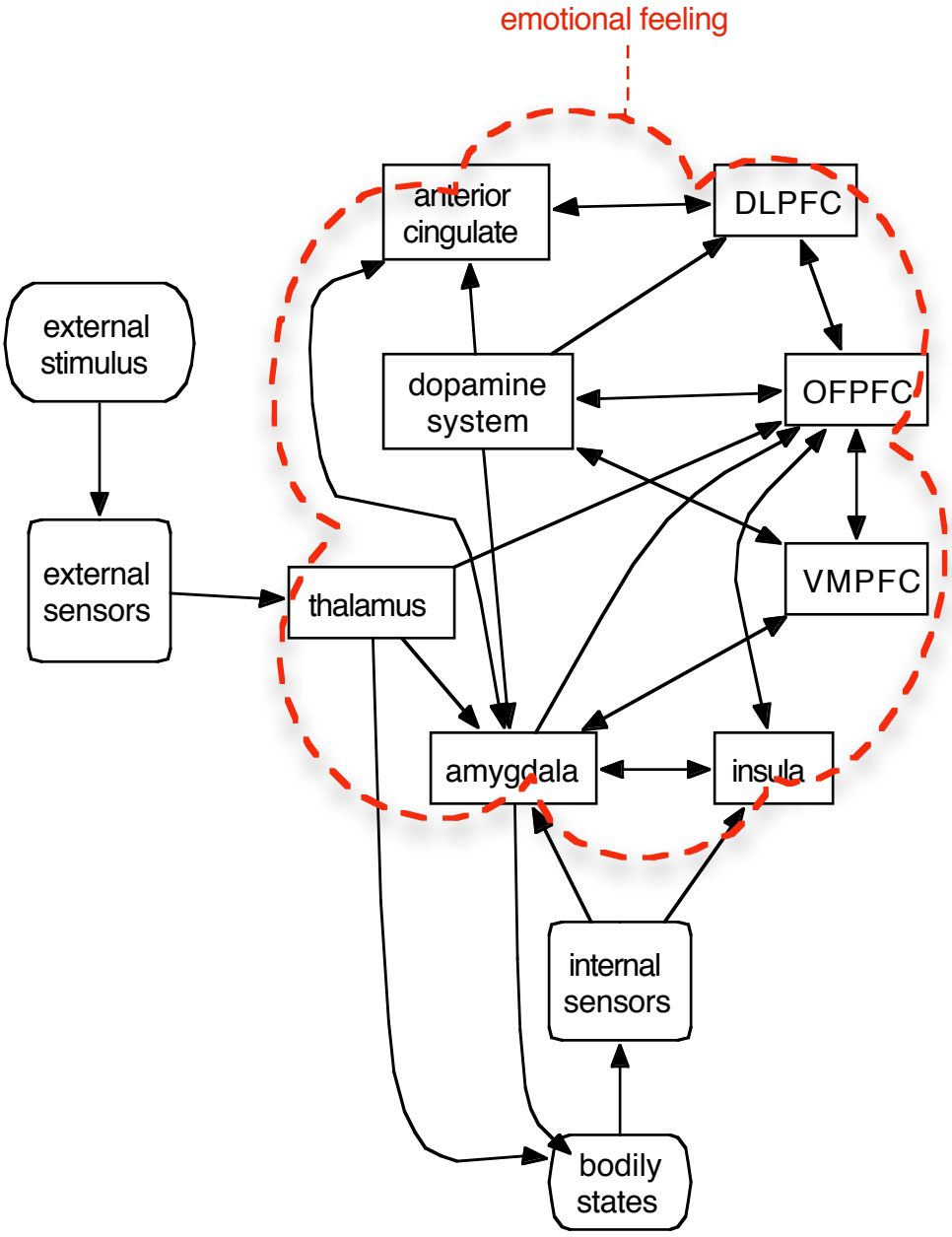

Figure 1. The EMOCON model of emotional consciousness, from Thagard (forthcoming-b). Abbreviations are PFC for prefrontal cortex, DL for dorsolateral, OF for orbitofrontal, and VM for ventromedial,

