CREATIVE INTUITION: HOW *EUREKA* RESULTS FROM THREE NEURAL MECHANISMS

Paul Thagard
University of Waterloo
Draft 4, July, 2012

Thagard, P. (in progress). For L. M. Osbeck & B. S. Held (Eds.), *Rational intuition: Philosophical roots, scientific investigations.* Cambridge: Cambridge University Press.

Introduction

When Archimedes shouted *Eureka*, "I have found it", he was experiencing self-consciousness of creativity: he became aware and excited that he had produced a new and valuable idea. Understanding this phenomenon is the ultimate challenge for cognitive science, because it requires simultaneous solution of three of its major problems: the nature of the self, consciousness, and creativity. This chapter will argue that all three problems have the same solution based on three fundamental brain mechanisms: neural representation, recursive binding, and interactive competition. Creative intuition is not a mysterious process of divine inspiration or Platonic apprehension of ideas, but rather the result of identifiable neural processes that operate in all humans. These processes are mechanistic, in that they result from the interactions of parts that produce regular changes (see e.g. Bechtel, 2008).

The historical record is insufficient to determine whether Archimedes really did shout Eureka (when taking a bath gave him an idea for measuring the volume of irregular solids), but there are undoubtedly real examples. For example, Darwin (1987) recorded in his notebook his realization in 1838 that biological evolution could result from natural selection among competing organisms. Many of us have experienced lesser moments of illumination with the same cognitive and emotional structure. For example, here is how I got the idea for my theory of explanatory coherence (Thagard 1989, 1992). On a

Saturday night in spring of 1987, I was in a movie theatre watching a boring movie, *Beverly Hills Cop 2*. For the previous few weeks, I had been excitedly programming a neural network model based on the insight of my collaborator Keith Holyoak that analogical mapping might be a process of parallel constraint satisfaction (Holyoak and Thagard, 1989). The movie was tedious, so I got to thinking about how well the computer program was working and wondering what other problems might be amenable to similar techniques. Suddenly it occurred to me that the main problem of my PhD thesis, evaluation of scientific theories, might also be a matter of satisfying multiple constraints. That evening and the next day, I worked out the details. A creative intuition concerning the connection between analogy and explanatory inference provided me with a new theory and computational model of explanatory coherence, generating my excited Eureka reaction.

Before getting into the details of the neural mechanisms that I think are responsible for such reactions, let me deal with some terminological preliminaries. I take intuitions to be conscious realizations that result from unconscious processes hard to identify. An intuition is something that pops into your head in a much less deliberative manner than the results of a verbal argument. I take intuitions to be creative if the realizations they produce are both new and valuable. Many characterizations of creativity have been given (e.g. Boden, 2004; Kaufman & Sternberg, 2010), but they all boil down to the recognition that creative leaps are both new (novel, surprising, original, etc.) and valuable (important, useful, appropriate, etc.). Hence a creative intuition is a suddenly conscious realization concerning something that is new and valuable.

Many writers on creativity have seen it as resulting from combinations of previously unconnected mental representations (e.g. Boden, 2004; Koestler, 1967; Mednick, 1962; Stewart, 1792; Thagard 1988). Evidence for this claim is usually anecdotal, but examination of 200 examples of scientific discovery and technological invention revealed combinations required for all 200 breakthroughs (Thagard, 2012). A mental representation is a structure or process in the mind that stands for something.

This chapter begins with an outline of an emerging theory of mental representations as patterns of activity in populations of neurons. Then it describes current views about how such representations can be combined into new ones by processes of binding that are performed by neural operations, resulting in high-level neural representations that Chris Eliasmith (forthcoming) calls *semantic pointers*. Such representations are usually unconscious but they can become conscious through a process of interactive competition among them, with the most important of them entering awareness. On this view, creative intuition is the result of competition between semantic pointers that are formed by binding other representations. In a slogan:

Eureka = representation + binding + competition.

Neural Representation

The most familiar kinds of representations are linguistic ones such as words and sentences. From a cognitive perspective, concepts are mental representations on the same scale as words, and propositions are mental representations on the same scale as sentences: propositions are formed out of concepts just as sentences are formed out of words. Mental representations, however, are not restricted to linguistic formats, as people are also capable of many kinds of images corresponding to different sensory

modalities, including vision, sound, touch, taste, smell, pain, and muscular motion (kinesthesia). There is much evidence, at least for visual representations, that these kinds of imagery cannot be reduced to verbal structures (Kosslyn, Ganis, and Thompson, 2003).

