THE AHA! EXPERIENCE: CREATIVITY THROUGH EMERGENT BINDING IN NEURAL NETWORKS

Paul Thagard and Terry Stewart
University of Waterloo
Draft 6, September, 2009

Abstract

Many kinds of creativity result from combination of mental representations. This paper provides a computational account of how creative thinking can arise from combining neural patterns into ones that are potentially novel and useful. We defend the hypothesis that such combinations arise from mechanisms that bind together neural activity by a process of convolution, a mathematical operation that interweaves structures. We describe computer simulations that show the feasibility of using convolution to produce emergent patterns of neural activity of the sort that can support human creativity.

CREATIVE COGNITION

Creativity is evident in many human activities that generate new and useful ideas, including scientific discovery, technological invention, social innovation and artistic imagination. Understanding is still lacking of the cognitive mechanisms that enable people to be creative, especially about the neural mechanisms that support creativity in the brain. How do people's brains come up with new ideas, theories, technologies, organizations, and aesthetic accomplishments? What neural processes underlie the wonderful AHA! experiences that creative people sometimes enjoy?

We propose that all human creativity requires the combination of previously unconnected mental representations constituted by patterns of neural activity. Then creative thinking is a matter of combining neural patterns into ones that are both novel and useful. We advocate the hypothesis that such combinations arise from mechanisms

that bind together neural patterns by a process of convolution rather than synchronization, which is the currently favored way of understanding binding in neural networks. We describe computer simulations that show the feasibility of using convolution to produce emergent patterns of neural activity of the sort that can support human creativity.

One of the advantages of thinking of creativity in terms of neural representations is that they are not limited to the sort of verbal and mathematical representations that have been used in most computational, psychological, and philosophical models of scientific discovery. In addition to words and other linguistic structures, the creative mind can employ a full range of sensory modalities derived from sight, hearing, touch, smell, taste and motor control. Creative thought also has vital emotional components, including the reaction of pleasure that accompanies novel combinations in the treasured AHA! experience. The generation of new representations involves binding together previously unconnected representations in ways that also generate new emotional bindings.

Before getting into neurocomputational details, we illustrate the claim that creative thinking consists of novel combination of representations with examples from science, technology, social innovation, and art. We then show how multimodal representations can be combined by binding in neural populations using a process of convolution in which neural activity is "twisted together" rather than synchronized. Emotional reactions to novel combinations can also involve convolution of patterns of neural activity. After comparing our neural theory of creativity with related work in cognitive science, we place it in the context of a broader account of multilevel mechanisms – including molecular, psychological, and social ones – that together

contribute to human creativity.

We propose a theory of creativity encapsulated in the following theses:

1. All creativity results from novel combinations of representations.

2. In humans, mental representations are patterns of neural activity.

3. Neural representations are multimodal, encompassing information that can be visual, auditory, tactile, olfactory, gustatory, kinesthetic, and emotional, as well as verbal.

4. Neural representations are combined by convolution, a kind of twisting together of existing representations.

5. The causes of creative activity reside not just in psychological and neural mechanisms, but also in social and molecular mechanisms.

Thesis 1, that creativity results from combination of representations, has been proposed by many writers, including Koestler (1967) and Boden (2004). Creative thinkers such as Einstein, Coleridge, and Poincaré have described their insights as resulting from combinatory play (Mednick, 1962). Thesis 2, that human mental representations are patterns of neural activity, is defended at length elsewhere (Thagard, 2010). The major thrust of this paper is to develop and defend thesis 4 by providing an account of how patterns of neural activity can be combined to constitute new ones. Our explanation of creativity in terms of neural combinations could be taken as an additional piece of evidence that thesis 2 is true, but we need to begin by providing some examples that provide anecdotal evidence for theses 1 and 3. Thesis 5, concerning the social and molecular causes of creativity, will be discussed only briefly.

CREATIVITY FROM COMBINATION OF REPRESENTATIONS

Fully defending thesis 1, that creativity results from combination of representations, would take a comprehensive survey of thousands of acknowledged instances of creative activity in many domains. We can only provide a couple of supporting examples from each of the four primary areas of human creativity: scientific discovery, technological invention, social innovation, and artistic imagination. We invite others to pose possibly refuting counterexamples to our slogan: No creation without representation combination.

Many scientific discoveries can be understood as instances of conceptual combination, in which new theoretical concepts arise by putting together old ones (Thagard, 1988). Two famous examples are the wave theory of sound, which required development of the novel concept of a sound wave, and Darwin's theory of evolution, which required development of the novel concept of natural selection. The concepts of sound and wave are part of everyday thinking concerning phenomena such as voice and water waves. The ancient Greek Chrysippus put them together in to create the novel representation of a sound wave that could explain many properties of sound such as propagation and echoing. Similarly, Darwin combined familiar ideas about selection done by breeders with the natural process of struggle for survival among animals to generate the mechanism of natural selection that could explain how species evolve.

One of the cognitive mechanisms of discovery is analogy, which requires putting together the representation of a target problem with the representation of a source (base) problem that furnishes a solution. Hence the many examples of scientific discoveries arising from analogy support the claim that creativity arises from combination of representations. See Holyoak and Thagard (1996) for a long list of analogy's greatest

successes in the field of scientific discovery.

Cognitive theories of conceptual combination have largely been restricted to verbal representations (e.g. Smith and Osherson, 1984; Costello and Keane, 2000), but conceptual combination can also involve perception (Wu and Barsalou, 2009; see also Barsalou et al., 2003). Obviously, the human concept of sound is not entirely verbal, possessing auditory exemplars such as music, thunder, and animal noises. Similarly, the concept of wave is not purely verbal, but involves in part visual representations of typical waves such as those in large bodies of water or even in smaller ones such as bathtubs. Hence a theory of conceptual combination, in general and in specific application to scientific discovery, needs to attend to non-verbal modalities.

Technological invention also has many examples of creativity arising from combination of representations. In our home town of Waterloo, Ontario, the major economic development of the past decade has been the dramatic rise of the company Research in Motion (RIM), maker of the extremely successful BlackBerry wireless device. The idea for this device originated in the 1990s as the result of the combination of two familiar technological concepts: electronic mail and wireless communication. According to Sweeny (2009), RIM did not originate this combination, which came from a Swedish company, Ericsson, where an executive combined the concepts of *wireless* and *email* into the concept of *wireless email*. Whereas the concept of sound wave was formed to explain observed phenomena, the concept of wireless email was generated to provide a new target for technological development and financial success (see Saunders and Thagard, 2005, for an account of creativity in computer science). Thus creative conceptual combination can produce representations of goals as well as of theoretical

entities. RIM's development of the BlackBerry depended on many subsequent creative combinations such as two-way paging, thumb-based typing, and an integrated single mailbox.

Another case of technological development by conceptual combination is the invention of the stethoscope, which came about by an analogical discovery in 1816 by a French physician, Théophile Laennec (Thagard, 1999, ch. 9). Unable to place his ear directly on the chest of a modest young woman with heart problems, Laennec happened to see some children listening to a pin scratching through a piece of wood, and came up with the idea of a *hearing tube* that he could place on the patient's chest. The original concepts here are multimodal, involving sound (hearing heartbeats) and vision (rolled tube). Putting these multimodal representations together enabled Laennec to create the concept we now call the stethoscope. It would be easy to document dozens of other examples of technological invention by representation combination.

Social innovations have been less investigated by historians and cognitive scientists than developments in science and technology (Mumford, 2002), but also result from representation combination. Two of the most important social innovations in human history are public education and universal health care. Both of these innovations required establishing new goals using existing concepts. Education and healthcare were private enterprises before social innovators projected the advantages for human welfare if the state took them on as a responsibility. Both innovations required novel combinations of existing concepts concerning government activity plus private concerns, generating the combined concepts of *public education* and *universal health care*. Many other social innovations, from universities, to public sanitation, to Facebook, can also be seen as

resulting from the creative establishment of goals through combinations of previously existing representations. The causes of social innovation are of course social as well as psychological, as we will make clear later when we propose a multilevel system view of creativity.

There are many kinds of artistic creativity, in domains as varied as literature, music, painting, sculpture, and dance. Individual creative works such as Beethoven's *Ninth Symphony*, Tolstoy's *War and Peace* , and Manet's *Le déjeuner sur l'herbe* are clearly the result of the cognitive efforts of composers, authors, and artists to combine many kinds of representations: verbal, auditory, visual, and so on. Beethoven's *Ninth,* for example, combines auditory originality with verbal novelty in the famous last movement known as the *Ode to Joy,* which also generates and integrates emotional representations. Hence illustrious cases of artistic imagination support the claim that creativity emanates in part from cognitive operations of representation combination.

Even kinesthetic creativity, the generation of novel forms of movement, can be understand as combination of representations as long as the latter are understood very broadly to include neural encodings of motor sequences (e.g. Wolpert and Ghahramani, 2000). Historically novel motor sequences include the slam dunk in basketball, the over-the-shoulder catch in baseball, the Statue of Liberty play in football, the bicycle kick in soccer, and the pas de deux in ballet. All of these can be described verbally and may have been generated using verbal concepts, but it is just as likely that they were conceived and executed using motor representations that can naturally be encoded in patterns of neural activity.

We are primarily interested in creativity as a mental process, but cannot neglect

the fact that it also often involves interaction with the world. Manet's innovative painting arose in part from his physical interaction with the brush, paint, and canvas. Similarly, invention of important technologies such as the stethoscope and the wheel can involve physical interactions with the world, as when Laennec rolled up a piece of paper to produce a hearing tube. External representations such as diagrams and equations on paper can also be useful in creative activities, as long as they interface with the internal mental representations that enable people to interact with the world.