The EMOCON model shows how to integrate somatic and cognitive theories of emotion. The model includes somatic perception, with areas such as the amygdala and insula representing bodily states by receiving information from internals sensors. Again by representation I mean neural activity that is tuned in such a way that it causally correlates with something else, in this case bodily states such as blood pressure, heart rate, and blood levels of cortisol and glucose. However, human emotion is not just perception of bodily states, as it also involves appraisal performed by brain areas known to be involved in higher-level cognition, such as the dorsolateral prefrontal cortex. The involvement of cognitive appraisal in emotions is especially evident with the social emotions such as shame, guilt, embarrassment, and pride, all of which require a representation of how you are doing with respect to your place in a social group. But even basic emotions such as happiness involve appraisal of the degree to which one's goals are being accomplished. Thus the EMOCON model, based on well-known brain areas and connections, provides a way of seeing how emotions can be both cognitive appraisals and somatic perceptions.

I argue elsewhere that the EMOCON model can explain numerous aspects of emotion, including our ability to differentiate many emotions, our integration of cognition and emotion, and emotional change (Thagard, forthcoming-b). My primary concern here is understanding the role of emotions in epistemology. Emotions are patterns of neural activity involving multiple brain areas that perform both cognitive appraisal and bodily perception. This activity depends on extensive feedback connections among the relevant brain areas. Emotions involve representations of bodily

states and sensory inputs, and in sophisticated brains such as those possessed by humans they also involve representations of the world and of persons, including the person who has the emotion. This account of emotions as neural activity fits perfectly with my previous account of beliefs and desires as patterns of neural activity: the same kinds of representation underlie emotional states as underlie cognitive beliefs. Thus the common currency of beliefs, desires, and emotions is neural activity, not the abstract propositions assumed in the traditional philosophical account. As for beliefs, emotions have both occurrent and dispositional forms, where the latter understands them as neural structures that generate neural activity when stimulated. For example, Jennifer's loving Vince may be either neural activity that occurs when she is thinking about him, or it may be neural structures (neurons and synaptic connections) that dispose her toward that neural activity even when she is not thinking about him.

As we saw for desires, the objects of emotional states can be representations of objects and concepts as well as states of affairs. I may fear *that* the skidding car will hit me, but it is just as psychologically plausible to say that I fear the car itself, or that I fear dying. On the positive side, I can feel happy about Jennifer or about actors, as well as happy about the fact that Jennifer is an actor. My account of emotions as neural activity is compatible with associating them with representations of objects and concepts as well as states of affairs, which is another advantage over the narrow view that emotions are propositional attitudes.

IMPLICATIONS FOR EPISTEMOLOGY

The neural-activity view of mental states makes possible a deeper understanding of the positive and negative contributions of emotion to the development of knowledge.

The accounts of beliefs, desires, emotions, and consciousness sketched above make it clear that from a neural perspective there is no sharp distinction between cognition and emotion. Cognition involves a panoply of kinds of representations, not just sentence-like ones, and all of these can have associated emotions. For example, Fazio (2001) reviews evidence for the automatic activation of emotional attitudes attached to concepts. The emotional nature of concepts is most evident with extreme examples such as *cockroach* and *ice cream,* but other concepts can have more moderate emotional associations. Representations of objects and visual scenes are also often associated with emotional reactions. The brain has many interconnections, shown roughly in figure 1, between areas of the brain associated with emotional processing and areas associated with cognitive processing. Thus from a neural perspective, cognition and emotion are highly interrelated.

Hence it is not surprising that even the most high-level cognitive processes, such as the discovery and evaluation of scientific hypotheses, are highly emotional. Ambitious scientists have intense goals, which are not abstract aims but rather desires associated with reward-related parts of the brain that also are involved in positive emotions. Beliefs associated with the prospect of accomplishing cognitive and personal goals naturally have positive emotions associated with them. Failure to achieve these goals activates negative emotions involving feelings of sadness, worry, or even anger. Because mental representations generally have an affective component, and this component unavoidably influences how the representations are processed, emotions are highly relevant to epistemic progress.