A unified account of mental representation can be given in terms of neural networks, although early accounts of artificial neural networks were insufficient to capture the full representational power of human brains. Simple connectionist networks consisted of neuron-like units that correspond to whole concepts or propositions. Like neurons, such units are connected to each other by excitatory and inhibitory links, and processing occurs by spreading of activation among the units based on the connections they have to each other. Although there may be some neurons in the brain that respond to specific stimuli, most representations are thought to be distributed among thousands or millions of neurons. Hence learning methods that produce distributed representations in artificial neural networks were a significant advance (Rumelhart and McClelland, 1986). The resulting PDP (parallel distributed processing) networks, however, were still limited in their ability to represent linguistically complex information, such as the proposition that if a woman loves a man, then the man may or may not love the woman.

To overcome this problem, techniques were developed to enable complex verbal information to be represented in vectors, which can be lists of numbers such as (.2, .5, .8 …). A vector can capture the activation of a whole set of neurons: in the example just given, the first number stands for the relative firing rate of neuron 1 equal to 20% of capacity, the second for the firing rate of neuron 2, and so on. If a neuron is capable of firing 100 times per second, then .2 indicates that it is firing 20 times per second. Hence

the mathematical objects called vectors can provide a simple approximation to populations of neurons, but more complex approximations with spiking neurons are described below. Paul Smolensky (1990) and Tony Plate (2003) developed powerful methods for translating complex verbal information into vectors that are built up out of other vectors (for a tutorial on how this works, see Eliasmith and Thagard, 2001). Hence neural networks based on these kinds of vector constructions do not have the representational limitations of previous connectionist and PDP approaches.

Eliasmith and Anderson (2003) showed how vectors can be represented in populations of biologically realistic neurons. The activity of neurons in connectionist and PDP models was limited to activation, the rate of firing. Real neurons, however, carry information not just by their rate of firing, but also by their pattern of firing, which in this context is usually called spiking. For example, here are two patterns of firing that both have the same rate, but different patterns: (fire, rest, fire, rest …) vs. (fire, fire, rest, rest, …). Spiking neurons have significantly enhanced representational and computational power compared to rate neurons (Maass, 1999). Groups of neurons need to work together with temporal coordination, just as musicians in a band need to interact and coordinate to produce an effective song.

Because we now have computational neural network models that closely mimic the operations of the brain, I think the best current account of mental representations is that they consist of patterns of spiking activity in populations of millions of neurons. This account can accommodate both information that is verbal and information from all sensory modalities, and shows how these forms of information can interact via the same processing format of neuronal spiking resulting from synaptic connections. Emotional

information can also be captured as spiking activity in populations of neurons, as will be described below.

In order to use this account of neural representation as the basis of a theory of creative intuition, we need to show that it can handle all the components of self-consciousness of creativity, including what is discovered, the self that did the discovering, and the emotion that accompanies realization of the accomplishment. Building up such complex representations requires the process of binding.

Recursive Binding

The importance of binding of representations in cognitive processing is evident in basic operations of vision and language (Revonsuo, 2009). When you see an apple, you do not see its color and shape as independent, but rather as bound together as properties of a single object. Similarly, when you process a simple sentence such as "Eve ate the apple", you need to bind the action of eating with the agent Eve and the object apple. Binding also makes possible the mingling of modalities, as when you coordinate the color and shape of an apple with its taste and smell, and when you verbally describe the taste and smell. Human mental representation is multimodal, including: words and other verbal representations built from them; sensory representations such as pain, vision, hearing, touch, smell, taste, and kinesthesia; and emotions, which synthesize cognitive appraisals and bodily perceptions (Thagard and Aubie, 2008). Without binding, thinking would be an overwhelming jumble of unconnected representations incapable of producing thought and action.