We have provided examples to show that combination of representations is a crucial part of creativity in the domains of scientific discovery, technological invention, social innovation, and artistic imagination. Often these domains intersect, for example when the discovery of electromagnetism made possible the invention of the radio, and when the invention of the microscope made possible the discovery of the cell. Social innovations can involve both scientific discoveries and technological invention, as when public health is fostered by the germ theory of disease and the invention of antibiotics. Artistic imagination can be aided by technological advances, for example the invention of new musical instruments.

We hope that our examples make plausible the hypothesis that creativity requires the combination of mental representations operating with multiple modalities. We now will describe neural mechanisms for such combinations.

NEURAL COMBINATION AND BINDING

Combination of representations has usually been modeled with symbolic techniques common in the field of artificial intelligence. For example, concepts can be modeled by schema-like data structures called frames (Minsky, 1975), and combination

of concepts can be performed by amalgamating frames (Thagard, 1988). Other computational models of conceptual combination have aimed at modeling the results of psycholinguistic experiments rather than creativity, but they also take concepts to be symbolic, verbal structures (Costello and Keane, 2000). Rule combination has been modeled with simple rules consisting of strings of bits and genetic algorithms (Holland , Holyoak, Nisbett, and Thagard, 1986), and genetic algorithms have also been used to produce new combinations of expressions written in the programming language LISP (Koza, 1992). Lenat (1984) produced discovery programs that generated new LISP-defined concepts out of chunks of LISP code. Rule-based systems such as ACT and SOAR can also employ learning mechanisms in which rules are chunked or compiled together to form new rules (Anderson, 1990; Laird, Rosenbloom, and Newell, 1986).

These symbolic models of combination are powerful, but they lack the generality to handle the full range of representational combinations that include sensory and emotional information. Hence we propose viewing representation combination at the neural level, since all kinds of mental representations – concepts, rules, sensory encodings, and emotions - are produced in the brain by the activity of neurons. Evidence for this claim comes from the vast range of psychological phenomena such as perception and memory that are increasingly being explained by cognitive neuroscience (see e.g. Smith and Kosslyn, 2007; Chandrasekharan, 2009; Thagard, 2010).

The basic idea that neural representations are constituted by patterns of activity in populations of neurons dates back at least to Donald Hebb (1949, 1980), and is implicit in many more recent and detailed neurocomputational accounts (e.g. Rumelhart and McClelland, 1986; Churchland and Sejnowsky, 1992; Dayan and Abbott, 2001; Eliasmith

and Anderson, 2003). If this basic idea is right, then combination of representations should be a neural process involving generation of new patterns of activity from old ones.

Hebb today is largely remembered for the eponymous idea of learning of synaptic connections between neurons that are simultaneously active, which has its roots in the learning theories of eighteenth-century empiricist philosophers such as David Hartley. But Hebb's most seminal contribution was the doctrine that all thinking results from the activity of *cell assemblies*, which are groups of neurons organized by their synaptic connections and capable of generating complex behaviors.

Much later, Hebb (1980) sketched an account of creativity as a normal feature of cognitive activity resulting from the firing of neurons in cell assemblies. Hebb described problem solving as involving many "cell-assembly groups which fire and subside, fire and subside, fire and subside, till the crucial combination occurs" (Hebb 1980, p. 119). Combination produces a new scientific idea that sets off a new sequence of ideas and constitutes a different way of seeing the problem situation by reorienting the whole pattern of cortical activity. The new combination of ideas that result from connection of cell-assemblies forms a functional system that excites the arousal system, producing the Eureka! emotional effect. Thus Hebb sketched a neural explanation of the phenomenon of insight that has been much discussed by psychologists interested in problem solving (e.g. Bowden and Jung-Beeman, 2003; Sternberg and Davidson, 1995).

Hebb's conception of creative insight arising from cell assembly activity is suggestive, but rather vague, and raises as many questions than it answers. How are cell assemblies related to each other, and how is the information they carry combined? For example, if there is a group of cell assemblies (a neural population) that encodes the

concept of *sound*, and another that encodes the concept of *wave*, how does the combined activity of the overall neural population encode the novel conceptual combination of *sound wave?*

We view the problem of creative combination of representations as an instance of the ubiquitous *binding problem* that pervades cognitive neuroscience (e.g. Treisman, 1996; Roskies, 1999). This problem was first recognized in studies of perception, where it is problematic how the brain manages to integrate various features of an object into a unified representation. For example, when people see a stop sign, they see the color red, the octagonal shape, and the white letters as all part of the same image, which requires the brain to bind together what otherwise might be several disparate representations. The binding problem is also integral to explaining the nature of consciousness, which has a kind of unity that may seem mysterious from the perspective of the variegated activity of billions of neurons processing many different kinds of information.

The most prominent suggestions of how to deal with the binding problem have concerned synchronization of neural activity, which has been proposed as a way to deal with the kind of cognitive coordination that occurs in consciousness (Crick, 1994; Engel, 1999; Grandjean, Sander, and Scherer, 2008; Werner and Maye, 2007). At a more local level, neural synchrony has been proposed as a way of integrating crucial syntactic information needed for the representation of relations, for example to mark the difference between *Romeo loves Juliet* and *Juliet loves Romeo* (Shastri and Ajjanagadde, 1993; Hummel and Holyoak, 1997, 2003). If synchrony is the key to all sorts of binding in the brain, then it should be possible to develop an account of representation combination as synchronization of neural populations. In Hebb's terminology, the creative combination

of cell assemblies would results from synchronization of the firings of the neurons in the original assemblies.

There are neurophysiological and computational arguments that cast doubt on the plausibility of synchronization as the fundamental binding mechanism (Stewart and Eliasmith, 2009; Eliasmith and Stewart, forthcoming). But we will not pursue a critique here. Our aim in this paper is to develop an alternative account based on convolution rather than synchronization. We will not propose a general theory of binding, but rather defend a more narrow account of how the information encoded in patterns of neural activity gets combined into new representations that may turn out to be creative. We draw heavily on Eliasmith's (2004, 2005) work on neurobiologically plausible simulations of high-level inference.

BINDING BY CONVOLUTION

Tony Plate (2003) developed an alternative way of thinking about binding in neural networks, using vector-based representations similar to but more computationally efficient than the tensor-product representations proposed by Smolensky (1990). Our presentation of Plate's idea, which we call binding by convolution, will be largely metaphorical in this section, but technical details are provided later. Eliasmith and Thagard (2001) include a relatively gentle introduction to Plate's method.

To get the metaphors rolling, consider the process of braiding hair. Thousands of long strands of hair can be braided together by twisting them systematically into one or more braids. Similar twisting can be used to produce ropes and cables. Another word for twisting and coiling up things in this way is *convolve*, and things are convolved if they are all twisted up together. Another word for "convolve" is "convolute", but we

avoid this term because in recent decades the term "convoluted" has come to mean "excessively complicated."

In mathematics, a convolution is an integral function that expresses the amount of overlap of one function *f* as it is shifted over another function *g*, expressing the blending of one function with another (http://mathworld.wolfram.com/Convolution.html). This notion gives a mathematically precise counterpart to the physical process of braiding, as we can think of mathematical convolution as blending two signals together (each represented by a function) in a way roughly analogous to how braiding blends two strands of hair together.

Plate developed a technique he called *holographic reduced representations* that applies an analog of convolution to vectors of real numbers. It is natural to think of patterns of neural activity using vectors: if a neural population contains *n* neurons, then its activity can be represented by a sequence that contains *n* numbers, each of which stands for the firing rate of a neuron. For example, if the maximum firing rate of a neuron is 200 times per second, the rate of a neuron firing 100 times per second could be represented by the number .5. Then the vector (.5, .4, .3, .2, .1) corresponds to the firing rates of this neuron and 4 additional ones with slower firing rates.

So here is the basic idea: if we abstractly represent the pattern of activity of two neural populations by vectors A and B, then we can represent their combination by the mathematical convolution of A and B, which is another vector corresponding to a third pattern of neural activity. For the moment, we ignore what this amounts to physiologically – see the simulation section below. The resulting vector has emergent properties, i. e. properties not possessed by (or simple aggregates of) either of the two

vectors out of which it is combined. The convolved vector combines the information included in each of the originating vectors in a nonlinear fashion that enables only approximate reconstruction of them. Hence the convolution of vectors produces an emergent binding, one which is not simply the sum of the parts bound together (on emergence, see Bunge, 2003, and Wimsatt, 2007).

Talking about convolution of vectors still does not enable us to grasp the convolution of patterns of neural activity. To explain that, we will need to describe our simulation model of how combination can occur in a biologically realistic, computational neural network. Before getting into technical details, however, we need to describe how the AHA! experience can be understood as convolution of a novel combined representation with patterns of brain activity for emotion.

EMOTION AND CREATIVITY

Cognitive science must not only explain how representations get combined into creative new ones, it should also explain how such combinations can be intensely emotional. Here are some quotes from eminent scientists that attest to the emotional component of scientific discovery (for further discussion, see Thagard 2006, ch. 10). We expect that breakthroughs in technological invention, social innovation, and artistic imagination are just as exciting.

James Watson:

> As the clock went past midnight I was becoming more and more pleased. There had been far too many days when Francis and I worried that DNA structure might turn out to be superficially very dull, suggesting nothing about either its replication or its function in controlling cell biochemistry.

> But now, to my delight and amazement, the answer was turning out to be profoundly interesting. For over two hours I happily lay awake with pairs of adenine residues whirling in front of my closed eyes. Only for brief moments did the fear shoot through me that an idea this good could be wrong. (Watson 1969, p. 118)

Carl Djerassi:

> I'm absolutely convinced that the pleasure of a real scientific insight - it doesn't have to be a great discovery - is like an orgasm (Wolpert and Richards 1997, p. 12).