From the perspective of traditional epistemology, considering the role of emotions in the development of knowledge might seem to conflate the descriptive with the normative. But naturalistic epistemology sees the need to base normative considerations in part on how human thinking actually works. Quine (1986, pp. 665-665) writes:

> Naturalization of epistemology does not jettison the normative and settle for the indiscriminate description of ongoing procedures. For me, normative epistemology is a branch of engineering. It is the technology of truth-seeking, or, in more cautiously epistemological terms, prediction. Like any technology, it makes free use of whatever scientific findings may suit its purpose. It draws upon mathematics in computing standard deviation and probable error and in scouting the gambler's fallacy. It draws upon experimental psychology in exposing perceptual illusions, and upon cognitive psychology in scouting wishful thinking. It draws upon neurology and physics, in a general way, in discounting testimony from occult or parapsychological sources. There is no question here of ultimate value, as in morals; it is a matter of efficacy for an ulterior end, truth or prediction. The normative here, as elsewhere in engineering, becomes descriptive when the terminal parameter is expressed.

Because cognitive psychology and neuroscience provide ample evidence of the role of emotions in cognition, naturalistic epistemology cannot afford to ignore how emotions affect the development of knowledge.

Even acceptance of hypotheses, not just their discovery, has an emotional component. This component would be avoidable if people were Bayesian engines,

making decisions about acceptance and rejection based on mathematical probabilities. But we rarely have the information that would allow us to be Bayesians: in scientific, legal, and other real life situations, the relevant conditional probabilities just are not known, and we lack the resources to acquire them (Thagard, 2000, 2004). Instead, we rely on a less rigorous but still effective method of accepting hypotheses if they cohere with the overall best explanation we can give. Simple neural networks can be used to represent both the maximization of coherence and the emotional inputs and outputs that can influence this process (Thagard, 2000, 2006).

Recently, Aubie and Thagard (forthcoming) have shown how coherence can be computed in more neurally realistic networks that better approximate to how brains work. We describe NECO (Neural Engineering Coherence), a model that uses populations of spiking neurons to represent the acceptability of hypotheses and evidence. This model is distributed in that acceptability of a hypothesis is represented by the activity of a whole group of neurons, and each neuron is involved in representing the acceptability of multiple hypotheses. NECO also models part of the emotional side of hypothesis evaluation, in that it ties the neural representation of a hypothesis to activity in populations of neurons intended to correspond to human emotional areas such as the amygdala. Thus inference to the best explanation, implemented as a neural process of computing explanatory coherence, can be understood at the neural level as closely tied to emotional processing. Theory evaluation can have emotional inputs consisting of the positive and negative attitudes associated with the concepts and beliefs that constitute hypotheses. It can also have emotional outputs, including the appropriately good feeling

we get when we achieve coherence and the annoyingly bad feeling we get when a proposed theory does not fit with our other beliefs.

There is also a negative side to the close connection in this model between cognitive and emotional coherence. When we feel good about how well our beliefs fit together, there is no way for us to tell whether in fact the coherence is really a matter of the goodness of fit of hypotheses with the evidence, or instead a matter of goodness of fit of hypotheses with our personal goals. Am I accepting a theory T because it fits best with the evidence or because it is my own theory and thereby fits with my personal goal of becoming famous? The inability of individual scientists to tell introspectively whether their judgment is objective shows the need for naturalistic epistemology to be social. If I have the goal of achieving reliable knowledge not biased by my inevitable personal goals, then I ought to submit my ideas to peer review by other people who do not have the same personal goals. Only then can I be confident that my theory evaluation is not emotionally skewed by personal desires that are unavoidably as much a part of the neural assessment of explanatory coherence as are considerations of evidence.

Thus my neural account of mental states can explain the interactions of beliefs and emotions in situations where emotional contributions are negative as well as where they are positive. If we adopt Quine's engineering role for naturalistic epistemology, we have to work with the natural system that evolution has given us, which has emotions intertwined with cognitions. Just as civic engineers cannot build ideal bridges over ideal rivers, naturalistic epistemologists cannot prescribe ideal inference patterns for ideal representations such as propositions. Instead, we need to consider various cognitive and social means to amplify the positive epistemic effects of emotion, such as providing

motivation and energy for scientific progress. We also need to develop cognitive and social means to diminish the negative epistemic effects, such as motivated inference. For both ameliorative projects, it is crucial to understand the underlying psychological terrain, which consists of many different kinds of representations all entangled with emotional processing.

The entanglement of cognition and emotion is even more important for establishing norms for practical reasoning, which includes both decision making and ethical judgment. There is rapidly expanding evidence that both ordinary and ethical decisions are made by processes that have a large and ineliminable emotional component. Hence moral epistemology, concerned with the justification of ethical claims, can also benefit from the deeper kind of understanding of cognitive-emotional interactions that neural theories can provide.