Any animal with a brain and sensory system presumably has a binding mechanism, but it takes a large brain with many interconnected neurons to be capable of

repeated, embedded bindings, which I will call *recursive bindings*: bindings of bindings of bindings. Human language can manage many layers of recursive binding, as in the song "There was an old woman who swallowed a cow to catch the goat to catch the cat to catch the bird to catch the spider that wriggled and jiggled and tickled insider her." Recursive binding is evident in Eureka phenomena that require combining self, discovery, and emotions, where each of these involves a binding of bindings. For example, Archimedes's "I found it" requires binding of the action of finding with the agent himself and the discovery concerning using buoyancy to measure volume of an irregular piece of metal. I will return to the question of self-representation in a later section.

How binding works is relatively clear in formal logic and linguistics, but these deal only with syntactic structures, not with the full range of representations that are part of human thought. How does the brain bind together representations when these are considered as patterns of activation in populations of spiking neurons? There are currently two candidate accounts of neural binding: synchrony and convolution.

Many cognitive scientists have endorsed the view that the brain performs binding by synchronizing the firing of neurons in different populations (see, for example Engel et al., 1999; Hummel and Holyoak, 2003). Suppose, for example, that *red* is represented by neurons firing in one population, and *apple* is represented by firing of neurons in another population. Then the binding *red apple* can result from coordination of firing in neurons in the two populations, just as two bands could listen to each other and start playing the same song together. It is not obvious, however, that synchrony is the correct or only mechanism for neural binding, for both empirical and theoretical reasons.

Whereas neural oscillations undoubtedly occur in brains systems as measured by electroencephalography, there is no direct evidence that the synchronization accomplished in these oscillations performs binding. Moreover, there are computational reasons for doubting that synchrony suffices for binding, having to do with the capacity of synchronizations to combine and take apart sufficiently complex representations (Stewart and Eliasmith, 2009; Thagard and Stewart, 2011).

Convolution is an alternative mechanism that Plate (2003) applied to vectors, which I explained earlier as lists of numbers. Any kind of representation can be translated into a vector. For example the word "cat" might be represented by the three-dimensional vector (3, 1, 20) because it contains the third, first, and twentieth letters of the alphabet. The picture on a high-definition television screen can be represented by a vector with 1920 X 1080 dimensions (more than 2 million), because such a screen has that number of pixels which have intensities describable by numbers. The mathematical method that Plate employed for producing new vectors by combining old ones is rather technical, so I avoid explaining it here (see Plate, 2003; Eliasmith and Thagard, 2001; Thagard and Stewart, 2011).

Instead, let me provide some metaphors that I hope will give at least a rough impression of how this kind of convolution works. Convolution of representations is something like braiding hair, which typically combines three separate strands. The strands are woven together into a single strand which looks different from the original hair, but which can eventually be unbraided to return the hair to its original shape. Similarly, convolution takes two or more vectors and "braids" them into a new vector that can operate as a whole and can also, when desired, be unbraided into the vectors that

compose it. For vectors, this process of deconvolution is only approximate in that the vectors that result from taking the convolved vector apart are not exactly the same as the ones that put it together.

There is currently no direct evidence that convolution is *the* mechanism used by the brain to perform binding of representations, but it is theoretically powerful for explaining how binding might work for any kind of representation for which there is a corresponding vector. Eliasmith (2005, forthcoming) developed a general method for enabling convolution of vectors to be performed by spiking neurons operating on neural representations. So there are now running computational models of spiking neurons that perform convolution with all the generality and efficiency needed to account for known kinds of binding. It is of course possible that the brain also uses synchronization and other unknown mechanisms for performing binding: biological systems often have several ways of performing important functions. But from now on I will assume that convolution is the most important mechanism for binding representations. We can then take advantage of a recent discovery by Chris Eliasmith (forthcoming) that convolution can be used to build powerful representations he calls *semantic pointers*.

Semantic Pointers

At the 2010 conference of the Cognitive Science Society, the ten winners of the Rummelhart prize to that date were asked to discuss the most important problems in cognitive science. Several of them mentioned the problem of figuring out how billions of neurons in the human brain are capable of processing symbols of the sort that operate in human language and thought. What brain processes produce concepts such as *apple* and meaningful sentences such as *Eve ate the apple?* Eliasmith's new idea of semantic

pointers seems to me to be the most plausible proposal to date for answering this fundamental question. This idea is very useful for understanding the nature of concepts (Blouw, Solodkin, Eliasmith, and Thagard, forthcoming; Thagard, 2012, ch. 18), emotions (Thagard and Schröder, forthcoming), intentions (Schröder, Stewart, and Thagard, forthcoming), and behavioral priming (Schröder and Thagard, forthcoming).