Gerard Edelman:

> After all, if you have been filling in the tedium of everyday existence by blundering around a lab, for a long time, and wondering how you're going to get the answer, and then something really glorious happens that you couldn't possibly have thought of, that has to be some kind of remarkable pleasure. In the sense that it's a surprise, but it's not too threatening, it is a pleasure in the same sense that you can make a baby laugh when you bring an object of nowhere. Breaking through, getting various insights is certainly one of the most beautiful aspects of scientific life. (Wolpert and Richards 1997, p. 137)

Carlo Rubbia:

> The act of discovery, the act of being confronted with a new phenomenon, is a very passionate and very exciting moment in everyone's life. It's the reward for many, many years of effort and, also, of failures (Wolpert and

Richards 1997, p. 197).

François Jacob:

> I had seen myself launched into research. Into discovery. And, above all, I had grasped the process. I had tasted the pleasure (Jacob, 1988, p. 196-197).
>
> These hypotheses, still rough, still vaguely outlined, poorly formulated, stir within me. Barely have I emerged that I feel invaded by an intense joy, a savage pleasure. A sense of strength, as well, of power (Jacob 1988, p. 298).

Richard Feynman:

> The prize is the pleasure of finding a thing out, the kick of the discovery, the observation that other people use it [my work]-those are the real things, the others are unreal to me (Feynman 1999, p. 12).

What are the neuropsychological causes of the "kick of the discovery"?

To answer that question, we need a neurocomputational theory of emotion that can be integrated with the account of representation generation provided earlier in this paper. Thagard and Aubie (2008) have hypothesized how emotional experience can arise from a complex neural process that integrates cognitive appraisal of a situation with perception of internal physiological states. Figure 1 shows the structure of the EMOCON model, with the interaction of multiple brain areas generating emotional consciousness, requiring both appraisal of the relevance of a situation to an agent's goals (performed by the prefrontal cortex and mid-brain dopamine system) and internal perception of physiological changes (performed by the amygdala and insula). For defense of the

neural, psychological, and philosophical plausibility of this account of emotional experience, see also Thagard (2010).

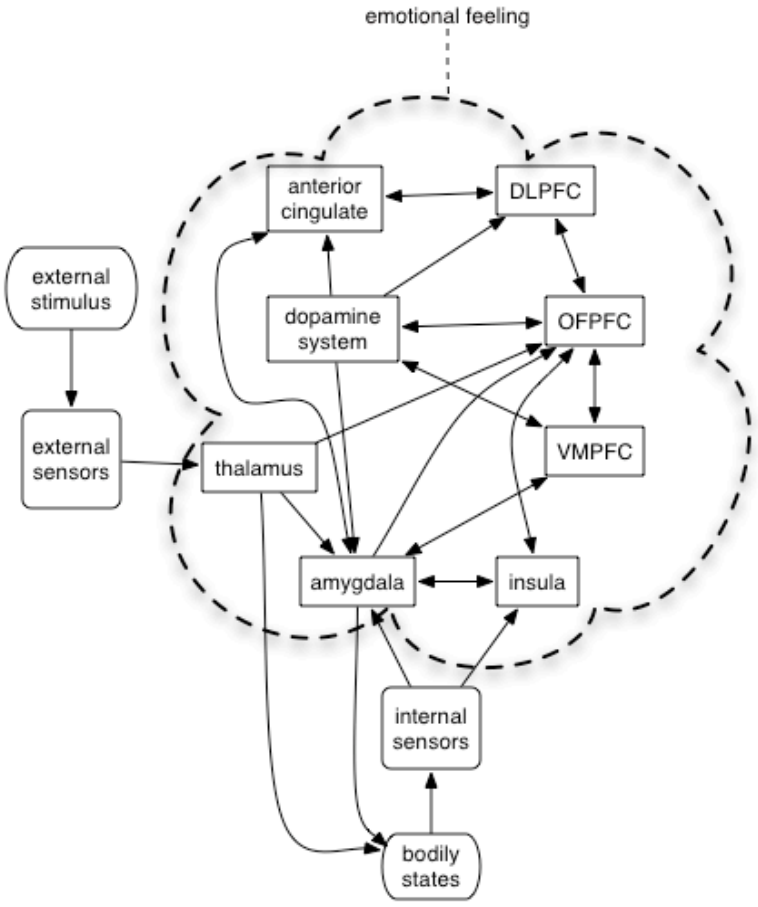

Figure 1. The EMOCON model of Thagard and Aubie (2008), which contains details and partial computational modeling. Abbreviations: DLPFC is dorsolateral prefrontal cortex. OFPFC is orbitofrontal prefrontal cortex. VMPFC is ventromedial prefrontal cortex. The dotted line is intended to indicate that emotional consciousness emerges from activity in the whole system.

If the EMOCON model is correct, then emotional experiences such as the ecstasy

of discovery are patterns of neural activity, just like other mental representations such as concepts and rules. Now it is easy to see how the AHA! or Eureka! experience can arise. When two representations are combined by convolution into a new one, the brain automatically performs evaluation of the relevance of the new representation to its goals. Ordinarily, such combinations are of little significance, as in the ephemeral conceptual combinations that take place in all language processing. There need be no emotional reaction to mundane combinations such as "brown cow" and "tall basketball player". But some combinations are surprising, such as "cow basketball" and may illicit further processing to try to make sense of them (Kunda, Miller, and Clarie, 1990).

In extraordinary situations, the novel combination may be not only surprising but actually exciting, if it has strong relevance to accomplishing the longstanding goals of the thinker. For example, Darwin was thrilled when he realized that the novel combination *natural selection* could explain facts about species that had long puzzled him, and the combination *wireless email* excited the inventors of the Blackberry when they realized its great commercial potential.

Figure 2 shows how representation combination can be intensely emotional, when patterns of neural activity corresponding to concepts become convolved with patterns of activity that constitute the emotional evaluation of the new combination. A new combination such as *sound wave* is exciting because it is highly relevant to accomplishing the discoverer's goals. But emotions are not just a purely cognitive process of appraisal, which could be performed dispassionately, as in a calculation of expected utility as performed by economists. AHA! is a very different experience from "Given the probabilities and expected payoffs, the expected value of option X is high."

Physiology is a key part of emotional experience – racing heartbeats, sweaty palms, etc., as pointed out by many theorists (e.g. James, 1894; Damasio, 1994; Prinz, 2004). But physiology cannot be the only part of the story, as there are many reasons to see cognitive appraisal as crucial too (Thagard, 2010). The EMOCON model shows how reactions can combine both cognitive appraisal and physiological perception. Hence we propose that the AHA! experience requires a triple convolution, binding: (1) two representations into an original one, (2) cognitive appraisal and physiological perception into a combined assessment of significance, and (3) the combined representation and the integrated cognitive/physiological emotional response into a unified representation (pattern of neural activity) of the creative representation and its emotional value.

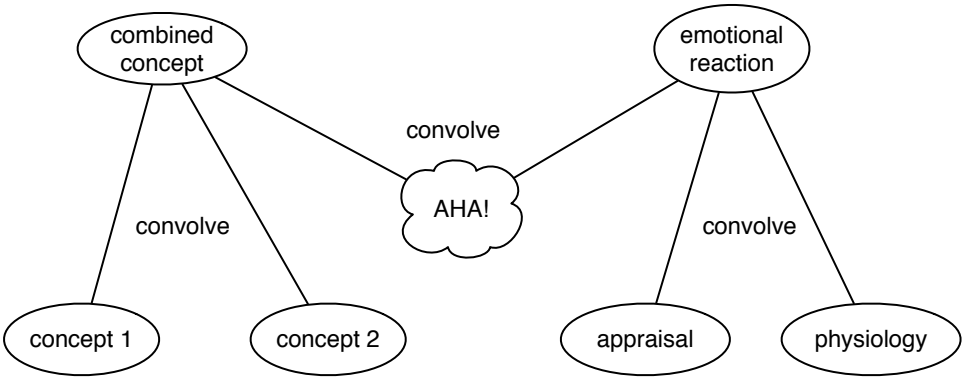

Figure 2. How the AHA! experience arises by multiple convolutions of representation combination and emotion.

Because the brain's operations are highly parallel, it would be a mistake to think of what we have just described as a serial process of first combining the representations, then integrating the appraisal and physiological perception, and finally convoluting the new representation and the emotional reaction. Rather, all these processes take place concurrently, with a constant flow of activation among the regions of the brain crucial for

different kinds of cognition, perception, and emotion. The AHA! experience seems mysterious because we have no conscious access to any of these processes or their integration. But figure 2 shows a possible mechanism for how the wonderful AHA! experience can emerge from neural activity.

SIMULATIONS

Our discussion of convolution so far has been largely metaphorical. We will now describe neurocomputational simulations that use the methods of Eliasmith (2004, 2005) to show: (1) how patterns of neural activity can represent vectors; (2) how convolution can bind together two representations to form a new one; and (3) how convolution can bind together a new combined representation with patterns of brain activity that correspond to emotional experience arising from a combination of cognitive appraisal and physiological perception.

Simulation 1: Neurocomputational model of visual patterns

The first requirement for a mechanistic neural explanation of conceptual combination is specifying how a pattern can be represented by a population of neurons. Earlier, we gave a simple, unrealistic example where a vector of five elements corresponded to the firing rates of five neurons; for example, the value .5 could indicate an average firing rate of 100Hz (i.e. 100 times per second). We now move away from thinking about how a vector can represent a set of neurons, as is common in connectionist models, to thinking about how a population of neurons can represent a vector. Linguistic representations can be translated into vectors using Plate's method of holographic reduced representation, and there are also natural translations of visual, auditory, and olfactory information into the mathematical form of vectors. So a general,

multimodal theory of representation combination need consider only how vectors can be neurally represented and combined.