Even theoretical reason has a practical side, in that no intelligent agent aims only to accumulate a vast number of trivial truths. It would be pointless to accumulate a potentially infinite number of truths such as that our solar system has fewer than 12 planets, fewer than 13 planets, fewer than 14 planets, and so on. Rather, scientists and people in general aim to accumulate *important* truths, ones that are relevant to their goals, which may include explanation as well as practical goals such as survival and success. A scientist, for example, may put enormous effort into pursuing a theory because of a combination of its potential for explanatory importance and promotion of the scientist's career. In animal brains, the value of actual and potential states of affairs is assessed by the emotional system, operating with the kinds of interactions shown in the EMOCON

model. Hence emotions are crucial to epistemology in that they can help to guide us toward the acquisition of important truths.

CONCLUSION

I have tried to show how construing beliefs, desires, and emotions as neural activity illuminates the relevance of emotions to epistemology. I argued that the traditional philosophical theory of propositional attitudes blocks rather than enhances understanding of mental states: the propositions that it invokes as abstract entities fail to explain interactions of beliefs and emotions. A traditionalist could respond that without propositions we have no way of talking about what is shared by two people who have the same beliefs. But the fact that my son Dan and I have the same height does not mean that there is an abstract thing – height – that we both share. Similarly, when he and I have the same belief that Waterloo is located in Ontario, it is not because we both have a relation to some dubious abstract entity, the proposition that Waterloo is in Ontario. Rather, the sameness of belief is the result of systematic similarities between our patterns of neural activity. These patterns should not be expected to be exactly the same, given that our life experiences have been different. But our experiences of Waterloo and Ontario have been sufficiently similar that we can estimate that there is enough in common between our neural representations of Waterloo and Ontario to attribute to us approximately the same belief. Height is approximate too, as Dan and I are not the same height if you worry about millimeters. Prescientifically, weight was thought of a property of objects, but Newton reconceived weight as a relation between objects that gravitationally attract each other. Similarly, we need to reconceive meanings as relations that neural representations have to each other and the world, not as things.

Given the complexity of brains, there is no prospect for giving strict conditions for identifying a particular belief, desire or emotion in one person, let alone saying precisely what constitutes the same mental state in different persons. Current brain scanning technology only permits identification of the joint activity of many thousands of neurons, and there is no immediate prospect of being able to say exactly what neural structures constitute my belief that Waterloo is in Ontario. But the neural-activity view of mental states does not imply solipsism, because there is ample scientific evidence that people have very similar brains that undergo very similar causal interactions with the world. Further advances in experimental and theoretical neuroscience can be expected to fill in many gaps in our understanding of the nature of neural representations, especially through computer models that simulate important kinds of inference.

Translation is also approximate. Philosophers have claimed that *Snow is white* and *La neige est blanche* express the same proposition and hence can constitute the same belief. From the neural perspective, the attribution of belief across languages is again approximate. We should not expect an exact correspondence between concepts of snow held by people in Canada and France, but just something close enough that we can attribute approximately the same belief to them. Similarly, the vaunted role of propositions as bearers of truth needs to give way to a looser relation between neural representations and the world. This relation will be more complex than the strict binary division of true and false, but such approximation is all that is needed for the epistemological purpose of marking a difference between belief and knowledge. Replacing strict notions of sameness of belief with the more approximate notions appropriate to neural activity also enables us to evade the notorious puzzles about

propositional attitudes raised by Frege and Kripke. It becomes not all mysterious how someone can believe that Venus is the evening star but not believe that Venus is the morning star.

There are thus philosophical as well as scientific advantages to replacing the doctrine of propositional attitudes by a view of mental states as patterns of neural activity. Only recently has there been enough experimental evidence and theoretical progress to make plausible the claim that beliefs, desires, and feelings are neural processes. A crucial part of this progress was development of plausible ideas about how neural representations corresponding to beliefs can be built out of neural representations of objects and concepts. It now becomes possible to describe in neural terms the interactions of beliefs with desires and emotional feelings, and to use the resulting physical system to explain how the development of knowledge depends on emotional as well as cognitive processes. Because cognition meets emotion in the brain, emotions can be integral to epistemology.