A mathematical treatment of semantic pointers can be given by interpreting them as vectors produced by convolution, but here I want to give a more qualitative and metaphorical explication that is accessible to non-mathematicians. Semantic pointers are neural processes that (1) provide *shallow meaning* through symbol-like relations to the world and other representations, (2) expand to provide *deeper meaning* with relations to perceptual, motor, and emotional information, (3) support complex syntactic operations, and (4) help to control the flow of information through a cognitive system to accomplish its goals. A semantic pointer consists of spiking patterns in a large population of neurons that provide a kind of compressed representation analogous to JPEG picture files or iTunes audio files. Just as a JPEG file can be expanded to produce a picture and an iTunes audio file can be expanded to produce music, so the semantic pointer can be used as a whole unit but also can be expanded to carry all the meaning that goes with the multimodal information that went into it.

The term "pointer" comes from computer science where it refers to a kind of data structure that gets its value from a machine address to which it points. A semantic pointer is a neural process that compresses information in other neural processes to which it points and into which it can be expanded when needed. For example, the concept *apple* can be understood as a semantic pointer consisting of spiking neurons that

compress, point to, and expand into other populations of spiking neurons that contain a wide range of information in various modalities such as vision and touch. In mathematical terms, the semantic pointer achieves its compressed representation by being formed through the convolution of other representation, and decompression is accomplished by means of deconvolution.

To use my earlier metaphor, the semantic pointer is the braid that results from weaving together several strands of information in different formats. That metaphor is limited by the fact that braids do not have any dynamic function, so let me try a different one: a semantic pointer is like a suitcase full of suitcases that can contain other suitcases. It is much more convenient to carry around a suitcase full of suitcases than a messy bunch of suitcases, but it is important that the big suitcase can be unpacked to find other suitcases that eventually can be unpacked to reveal contents such as clothes. Similarly, semantic pointers provide convenient ways of carrying out important functions such as syntactic processing and inferences, but can be unpacked to reveal the sources of meaning arising from sensory and motor processes. Semantic pointers are thus the most powerful kind of representations that result from recursive binding of neural representations.

It is plausible, therefore, that the thought Eureka! operated in Archimedes' mind as a semantic pointer built out of semantic pointers for the self, the discovery he made, and his emotional reaction to it. But I need to describe how self representations can be understood as semantic pointers.

Semantic Pointers for the Self

The understanding of Eureka as a semantic pointer presumes that the I or self can be represented by a semantic pointer that is bound into the larger pointer for "I have found it." This interpretation runs counter to the two main ways in which philosophers have understood the self. Idealist philosophers such as Plato and Kant have assumed that selves are transcendental, unified entities – souls. In contrast, some naturalistic philosophers like Hume have been skeptical of the existence of the self, viewing it as a misleading concoction from diverse kinds of experience. Thinking of the self as represented by a semantic pointer shows how it can have both unity and diversity, although social and molecular mechanisms are relevant in addition to psychological and neural ones (Thagard, in press; Thagard and Wood, forthcoming).

The unity comes from the fact that the semantic pointer representation of the "I" or "Paul Thagard" is sufficiently compressed that it can figure in syntactic structures like "I have found it" and "I am a cognitive scientist" as well as contributing to many kinds of inferences. The diversity comes from the capacity of the pointer to decompress or unpack into many other kinds of information including current sensory experiences (I am listening to music), memories (I got my PhD at the University of Toronto), and general concepts that apply to myself and others (I am a Canadian professor). All these are bound together into the convenient unifying representation that the semantic pointer provides. By virtue of the compressed representation that can be decompressed, the semantic pointer representation of self gets the convenient package that can be bound with actions and objects, but also carries the diverse range of information that comes from previous and current experience. In mathematical terms, the self-representation is a

vector that can be manipulated as a whole but can also be deconvolved into the various vectors that were joined together to produce it by convolution.