Instead of using a single neuron for each value in a vector, we assume that values are represented by many neurons, so that the loss of a single neuron will not destroy the value being represented. For a visual example, think of a simple 5X5 grid, with a total of 25 pixels. This grid can be represented by a vector with 25 values (dimensions), each of which is a real number standing for the brightness of the pixel. Figure 3 shows a simple encoding of three different visual patterns using 100 artificial neurons for each value of the vector that represents the visual pattern. For each pixel in the three small input patterns on top of the diagram, there is a 10X10 array of neurons in the 2500-neuron networks shown in the large boxes below them.

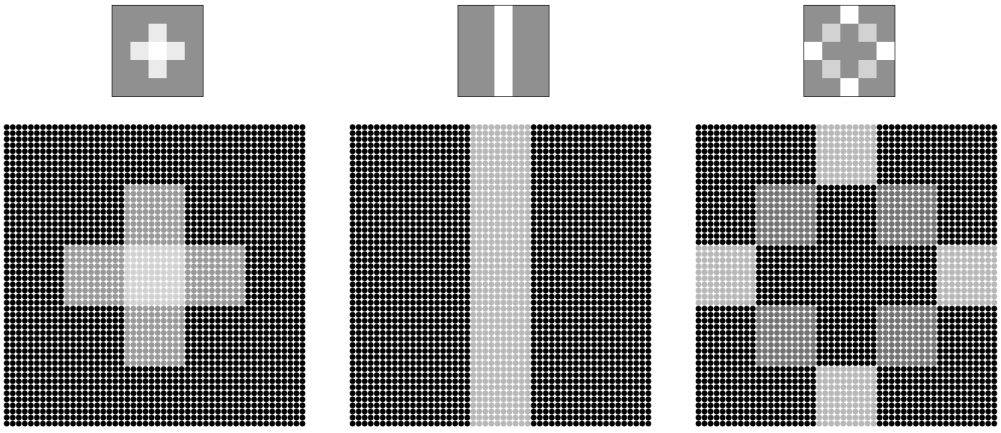

Figure 3. Simple neural network encoding using 100 neurons per pixel value, for each of three 25-dimensional vectors. Each small square on top is a 5X5 grid of pixels. In the large bottom squares, each circle is an individual neuron and the lighter shading indicates a faster firing rate. Encodings for three different 25-dimensional vectors are shown.

For improved neurological realism, instead of the strict one-to-one mapping between neurons and elements in the vector, we can introduce neurons that respond to nearby elements as well. This produces a smooth topological map, as seen in Figure 4. But such maps generally do not capture neuron firing patterns that are known to occur in the brain. For example, Georgopoulos, Scwhartz, and Kettner (1986) demonstrated that neurons in the motor cortex of monkeys encode reaching direction by each one having its own preferred direction vector. That is, for each neuron, there is a particular vector for which it fires most quickly, and the firing rate decreases as the vector represented becomes more dissimilar to that direction vector. This blurring results in an efficient distributed encoding of the vector across the set of neurons.

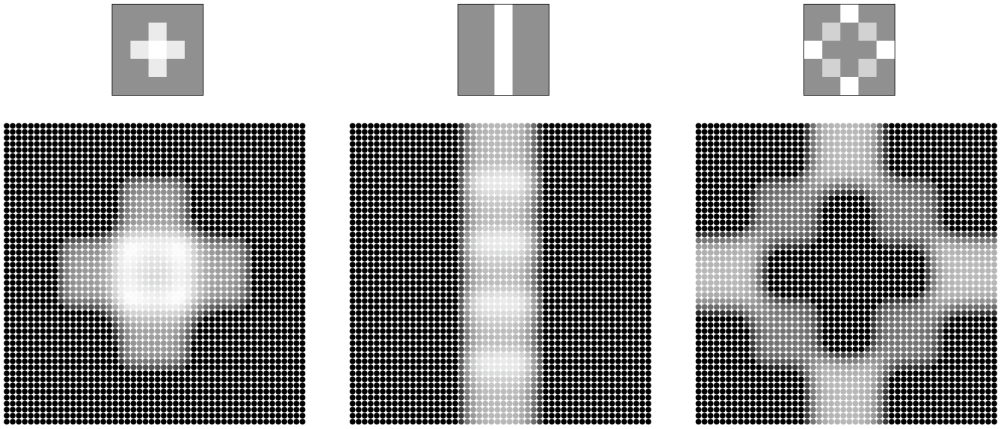

Figure 4. Encoding using 100 neurons per pixel value where each neuron responds to nearby elements in the vector. Each circle is an individual neuron and the lighter shading indicates a faster firing rate. Encodings for three different 25-dimensional vectors are shown.

This idea of neurons having preferred vectors can be directly extended to the higher-dimension patterns that we use in this paper, following Eliasmith and Anderson

(2003). For each neuron, we randomly generate a vector (pattern) that it prefers (fires most strongly in response to). The flow of current into the neuron is set to be proportional to the similarity between the represented vector and this preferred vector. To further constrain the model to be realistic, each neuron also has a random background current constantly flowing into it (so that it can fire even when there is no value being represented, albeit slowly) and a randomly chosen maximum firing rate in the range 100Hz to 200Hz.

To display these neurons, we can organize them so that neurons with similar preferred direction vectors are placed near each other, as is generally found throughout the brain. Using this approach, Figure 5 shows a group of 2500 neurons representing three different 25-dimensional vectors. For further details on the creation of these neuron populations, see the appendix.

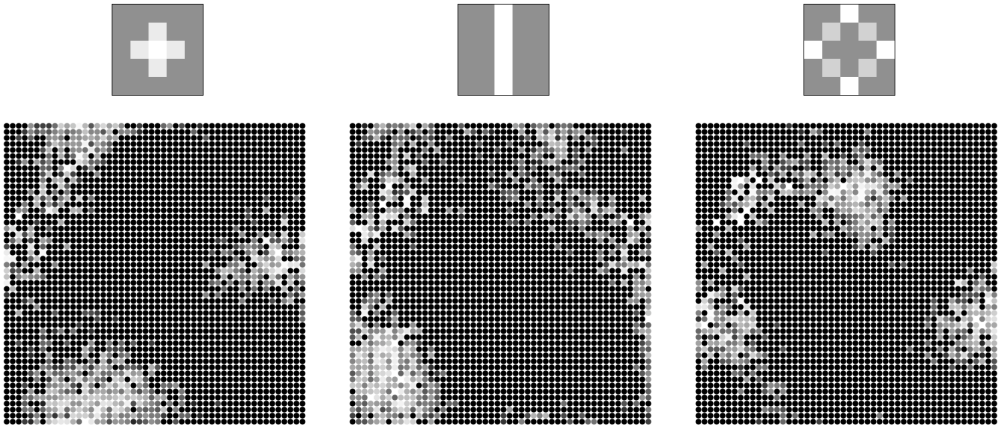

Figure 5. Distributed encoding using preferred direction vectors of a 25-dimensional vector with 2500 neurons. Each circle is an individual neuron and the lighter shading indicates a faster firing rate. Encodings for three different 25-dimensional vectors are shown.

Instead of just looking at the overall firing rate of each of these neurons over a long period of time, we can also examine neuron behavior at a single snapshot in time. In Figure 6 we show the neural behavior over a tenth of a millisecond. Here, the shading indicates the voltage of each neuron, with black indicating the neuron resting potential (-70mV) and white indicating that enough voltage has built up for the neuron to fire (around -45mV). When this voltage is reached that neuron fires, reducing its voltage back down to the resting level. The input current (due to the preferred direction vector) will then slowly increase that voltage until it fires again.

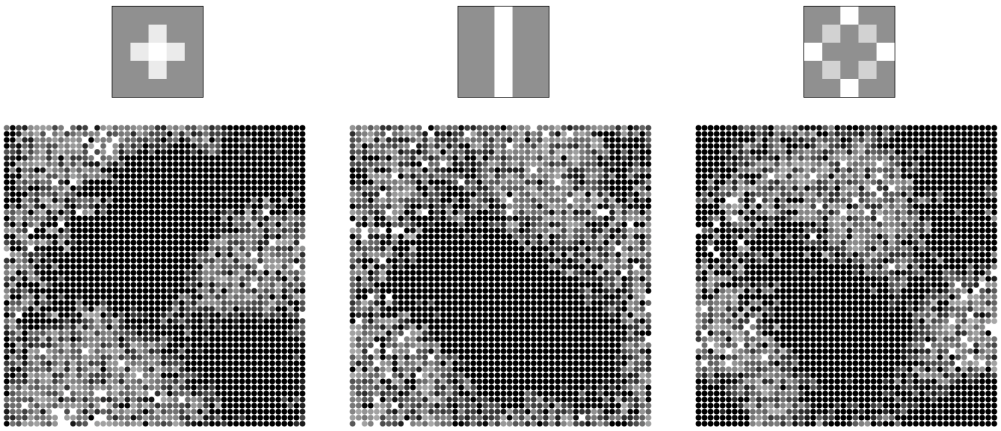

Figure 6. A snapshot in time showing 0.1ms of distributed encoding using preferred direction vectors of a 25-dimensional vector with 2500 neurons. Each circle is an individual neuron and the lighter shading indicates a higher membrane voltage. White circles indicate a neuron that is currently firing. Encodings for three different 25-dimensional vectors are shown.