Acknowledgements. I am grateful to Ulvi Doguoglu and Benoit Hardy-Vallée for comments on an earlier draft. The Natural Sciences and Engineering Research Council of Canada provided research support.

REFERENCES

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

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

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

Churchland, P. M. (1989). *A neurocomputational perspective*. Cambridge, MA: MIT Press.

Damasio, A. R. (1994). *Descartes' error*. New York: G. P. Putnam's Sons.

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.

Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science, 25*, 245-286.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion, 15*, 115-141.

Fitch, G. (2005). Singular propositions. *Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. Retrieved March 2, 2007, from http://plato.stanford.edu/archives/sum2005/entries/propositions-singular/.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(3-81).

Gale, G. (1967). Propositions, judgments, sentences, and statements. In P. Edwards (Ed.), *Encyclopedia of philosophy* (Vol. 6, pp. 494-505). New York: Macmillan.

Goleman, D. (1995). *Emotional intelligence*. New York: Bantam.

Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MA: MIT Press/Bradford Books.

Hempel, C. G. (1965). *Aspects of scientific explanation*. New York: The Free Press.

Iacona, A. (2003). Are there propositions? *Erkenntnis, 58*, 325-351.

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

King, J. C. (2006). Structured propositions. *Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. Retrieved Feb. 27, 2007, from http://plato.stanford.edu/archives/fall2006/entries/propositions-structured/

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). London: Routledge.

Maass, W., & Bishop, C. M. (Eds.). (1999). *Pulsed neural networks*. Cambridge, MA: MIT Press.

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, Mass.: MIT 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.

Panaccio, C. (2004). *Ockham on concepts*. Aldershot: Ashgate.

Plate, T. (2003). *Holographic reduced representations*. Stanford: CSLI.

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

Quine, W. V. (1960). *Word and object*. Cambridge, MA: MIT Press.

Quine, W. V. (1986). Reply to Morton White. In L. E. Hahn & P. A. Schilpp (Eds.), *The philosophy of W. V. O. Quine* (pp. 663-665). La Salle: Open Court.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature, 435*(7045), 1102-1107.

Richard, M. (1990). *Propositional attitudes; An essay on thoughts and how we ascribe them*. Cambridge: Cambridge University Press.

Rieke, F., Warland, D., de Ruyter van Steveninick, R. R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Ryder, D. (2004). SINBAD neurosemantics: A theory of mental representation. *Mind & Language, 19*, 211-240.

Salmon, N., & Soames, S. (Eds.). (1988). *Propositions and attitudes*. Oxford: Oxford University Press.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world.* Princeton: Princeton University Press.

Salmon, W. C. (1989). Four decades of scientific explanation. In P. Kitcher & W. C. Salmon (Eds.), *Scientific explanation (Minnesota Studies in the Philosophy of Science., vol. XIII)* (pp. 3-219). Minneapolis: University of Minnesota Press.

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

Schroeder, T. (2004). *Three faces of desire*. Oxford: Oxford University Press.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence, 46*, 159-217.

Smolensky, P., & Legendre, G. (2006). *The harmonic mind*. Cambridge, MA: MIT Press.

Thagard, P. (1988). *Computational philosophy of science*. Cambridge, MA: MIT Press/Bradford Books.

Thagard, P. (1992). *Conceptual revolutions*. Princeton: Princeton University Press.

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

Thagard, P. (2004). Causal inference in legal decision making: Explanatory coherence vs. Bayesian networks. *Applied Artificial Intelligence, 18*, 231-249.

Thagard, P. (2006a). Desires are not propositional attitudes. *Dialogue: Canadian Philosophical Review, 45*, 151-156.

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

Thagard, P. (forthcoming-a). Coherence, truth, and the development of scientific knowledge. *Philosophy of Science*.

Thagard, P. (forthcoming-b). Emotional consciousness: A neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience.

Thagard, P. (forthcoming-c). I feel your pain: Mirror neurons, empathy, and moral motivation.

Thagard, P. (forthcoming-d). The moral psychology of conflicts of interest: Insights from affective neuroscience. *Journal of Applied Philosophy*.

van Fraassen, B. (1980). *The scientific image*. Oxford: Clarendon Press.

Woodward, J. (2004). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.