Despite this unity, the self-concept is malleable – changeable based on social context (Markus and Kunda, 1986). People think of themselves in different ways depending on their social situations, for example as sociable while at a party but as introspective while reading a book. Such malleability is naturally explained by noticing that context will activate different concepts associated with the self, bringing them to consciousness. Similarly, it is an important ingredient of Eureka that it is a conscious experience. Neural representation and recursive binding are powerful mechanisms for building semantic pointers, but we need a third mechanism to explain how such pointers enter consciousness.

Consciousness and Interactive Competition

Psychologists and neuroscientists have long recognized that *attention* is a crucial part of conscious awareness. At any moment, there are many events occurring in our environments and many kinds of information being processed in our brains. Attention is limited, in that we can only be consciously aware of a few items at once. Perhaps this limitation is a side effect of shortage of processing capacity in active memory resulting from the large number of neurons it takes to perform bindings, or perhaps the limitation is a biological adaptation that serves to focus humans and other organisms on actions needed for survival and reproduction. Either way, attention selects a small subset of candidate representations as sufficiently important to enter consciousness (Braun, 2009).

It is widely maintained that attention functions by means of competition among representations (e.g. Desimone and Duncan, 1995; Maia and Cleeremans, 2005; Smith

and Kosslyn, 2007). One way in which competition might work is if the brain contained a process that ranks each representation for importance and them picks the ones with the highest rankings. A more neurologically plausible, parallel process is suggested by a connectionist mechanism of interactive competition that has been used to explain many psychological phenomena, including word recognition, concept application, and theory evaluation. In this mechanism, there is no central ranker of representations, but rather a collection of neuron-like units that compete to have the highest activation by inhibiting each other as well as by being excited by inputs. For example, suppose you are in a park and see an animal moving but are not sure from a brief glimpse whether it is a bird, squirrel, or cat. Figure 1 shows a simplified neural network that takes sensory input between units representing these three interpretations and chooses the most plausible concept by means of a parallel process of interactive competition among the units.

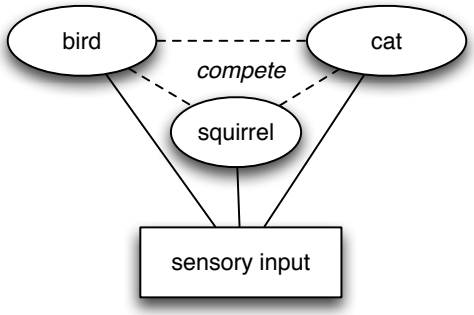

Figure 1. Neural network for competition between three concepts to determine which one best categorizes sensory input. Solid lines are excitatory links, but dotted lines are inhibitory.

Attention can operate by a similar process, except that the representations that compete need to be much more complex than the simple nodes in connectionist networks. The concept of bird or cat or squirrel that comes to conscious awareness carries with it a

large amount of meaning that arises through sensory information and connections with other concepts. Eliasmith's semantic pointer idea provides a plausible answer to how this might work, because concepts understood as semantic pointers decompress or unpack into patterns of neural activation with all the needed associations. Similarly, if what comes to consciousness is a proposition such as "I have found it", then we can hypothesize that the semantic pointer representing this proposition managed to win a competition among various other thoughts that are also candidates for consciousness. Figure 2 shows how this might work, with the semantic pointer for Eureka winning out in the interactive competition over semantic pointers for two other thoughts (perhaps Archimedes thinking about the soap on his face, or about his family).

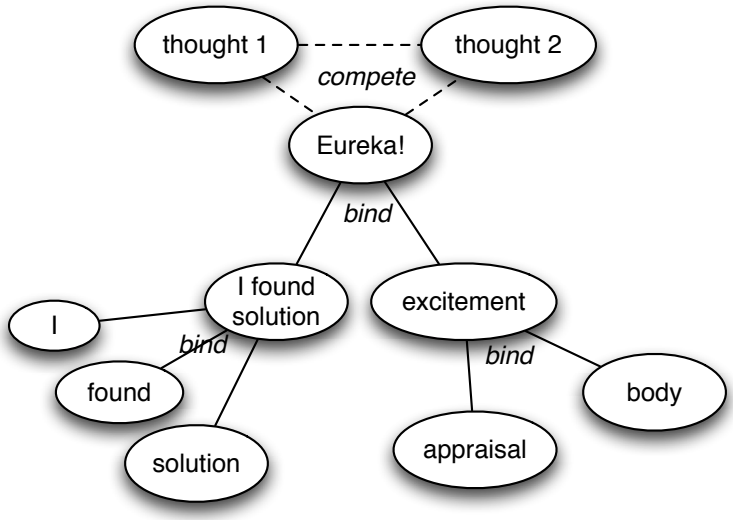

Figure 2. The Eureka semantic pointer is formed by binding processes indicated by the straight lines (which can also be interpreted as excitatory links), and becomes conscious because of competition against other