Simulation 2: Convolution of Patterns

Now that we have specified how neurons can represent vectors, we can organize neurons to perform the convolution operation (Eliasmith, 2004). We need a neural model where we can provide two different patterns as input (using the representation scheme above), and the resulting neural firing will cause a third group of neurons to represent the convolution of the two original patterns.

The Neural Engineering Framework (NEF) of Eliasmith & Anderson (2003) provides a methodology for converting a function such as convolution into a neural model by deriving the synaptic connection weights that will implement that function. Once these weights are found, we use the representation method discussed above to encode the two original vectors in neural groups A and B. When these neurons fire, the synaptic connections cause current to flow into any neurons to which they are connected. This flow in turn causes firing in neural group C that represents the convolution of A and B. Details on the derivation of these synaptic connections can be found in the appendix. Figure 7 shows how convolution combines the neural representation of two perceptual inputs, on the left, into a neural representation of their convolution, on the right. It is clear from the visual interpretation of the neural representation on the right that the convolution of the two input patterns is not simply the sum of those patterns, and therefore amounts to an emergent binding of them.

Importantly, one set of neural connection weights is sufficient for performing the convolution of *any* two input vectors. That is, the synaptic connections do not need to be changed if we need to convolve two new patterns; all that has to change is the firing patterns of the neurons in groups A and B. We do not rely on a slow, learning process of changing synaptic weights: convolution is a fast process response to changes in

perceptual inputs. If the synaptic connections in our NEF model correspond to a fast neurotransmitter AMPA, convolution can occur within five milliseconds.

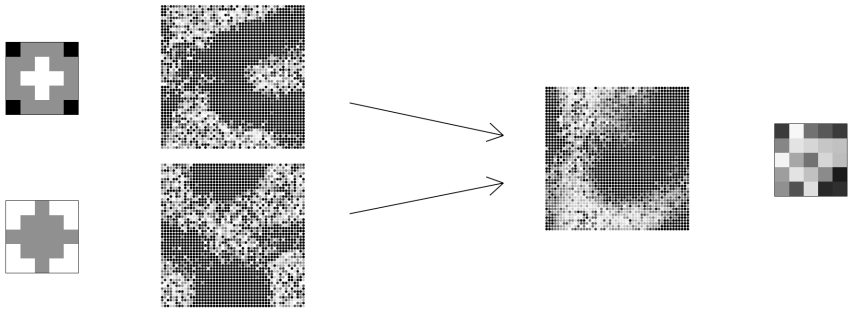

Figure 7. Convolution occurring in simulated neurons. The two input grids on the left are represented by the neural network firing patterns beside them. On the right is the neural network firing pattern that represents the convolution of the two inputs. The grid on the far right is the visual interpretation of the result of this convolution. Arrows indicate synaptic connections via intervening neural populations.

After two representations have been combined, it is still possible to extract the original information from the combined representation. The process of convolution can be reversed, using neural connections almost identical to those needed for performing the convolution in the first place, as shown in figure 8. There is a loss of information in that the extracted information is only an approximation of the original. However, by increasing the number of vector values and the number of neurons per value, we can make this approximation as accurate as is desired (Eliasmith and Anderson, 2003).

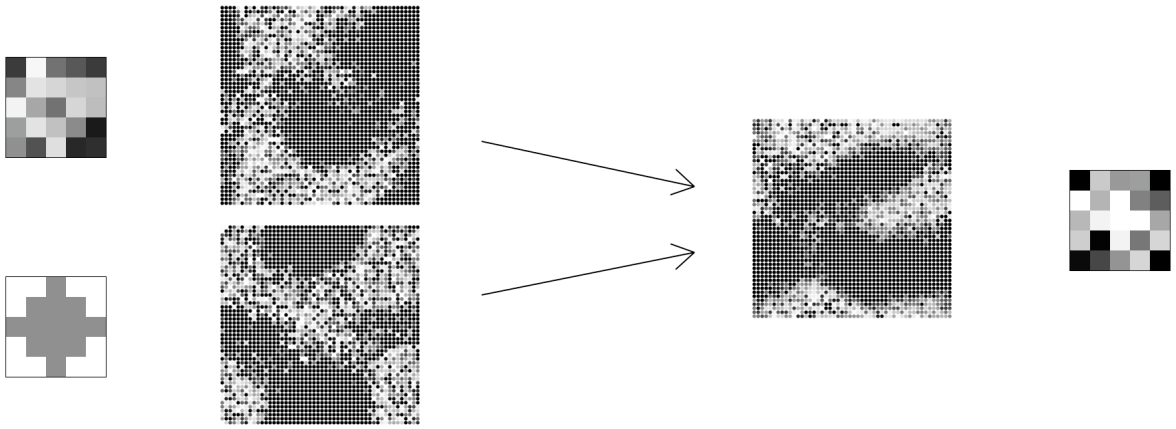

Figure 8. Deconvolution occurring in simulated neurons. Inputs are (top left) the output from figure 7 and (bottom left) the other input from figure 7. The result (far right) is an approximation of the first of the original inputs (top left in figure 7).

Simulation 3: Multimodal Convolution

Simulation 2 demonstrates the use of biologically realistic artificial neurons to combine two arbitrary vector-based representations, showing the feasibility of convolution as a method of conceptual combination. We used a visual example, but many other modalities can be captured by vectors. Hence we can create multimodal representations by convolving representations from distinct modalities. For example, we might combine a particular visual stimulus with a particular auditory stimulus, along with a symbolic label, and even an emotional valence.

Earlier we proposed that the AHA! experience required a convolution of at least four components: two representations that get combined into a new one, and two aspects

of emotional processing – cognitive appraisal and physiological perception. Figure 9 shows our NEF simulation of how this might work. Instead of just two representations being convolved, there are a total of four that could stand for various aspects of the cognitive and emotional content of a situation. Each convolution is implemented as before, and they are added by having all convolutions project to the same set of output neurons.

We do not need to assume that all of the vectors being combined are of the same length. For example, representing physiological perception may only require a few dimensions, whereas encoding symbolic content requires hundreds of dimensions. These vectors can still be combined by projecting the low-dimensional vector into the higher-dimensional space. This means that we can combine any representations of any size and any modality into a single novel representation.

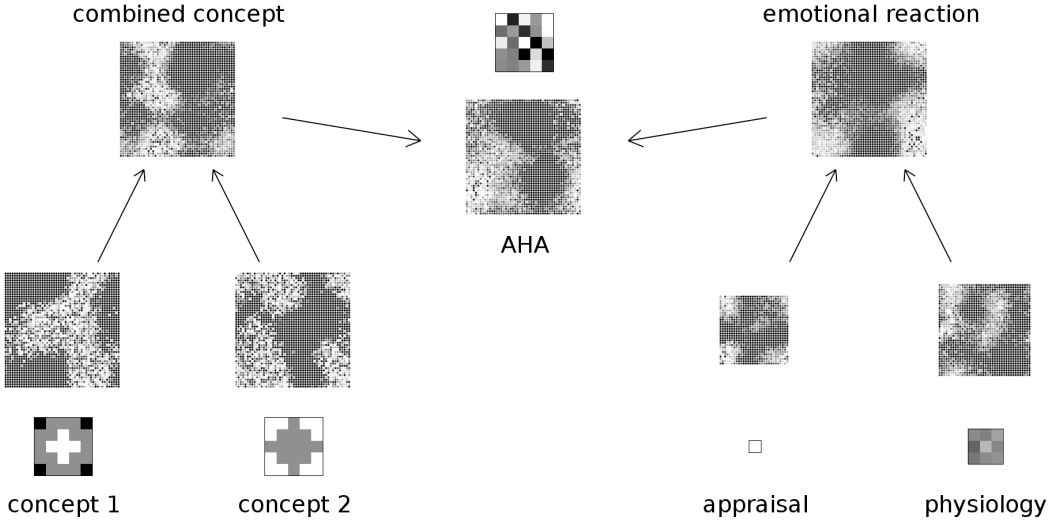

Figure 9: Combining four representations into a single result.

In our simulations, we have shown that vectors can be encoded using neural firing patterns. These vectors can represent information in any modality, from a visual stimulus

to an internal state to a symbolic term. Furthermore, these representations can be combined via a neural implementation of the convolution operation that produces a new vector that encodes an approximation of all the original vectors. We have thus shown the computational feasibility of using convolution to combine representations of concepts and emotional reactions to them, generating the AHA! experience.

LIMITATIONS

To our knowledge, we have presented the first detailed, neurocomputational account of psychological mechanisms that may contribute to creativity and the AHA! experience. But we acknowledge that our account has many limitations with respect to describing the neural and other kinds of processes that are involved in creativity and innovation. We see the models we have presented here as only a small part of a full theory, so we will sketch here what we see as some of the missing psychological, neural, and social ingredients.

By no means are we proposing a purely neural, ruthlessly reductionist account of creativity. We recognize the importance of understanding thinking in terms of multilevel mechanisms, ranging from the molecular to the psychological to the social, as shown in figure 10 (Thagard 2009, 2010; see also Bechtel, 2008, and Craver, 2007). Creativity has many social aspects, requiring interaction among people with overlapping ideas and interest. For example, scientific research today is largely collaborative, and many discoveries occur because of fruitful interactions among researchers (Thagard 1999, ch. 11; Thagard, 2006b). Neurocomputational models ignore the social causes of creative breakthroughs, which are often crucial in explaining how different representations come to be combined in a single brain. For example, Darwin's ideas about evolution and

breeding (artificial selection) were the result of many interactions he had with other scientists and farmers, and various engineers were involved in the development of the wireless technologies that evolved into the Blackberry. A full theory of creativity and innovation will have to flesh out the upward and downward arrows in figure 10 in a way that produces a more complete account of creativity.

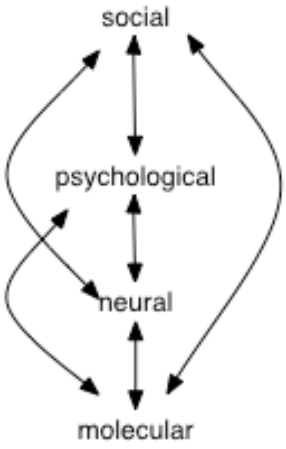

Figure 10. Causal interactions between four levels of analysis relevant to understanding creativity.

Figure 9 is incompatible with other, more common views of causality in multilevel systems. Reductionist views assume that causality runs only upwards, from the molecular to the neural to the psychological to the social. At the opposite extreme, antireductionist views assume that the complexity of systems and the nature of emergent properties (ones that hold of higher level objects and not of their constituents) means that higher levels must be described as largely independent of lower ones, so that social explanations (e.g. in sociology and economics) have a large degree of autonomy from psychological ones, and psychological explanations have a large degree of autonomy from neural and molecular ones.

From our perspective, reductionist, individualistic views are inadequate because they ignore ways in which objects and events at higher levels have causal effects on objects and events and lower levels. Consider, two scientists (perhaps Crick and Watson, who developed many of their ideas about DNA jointly), managing in conversation to make a creative breakthrough. This conversation is clearly a social interaction, with two people communicating in ways that may be both verbal and nonverbal. Such interactions can have profound psychological, neural, and molecular effects. When the scientists bring together two concepts or other representations that they have not previously connected, their beliefs may change dramatically, for example when they realize they have found a hypothesis that can solve the problem on which they have been jointly working. This psychological change is also a neural change, altering their patterns of neural activity. If the scientists are excited or even a little pleased by their new discovery, they will also undergo molecular changes such as increases in dopamine levels in the nucleus accumbens and other reward-related brain areas. Because social interactions can cause important psychological, neural, and molecular changes, the reductionist view that only considers how lower-level systems causally affect higher-level ones is implausible. Craver and Bechtel (2007) argue that there are no downward causes, but Thagard (forthcoming) disputes their arguments.

On the other hand, the holistic, antireductionist view that insists on autonomy of higher levels from lower ones is also implausible. Molecular changes even as simple as ingestion of caffeine or alcohol can have large effects on psychological and social processes; compare the remark that a mathematician is a device for turning coffee into theorems. If our account of creative representation combination as neural binding is on

the right track, then part of the psychological explanation of the creativity of individuals, and hence part of the social explanation of the productivity of groups, will require attention to neural processes. Genetic processes may also be relevant, as seen in the recent finding that combinations of genes involved in dopamine transmission have some correlation with artistic capabilities (Kevin Dunbar, personal communication).

One useful way to characterize multilevel relations is provided by Bunge (2003) who critiques both holism and individualism. He defines a system as a quadruple:

<Composition, Environment, Structure, Mechanism> where:

Composition = collection of parts;

Environment = items that act on the parts;

Structure = relations among parts, especially bonds between them;

Mechanism = processes that make the system behave as it does.

Our discussion in this paper has primarily been at the neural level: the composition is a collection of neurons; the environment consists of the physiological inputs such as external and internal perception that cause changes in firing of neurons, the structure consists of the excitatory and inhibitory synaptic connections among neurons, and the mechanism is the whole set of neurochemical processes involved in neural firing.

A complete theory of creativity would require specifying social, psychological, and molecular systems as well, not just on their own, but in relation to neural processes of creativity. We would need to specify the composition of social, psychological, neural, and molecular systems in a way that exhibits their part-whole relations. Much more problematically, we need to describe the relations among the processes that operate at different levels. This project of multilevel interaction goes far beyond the scope of the

current paper, but we mention it here to forestall the objection that neural convolution is only one aspect of creativity: we acknowledge that our neural account is only part of a full scientific explanation of creativity. Thagard (forthcoming) advocates the method of multilevel interactive mechanisms (MIM) as a general approach for cognitive science.

Various writers on the social processes of innovation have remarked on the importance of interpersonal contact for the transmission of *tacit knowledge,* which can be very difficult to put into words (e.g. Asheim and Gertler, 2005). Our neural account provides a non-mysterious way of understanding tacit knowledge, as the neural representations our models employ are compatible with procedural, emotional, and perceptual representations that may be hard to put into words. Transferring such information may require thinkers to work together in the same physical environment so that bodily interactions via manipulation of objects, diagrams, gestures, and facial expressions can provide sources of communication that may be as rich as verbal conversation. A full account of creativity and innovation that includes the social dimension should include explanation of how nonverbal communication can contribute to the joint production of tacit knowledge, which we agree is an important part of scientific thinking (Sahdra and Thagard, 2003).

Even at the psychological level, our account of creativity is incomplete. We have described how two representations can be combined into new ones, but we have not attempted to say what triggers such combinations. Specifying triggering conditions for representation combination would require full theories of language processing and problem solving. We have not investigated how conceptual combination can occur as part of language processing (e.g. Medin and Shobin, 1988; Wisniewski, 1997).

Conceptual combination is clearly not sufficient for creativity, as people make many boring combinations as part of everyday communication. Moreover, in addition to conceptual combination, there are other kinds of mental operations important for creativity, as we review in the next section on comparisons with other researchers' accounts of creativity.

For problem solving, Thagard (1988) presented a computational model of how conceptual combination could be triggered when two concepts become attributed to the same object during the attempt to solve explanation problems, and a similar account could apply to the kinds of creativity we have been concerned with here. When scientists, inventors, social activists, or artists are engaged in challenging activity, they naturally have mental representations such as concepts active simultaneously in working memory. Combining such representations occasionally produces creative results. But the neurocomputational models described in this paper deal only with the process of combination, not with triggering conditions for combination or with post-combination employment of newly created and combined representations.

A full account of creativity as representation combination would need to include both the upstream processes that trigger combination and the downstream processes that make use of newly created representations, as shown in figure 10. The downstream processes include assessing the ongoing usefulness of the newly created representations for whatever purpose inspired them. For example, the creation of a new theoretical concept such as *sound wave* becomes important if it enters the scientific vocabulary and becomes subsequently employed in ongoing explanations. Thus a full theory of creativity would have to include a description of problem solving and language

processing as both inputs to and outputs from the neural process of representation generation, which is all we have tried to model in this paper. Implementing figure 11 would require integrated neurocomputational models of problem solving and language processing that remain to be developed.

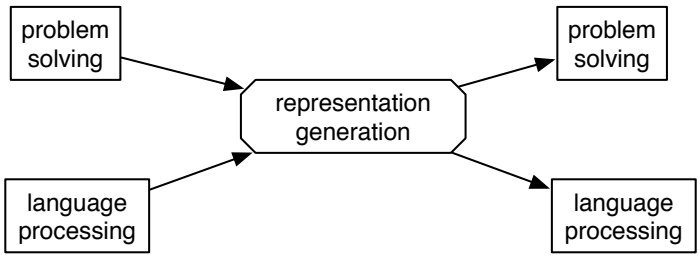

Figure 11. Problem solving and language processing as contributing to and affected by representation combination.

COMPARISONS WITH RELATED WORK

Our neural models of representation combination and emotional reaction are broadly compatible with the ideas of many researchers on creativity whose work has addressed different issues. Boden (2004) usefully distinguishes three forms of creativity: combinatorial, exploratory, and transformational. We have approached only the combinatorial form in which new concepts are generated, but have not dealt with the broader exploration and transformation of conceptual spaces. We have been concerned with what Boden calls psychological creativity, which involves the generation of representations new to particular individuals, and not historical creativity, which involves the generation of completely new representations not previously produced by any individual.

Our approach in this paper has also been narrower than Nersessian's (2008) discussion of mental models and their use in creating scientific concepts. She defines a

mental model as a "structural, behavioral, or functional analog representation of a real-world or imaginary situation, event or process" (p. 93). We agree with her contention that many kinds of internal and external representations are used during scientific reasoning, including ones tightly coupled to embodied perceptions and actions. As yet, no neurocomputational theory of mental models has been developed, so we are unable to situate our account of representation combination within the rich mental-models approach to reasoning. But our account is certainly compatible with Nersessian's claim (2008, p. 138) that conceptual changes arise from interactive processes of constructing, manipulating, evaluating and revising models.

We agree with Hofstadter (1995) that many computational models of analogy have relied too heavily on fixed verbal representations that prevent the kind of fluidity found in many kinds of creative thinking. In the models described in this paper, we use neural representations of a more biologically realistic sort, representing concepts by the activity of thousands of spiking neurons, not by single nodes as in localist connectionist simulations, nor by a few dozen non-spiking neurons as in PDP simulations. We acknowledge, however that we have not produced a new model of analogical processing. Eliasmith and Thagard (2001) was a first step toward a vector-based system of analogical mapping, but no full neural implementation of that system has yet been produced.

The scope of our project is also narrower than much artificial intelligence research on scientific discovery that develops more general accounts of problem solving (e.g. Langley, Simon, Bradshaw, and Zytkow, 1987; Bridewell, Langley, Todorovski, and Dzeroski, 2008). Our neural simulations are incapable of generating new scientific laws or representations of processes that are crucial for a general account of scientific

discovery. Another example of representation generation is the "crossover" operation that operates as part of genetic algorithms in the generation of new rules in John Holland's classifier systems (Holland, Holyoak, Nisbett, and Thagard, 1986). Our models could be seen as doing a kind of crossover mating between neural patterns, but they are not embedded in a general system capable of making complex inferences.