semantic pointers carried out by inhibitory connections shown as dotted lines.

Of course, the process in the brain is a lot more complicated than figure 2 displays. Each of the semantic pointers requires not just a single node, but a pattern of activation in large populations of neurons. Inhibitory connections are not between single nodes, but between complexes of neurons. Thagard and Aubie (2008) describe how the functions of simple connectionist networks including inhibition can be performed by larger, distributed populations of neurons. Competition need not occur in a single brain area, although it is possible that it is facilitated by convergence zones (association areas) where different kinds of information are conveyed (Driver and Noesselt, 2008). Dehaenes (2009) proposes that there is a *neuronal global workspace* for consciousness that is not a specific location in the brain but rather relies on cortical pyramidal neurons with long-distance connections.

To sum up, Table 1 shows the three neural mechanisms that I conjecture are most important to creative intuition in the Eureka reaction. In the three mechanisms, the interactions of parts of increasing complexity produce increasingly complex emergent results, in an utterly non-mystical sense of emergence. Emergent properties are ones possessed by the whole, not by the parts, and are not simple aggregates of the properties of the parts because they result from interactions of the parts (Bunge, 2003; Wimsatt, 2007). Of course there are other relevant mechanisms, such as the biochemical reactions involving proteins and messengers that operate within neurons and other cells. At a much higher level, social mechanisms are also relevant, because interactions between

people such as conversation produce changes in attention relevant to consciousness and intuition.

Parts	*Interactions*	*Emergent result*
Neurons	Excitation, inhibition, synaptic connections	Representation by firing patterns
Neural populations	Recursive binding	Semantic pointers
Semantic pointers	Interactive competition	Conscious experience

Table 1. Three mechanisms for creative intuition.

Objections

Many philosophers and even some psychologists and neuroscientists would find it outrageous to suggest that an exalted phenomenon like self-consciousness of creativity could be identified with or explained by neural mechanisms. I will now briefly consider four general objections to a neural explanation of creative intuition, concerning robots, category mistakes, conceivability, and what is it like to feel creative.

My contention that creative intuition results from neural representation, recursive binding, and competition among semantic pointers may seem arbitrarily to rule out the possibility that non-human agents such as computers could turn out to have creative intuition. Already there are computer programs capable of generating products that seem to be at least somewhat new and valuable (see Boden, 2004 and a 2009 issue of AI Magazine, vol. 30, no. 3, for candidate examples). The gap between humans and computers should close further as machines continue to increase in speed, memory, and software sophistication.

I certainly do not intend to argue that it is impossible for machines to be creative, and have argued that are already robots capable of representing the world (Parisien and Thagard, 2008). But the mechanisms by which they acquire and use these representations are very different from ones used by people, and I know of no current computers capable of building up representations of representations of representations by anything like the kind of recursive binding I have described as occurring by convolution. For example, Bayes nets are a powerful technology used in some of today's best robots (Thrun, Burgard, and Fox, 2005), but I know of only one discussion of how a Bayes net can represent other Bayes nets (Glymour and Danks, 2007). Moreoever, I am not aware of any analog of interactive competition in computer programs using Bayes nets.

Therefore, although something like creativity might be developed by computers more flexible and intelligent than current ones, I expect that the creative intuitions that robots might have would be very different from the ones that are generated by human brains. This difference has potentially large ethical consequences for the desirability of developing computers capable of intuitions, because their inclinations toward actions will likely be very different from humans: computers lack the biological goals and emotional reactions that are an important part of human ethical intuition known as conscience (Thagard and Finn, 2011). Computers may well someday have intuitions, but I would trust theirs even less than I do human intuitions (Thagard, forthcoming).