Our account of creativity as based on representation combination is similar to the idea of blending (conceptual integration) developed by Fauconnier and Turner (2002), which is modeled computationally by Pereira (2007). Our account differs in providing a neural mechanism for combining multimodal representations, including emotional reactions.

Brain imaging is beginning to yield interesting results about the neural correlates of creativity (e.g. Subranamiam, Kounios, Parrish, and Jung-Beeman, 2009; Bowden and Jung-Beeman, 2003). Unfortunately, our neural models of creativity are not yet organized into specific brain areas, so we cannot explain particular findings concerning these neural correlates.

CONCLUSION

Despite the limitations described in the last two sections, we think our account of representation generation is novel and interesting, perhaps even creative, in several respects. First, we have shown how conceptual combination can occur in biologically realistic populations of thousands of spiking neurons. Second, we employed a mechanism of binding – convolution – that differs in important ways from the synchrony mechanism that is more commonly advocated. Third, and perhaps most important, we have used convolution to show how the creative generation of new representations can

generate emotional reactions such as the much-desired AHA! experience.

Convolution also provides an alternative to synchronization as a potential naturalistic solution to the classic philosophical problem of explaining the apparent unity of consciousness. In figure 8, we portrayed a simulation that integrates the activity of seven neural populations, showing how there can be a unified experience of the combination of two concepts and an emotional reaction to them. We do not mean to suggest that there is a specific locus of consciousness in the brain, as there are many different convergent zones (also known as association areas) where information from multiple neural populations can come together. The dorsolateral prefrontal cortex seems to be an important convergence zone for working memory and hence for consciousness, as in the EMOCON model of Thagard and Aubie (2008).

Neural processes of the sort we have described are capable of encoding the full range of representations that contribute to human thinking, including ones that are verbal, visual, auditory, olfactory, tactile, kinesthetic, procedural, and emotional. This range should enable application of the mechanism of representation combination to all realms of human creativity, including scientific discovery, technological invention, social innovation, and aesthetic imagination. But much research remains to be done to build a full, detailed account of the creative brain.

APPENDIX

Representation using vectors across modalities

Using convolutions to combine neural representations makes the fundamental assumption that anything we wish to represent in the brain can be treated as a vector (i.e. a set of numbers of a fixed size). It is clear how this approach can be mapped to visual

representations; at the simplest level, the brightness of each pixel in an image can be mapped onto one value in the vector. This approach is used to create the figures in this paper. For colour images, three more values (representing the three primary colours) may be used per pixel. A more realistic approach can draw on extensive research on the primate visual cortex, showing many layers of neurons, each of which transforms the vector in the previous layer into a new vector, where, for example, individual values may indicate the presence of edges at particular locations.

Other modalities can also be treated in this manner. For audition, the cochlea contains sensory cells responsive to different sound frequencies (from 20Hz to 20,000Hz). The activity of these cells can be seen as a vector of values, each value corresponding to a different frequency. For olfaction, the mammalian nose contains approximately 2000 different types of receptor cells, each one sensitive to a different range of chemicals. This range can be treated as a large vector with 2000 values, one value for each type of receptor cell (e.g. Hopfield, 1999).

Verbal representations such as sentences, frames (sets of attribute-value pairs), and analogies can be converted into vectors using the holographic reduced representation method of Plate (2003; see also Eliasmith and Thagard, 2001).

Representing vectors using neurons

The simulations in this paper use the general approach to encoding a vector into a population of neurons of Eliasmith and Anderson (2003). Every neuron has a preferred direction vector $\tilde{\phi}$ such that the current entering the neuron is proportional to the similarity (dot product) between $\tilde{\phi}$ and the vector $\boldsymbol{x}$ being represented. If α is the sensitivity of the neuron and J^{bias} is a fixed background current, then the total current

flowing into cell i at any given point is

$$J_i = \alpha_i \tilde{\phi}_i \cdot \boldsymbol{x} + J_i^{bias} \tag{1}$$

In our simulations, we model each neuron with the standard leaky integrate-and-fire (LIF) model. This produces a series of spikes at times t_{in} for each neuron i. If the details of the LIF model (i.e. the relation between input current and spiking behaviour) are written as $G[\cdot]$ and the neural noise of variance σ^2 is $\eta(\sigma)$, then the encoding of any given $\boldsymbol{x}$ as the temporal spike pattern across the neural group is given as

$$\sum_n \delta(t - t_{in}) = G_i[\alpha_i \tilde{\phi} \cdot \boldsymbol{x}(t) + J_i^{bias} + \eta_i(\sigma)] \tag{2}$$

This formula allows us to determine the spiking pattern across a group of neurons which corresponds to a particular input $\boldsymbol{x}$. We can also perform the opposite operation: using the pattern of spikes to recover the original value of $\boldsymbol{x}$. We write this as $\hat{x}$ to indicate that it is an estimate, and this is used above to determine the output vectors from our simulations (the right-most image in Figure 7 and the central image in Figure 8).

The first step in calculating $\hat{x}$ is to determine the linearly optimal decoding vectors ϕ for each neuron as per Equation 3, where a_i is the firing rate for neuron i. This method has been shown to uniquely combine accuracy and neurobiological plausibility (e.g. Salinas and Abbot, 1994).

$$\phi = \Gamma^{-1} \Upsilon \qquad \Gamma_{ij} = \int a_i a_j \, dx \qquad \Upsilon_j = \int a_i x \, dx \tag{3}$$

Now we can determine $\hat{x}$ by weighting the activity of each neuron by the corresponding ϕ value. To improve the neural realism of the model, we can do this weighting in a manner that respects the causal influence one neuron can have on another. That is, instead of just considering the spike timing, we note that when one neuron spikes,

it affects the next neuron by causing a post-synaptic current to flow into it. These currents are well-studied, and different neurotransmitters have different characteristic shapes. If we denote the current caused by a single spike at time t=0 as $h(t)$ for a given neurotransmitter, then we can use these currents to derive our estimate of $\boldsymbol{x}$ as follows

$$\hat{\boldsymbol{x}}(t)=\sum_{in}\delta(t-t_{in})*h_i(t)\phi_i=\sum_{in}h(t-t_{in})\phi_i \tag{4}$$

Deriving synaptic connection weights

Given the decoding vectors ϕ derived above, we can derive the optimal synaptic connection weights for connecting two groups of neurons such that the value represented by the second group will be some given function of the value represented by the first group, providing the basis for defining the convolution operation given in this paper (Eliasmith, 2004). We start with a linear transformation. That is, if the first group of neurons represents $\boldsymbol{x}$ and the second group represents $\boldsymbol{y}$, we want $\boldsymbol{y}=M\boldsymbol{x}$, where M is an arbitrary matrix. Both $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of arbitrary size. To achieve this, Equation 1 dictates that the current entering the second group of neurons should be as follows, where we use the index j for the elements of the second group.

$$J_j=\alpha_j\tilde{\phi}_j\cdot\boldsymbol{x}+J_j^{bias} \tag{5}$$

If we substitute $\hat{x}$ for $\boldsymbol{x}$ using Equation 4, we can express the current coming into the second group of neurons as a function of the current leaving the first group.

$$J_j=\alpha_j\tilde{\phi}_j\cdot\sum_{in}h(t-t_n)\phi_i+J_j^{bias} \tag{6}$$

Rearranging this equation leads to an expression where the current leaving each neuron in the first group is weighted by a fixed value and summed to produce the current in the second group of neurons. These weights are the desired synaptic connection

weights.

$$J_j = \sum_{in} \alpha_j \tilde{\phi}_j \phi_i h(t - t_n) + J_j^{bias} \tag{7}$$

This approach allows us to derive optimal connection weights ω_{ij} for any linear operation. It should be noted that this process will also work without modification for adding together inputs from multiple neural groups.

$$\omega_{ij} = \alpha_j \tilde{\phi}_j \phi_i \tag{8}$$

To determine the connection weights for nonlinear operations, we return to the derivation of the decoding vectors in Equation 3. We modify this to derive a new set of decoding vectors which will approximate an arbitrary function of x, rather than x itself. Equation 3 can be seen as a special case of Equation 9 where $f(\boldsymbol{x})=\boldsymbol{x}$.

$$\phi^{f(x)} = \Gamma^{-1} \Upsilon^{f(x)} \qquad \Gamma_{ij} = \int a_i a_j dx \qquad \Upsilon_i^{f(x)} = \int a_i f(x) dx \tag{9}$$

Given these tools, we can derive the synaptic connections needed for performing the convolution operation. The definition of the circular convolution is given in equation 10. Importantly, we also see that it can be rewritten as a multiplication in the Fourier domain, as per the convolution theorem. We use this form to reduce the number of nonlinear operations required, thus increasing the accuracy of the neural calculation.

$$\boldsymbol{z} = \boldsymbol{x} * \boldsymbol{y} \qquad \boldsymbol{z}_i = \sum_{j=1}^{N} \boldsymbol{x}_j \boldsymbol{y}_{(i-j) \bmod N} \qquad \boldsymbol{z} = F^{-1}(F(\boldsymbol{x}) \cdot F(\boldsymbol{y})) \tag{10}$$

Our first step is to define a new group of neurons (indexed by k) to represent $\boldsymbol{v}$ which will contain the Fourier transforms of the vector $\boldsymbol{x}$ represented by one group of neurons (indexed by i) and the vector $\boldsymbol{y}$ represented by another group of neurons (indexed by j). That is, we want $\boldsymbol{v}=[F(\boldsymbol{x}),F(\boldsymbol{y})]$. Since the Fourier transform is a linear operation,

we can use Equation 8 to derive these weights, where F_D is defined to be the Discrete Fourier Transform matrix of the same dimensionality D as $\boldsymbol{x}$ and $\boldsymbol{y}$.

$$\begin{aligned} \omega_{ik} &= \alpha_k \tilde{\phi}_k M \phi_i \qquad M = [F_D, 0] \\ \omega_{jk} &= \alpha_k \tilde{\phi}_k M \phi_j \qquad M = [0, F_D] \end{aligned} \tag{11}$$

We next need to multiply the individual Fourier transform coefficients together. Since this is a nonlinear operation, this is done using Equation 9 where the function $f(\boldsymbol{v})$ is defined by taking the product of the corresponding elements in vector $\boldsymbol{v}$. That is, element 1 in the result will be the production of elements 1 and 1+D, element 2 will be the product of elements 2 and 2+D, and so on. We combine this new set of decoding vectors with the matrix corresponding to the inverse Fourier transform to produce the synaptic connection weights to the final set of neurons (indexed by m) representing $\boldsymbol{z}=\boldsymbol{x}*\boldsymbol{y}$.

$$\omega_{km} = \alpha_m \tilde{\phi}_m F_D^{-1} \phi_k^{f(v)} \tag{12}$$

These synaptic connection weights are used in all the simulations in this paper.