A second more philosophical objection to ascribing self-consciousness and creativity to brains is that these properties belong to persons and it is a category mistake to attribute them to a particular part of the body (see Bennett, Dennett, Hacker, and Searle, 2007). The history of science, however, provides ample evidence that categories

change as knowledge advances: for example, we learned from Count Rumford that heat is a kind of motion, not a substance; and we learned from Charles Darwin that humans are a kind of animal, not specially created. Similarly, evidence is mounting that creativity and self-consciousness are kinds of brain processes, not vague properties of vaguer entities called persons.

Another standard move against naturalistic accounts of mental phenomena is arguments of philosophers such as Descartes (1980) and Chalmers (1996) concerning conceivability. We can easily imagine, it is claimed, that there are beings capable of creative intuition that lack the mechanisms of neural representation, recursive binding, and competition among semantic pointers. Hence these neural mechanisms are not essential to creative consciousness. This argument has no force, however, for it would rule out many of the most important scientific discoveries (Thagard, 2010). We can imagine that heat is not the motion of molecules, that lightning is not electrical discharge, and that humans are not animals evolved by natural selection. But in all these cases, there is ample evidence from observation and experiment to conclude that imagination yields falsehoods. Thought experiments intended to block evidence-based theoretical conclusions are prime examples of *un*creative intuitions: they serve purely to maintain ideas that are old and useless rather than new and valuable.

Finally, I need to address the standard philosophical argument that any neural account of the Eureka phenomenon leaves out a crucial aspect of consciousness: what it feels like to have the self-conscious experience of Archimedes and other discoverers. Some philosophers even write of the "what-it-is-likeness" of experience, which they place beyond the reach of scientific explanation. Lumping all the richness of conscious

experience into something ineffable is akin to the strategy of nineteenth century biologists to explain life in terms of some mysteriousness vital force. Now we know that life is the result of many different mechanisms such as genetic transmission, metabolism, and cell division. Similarly, I predict that the varieties of conscious experience will someday be recognized as the result of various mechanisms including the ones for representation, binding, and competition that I have been discussing. For example, one of the important aspects of the Eureka feeling is the highly positive emotion of excitement. Much is already known about how positive emotions arise from brain activity, so this aspect of what it is like to have self-consciousness of creativity is already well on its way to being explained (Thagard and Aube, 2008; Rolls, 2005). It is reasonable to expect that other aspects of self-consciousness of creativity will also prove amenable to mechanistic explanation through advances in theoretical and experimental cognitive neuroscience. Once what-is-it-likeness is broken down into its components, it becomes explainable rather than ineffable.

General Discussion

This chapter has proposed that creative intuition involves self-consciousness of creativity, and that three neural mechanisms – representation, binding, and competition - are at the core of self, consciousness, and creativity. All require representation by populations of spiking neurons, binding of representations by a process like convolution into semantic pointers, and interactive competition among those pointers. Numerous important issues remain, such as the empirical evidence for this account and its implications for the general question of when intuitions are rational.

At the empirical level, it is natural to be concerned about the extent of experimental evidence for the three mechanisms proposed here. Ideally, it would be good to have both neurological evidence for the occurrence of the mechanisms in the brains of humans and other organisms, and psychological evidence concerning aspects of consciousness that are best explained by those mechanisms. There is abundant experimental evidence supporting the idea that neural representations operate by populations of spiking neurons that become tuned to occurrences in the environment (Gerstner and Kistler, 2002); but I know of no direct tests of the idea that such representations get bound together by a process like convolution. Similarly, although the process of competition among representations has often been used by psychologists to explain phenomena concerning attention, I have not seen any direct evidence based on observations of brains that support the existence of the mechanism. There have been some recent brain scanning experiments observing the neural correlates of some simple insight phenomena (Kounios and Beeman, 2009), but their relevance to the more general question of creativity are not clear. With respect to creative intuition, theoretical neuroscience seems to be out in front of experimental neuroscience and psychology, but I hope that this gap will shrink through future research.