Acknowledgements: Our research has been supported by the Natural Sciences and Engineering Research Council of Canada and SHARC/NET. Thanks to Chris Eliasmith for valuable advice on simulations and for helpful comments on a previous draft.

REFERENCES

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Asheim, B. T., & Gertler, M. S. (2005). The geography of innovation: Regional innovation systems. In J. Fagerberg, D. C. Mowery & R. R. Nelson (Eds.), *The Oxford handbook of innovation* (pp. 291-317). Oxford: Oxford University Press.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences, 7*, 84-91.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: Routledge.

Boden, M. (2004). *The creative mind: Myths and mechanisms* (2nd ed.). London: Routledge.

Bowden, E. M., & Jung-Beeman, M. (2003). Aha! Insight experience correlates with solution activation in the right hemisphere. *Psychonomic Bulletin and Review, 2003*(10), 730-737.

Bridewell, W., Langley, P., Todorovski, L., & Dzeroski, S. (2008). Inductive process modeling. *Machine learning, 71*, 1-32.

Bunge, M. (2003). *Emergence and convergence: Qualitative novelty and the unity of knowledge*. Toronto: University of Toronto Press.

Chandrasekharan, S. (2009?). Builiding to discovery: A common coding model. *Cognitive Science, 33*, 1059-1086.

Churchland, P. S., & Sejnowski, T. (1992). *The computational brain*. Cambridge, MA: MIT Press.

Costello, F. J., & Keane, M. T. (2000). Efficient creativity: constraint-guided conceptual combination. *Cognitive Science, 24*, 299-349.

Craver, C. F. (2007). *Explaining the brain*. Oxford: Oxford University Press.

Craver, C. F., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy, 22*, 547-663.

Crick, F. (1994). *The astonishing hypothesis: The scientific search for the soul*. London: Simon and Schuster.

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

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Eliasmith, C. (2004). Learning context sensitive logical inference in a neurobiological simulation. In S. Levy & R. Gayler (Eds.), *Compositional connectionism in cognitive science* (pp. 17-20). Menlo Park, CA: AAAI Press.

Eliasmith, C. (2005). Cognition with neurons: A large-scale, biologically realistic model of the Wason task. In B. Bara, L. Barasalou & M. Bucciarelli (Eds.), *Proceedings of the XXVII Annual Conference of the Cognitive Science Society* (pp. 624-629). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Eliasmith, C., & Stewart, T. C. (forthcoming). A new biologically implemented cognitive architecture. *unpublished manuscript, University of Waterloo Centre for*

Theoretical Neuroscience.
Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science, 25*, 245-286.
Engel, A. K., Fries, P., König, P., Brecht, M., & Singer, W. (1999). Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition, 8*, 128-151.
Fauconnier, G., & Turner, M. (2002). *The way we think*. New York: Basic Books.
Feynman, R. (1999). *The pleasure of finding things out*. Cambridge, MA: Perseus Books.
Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 233*(4771), 1416-1419.
Grandjean, D., Sander, D., & Scherer, K. R. (2008). Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and Cognition, 17*, 484-495.
Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.
Hebb, D. O. (1980). *Essay on mind*. Hillsdale, NJ: Lawrence Erlbaum.
Hofstadter, D. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. New York: Basic Books.
Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press/Bradford Books.
Holyoak, K. J., & Thagard, P. (1997). The analogical mind. *American Psychologist, 52*, 35-44.
Hopfield, J. (1999). Odor space and olfactory processing: Collective algorithms and neural implementation. *Proceedings of the National Academy of Sciences, 96*, 12506-12511.
Jacob, F. (1988). *The statue within* (F. Philip, Trans.). New York: Basic Books.
James, W. (1884). What is an emotion? *Mind, 9*, 188-205.
Koestler, A. (1967). *The act of creation*. New York: Dell.
Kounios, J., & Beeman, M. (2009). The *Aha!* moment: The cognitive neuroscience of insight. *Current directions in psychological science, 18*, 210-216.
Koza, J. R. (1992). *Genetic programming*. Cambridge, MA: MIT Press.
Kunda, Z., Miller, D., & Claire, T. (1990). Combining social concepts: The role of causal reasoning. *Cognitive Science, 14*, 551-577.
Laird, J., Rosenbloom, P., & Newell, A. (1986). Chunking in Soar: The anatomy of a general learning mechanism. *Machine Learning 1* 11-46.
Lenat, D., & Brown, J. S. (1984). Why AM and Eurisko appear to work. *Artificial Intelligence, 23*, 269-294.
Medin, D., & Shoben, E. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20*, 158-190.
Mednick, S. A. (1962). The associate basis of the creative process. *Psychological Review, 69*, 220-232.
Minsky, M. (1974). A framework for representing knowledge: MIT Artificial Intelligence Laboratory.
Mumford, M. D. (2002). Social innovation: Ten cases from Benjamin Franklin. *Creativity research journal, 14*, 253-266.
Nersessian, N. (2008). *Creating scientific concepts*. Cambridge, MA: MIT Press.
Pereira, F. C. (2007). *Creativity and artificial intelligence*. Berlin: Mouton de Gruyter.

Plate, T. (2003). *Holographic reduced representations*. Stanford: CSLI.
Prinz, J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford: Oxford University Press.
Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge MA: MIT Press/Bradford Books.
Sahdra, B., & Thagard, P. (2003). Procedural knowledge in molecular biology. *Philosophical Psychology, 16*, 477-498.
Salinas, E., & Abbott, L. F. (1994). Vector reconstruction from firing rates. *Journal of Computational Neuroscience, 1*, 89-107.
Saunders, D., & Thagard, P. (2005). Creativity in computer science. In J. C. Kaufman & J. Baer (Eds.), *Creativity across domains: Faces of the muse* (pp. 153-167). Mahwah, NJ: Lawrence Erlbaum Associates.
Smith, E., & Osherson, D. (1984). Conceptual combination with prototype concepts. *Cognitive SciencefR fI, 8*, 337-361.
Smith, E. E., & Kosslyn, S. M. (2007). *Cognitive psychology: Mind and brain*. Upper Saddle River, NJ: Pearson Prentice Hall.
Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence, 46*, 159-217.
Stewart, T. C., & Eliasmith, C. (2009). Compositionality and biologically plausible models. In W. Hinzen, E. Machery & M. Werning (Eds.), *Oxford handbook of compositionality*. Oxford: Oxford University Press.
Subranamiam, K., Kounios, J., Parrish, T. B., & Jung-Beeman, M. (2009). A brain mechanism for facilitation of insight by positive affect. *Journal of Cognitive Neuroscience, 21*, 415-432.
Sweeny, A. (2009). *BlackBerry planet*. Missisaugua, ON: Wiley.
Thagard, P. (1988). *Computational philosophy of science*. Cambridge, MA: MIT Press.
Thagard, P. (1999). *How scientists explain disease*. Princeton: Princeton University Press.
Thagard, P. (2006a). *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge, MA: MIT Press.
Thagard, P. (2006b). How to collaborate: Procedural knowledge in the cooperative development of science. *Southern Journal of Philosophy, 44*(177-196).
Thagard, P. (2009). Why cognitive science needs philosophy and vice versa. *Topics in Cognitive Science, 1*, 237-254.
Thagard, P. (2010). *The brain and the meaning of life*. Princeton, NJ: Princeton University Press.
Thagard, P. (forthcoming). Who are you? The self as a system of multilevel interacting mechanisms. *unpublished manuscript - University of Waterloo*.
Thagard, P., & Aubie, B. (2008). Emotional consciousness: A neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. *Consciousness and Cognition, 17*, 811-834.
Treisman, A. (1996). The bindng problem. *Current opinion in neurobiology, 6*(171-178).
Watson, J. D. (1969). *The double helix*. New York: New American Library.
Werning, M., & Maye, A. (2007). The cortical implementation of complex attribute and substance concepts: Synchrony, frames, and hierarchical binding. *Chaos and*

complexity letters, 2, 435-452.
Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings*. Cambridge, MA: Harvard University Press.
Wisniewski, E. J. (1997). Conceptual combination: Possibilities and esthetics. In T. B. Ward, S. M. Smith & J. Viad (Eds.), *Conceptual structures and processes: Emergence, discovery, and change* (pp. ADD). Washington, D. C.: American Psychological Association.
Wolpert, L., & Richards, A. (1997). *Passionate minds: The inner world of scientists*. Oxford: Oxford University Press.
Wu, L. L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychological, ADD*.