It is widely believed that creative new ideas often occur to people when they are relaxing after pursuing difficult problems, as when Kekulé reported dreaming the structure of benzene. The interactive competition view of consciousness might be able to explain this. When you are working hard on a problem, currently active ideas may suppress new semantic pointers that have been formed and prevent them from entering consciousness. At leisure, however, you may not have such pressing thoughts, enabling

the new combination to enter consciousness. Incubation is the process of unconscious combination of ideas (convolution of neural representations) eventually leading to awareness of discovery when problem-solving semantic pointers win the competition to become conscious.

At the theoretical level, the triple mechanisms account needs to be related to other explanations of consciousness. My proposal seems broadly compatible with philosophical accounts of consciousness in terms of higher order representation and perception (e.g. Carruthers, 2011), but is far more specific. Similarly, the three mechanisms proposed might be understood as a mechanistic specification of global workspace theories of consciousness that have been popular among psychologists and neuroscientists (e.g. Dehaene, 2009), but a systematic comparison remains to be done. An open question is whether the view of consciousness as interactive competition among semantic pointers also applies to other domains of consciousness, such as sensory experience, verbal thinking, and emotions.

I have proposed the three mechanisms of neural representation, recursive binding, and semantic pointer competition specifically to explain creative intuition, but would not hesitate to see them as important for intuition in general. Whereas creative intuition primarily functions to generating hypotheses, intuition is sometimes defended as a basis for believing them. For example, thought experiments that are used to produce intuitions are thought by some philosophers to contribute to the justification of theories in science as well as philosophy (Brown, 1991).

If intuition were divine inspiration or Platonic grasping of eternal entities, then there might be some justification in taking intuition as probative rather than merely

suggestive. But the three-mechanism account of intuition provides grounds for skepticism about the trustworthiness of intuitions. I have no objections to the creative side of intuition, because its purpose is just to generate ideas that can then be tested for value. But many ideas that initially seem to the inventor to be new and valuable eventually turn out be weak on novelty, importance, or both. Similarly, ideas that seem intuitive may just be the result of unconscious prejudices that emerge into consciousness with misleading force that is not commensurate with their value. Such illusions are common in philosophy, in both the analytic and phenomenological traditions, where advocates defend their theories based on intuitions that derive from stories (grandiosely called thought experiments) that they themselves have made up to support their own view. For critiques of the use of thought experiments in philosophy to generate intuitions that are mistaken for evidence, see Thagard (2010, forthcoming).

I am not saying that intuitions and thought experiments are useless. As many scientific cases show (e.g. Einstein thinking about relativity by imagining riding on a beam of light), thought experiments can contribute to valuable scientific discoveries. Moreover, they are sometimes useful in identifying difficulties in alternative theories, as when Galileo showed a serious problem with the Aristotelian view that heavy objects fall faster than lighter ones by imagining the fall of a heavy object and a light one tied together. Good scientific thought experiments stimulate inquiry, whereas philosophical intuitions often serve to block inquiry. There is no way of telling from the conscious aspects of intuition whether it is based on reliable evidence or feeble prejudice; hence intuitions should always be subject to rational scrutiny rather than taken at face value. In some cases, such as when an intuition results from large amounts of experience that

corresponds to reality, intuitions may in fact be veridical, but that can only be determined by subsequent investigation. My recommendation is not to trust anyone's intuitions, including your own, until you can evaluate the evidence that underlies them. Otherwise, the representations, bindings, and competition that produce the intuition may be misleading. Intuitions can just as easily result from motivated and fear-driven inference where your emotions distort your beliefs as from reliable patterns of inference (Thagard and Nussbaum, forthcoming).

To conclude, let me observe that the explanation of creative intuition in terms of neural mechanisms makes it clear why creativity has often been thought of as a divine or mysterious process. Common sense and introspection can tell us nothing about mechanisms like spiking neurons, binding, and competition, so it is not surprising that pre-scientific explanations of creativity have looked to supernatural factors such as the Muses and Platonic apprehension of ideas. Fortunately, because of empirical and theoretical advances in neuroscience, the veil of mystery is rapidly lifting from the face of creativity, and intuition can be elevated from the ineffable to the comprehensible. Self-consciousness of creativity in the Eureka experience is becoming a natural phenomenon open to mechanistic explanation.

Acknowledgements: I am grateful to Chris Eliasmith for comments on an earlier draft, and to the Natural Sciences and Engineering Research Council of Canada for funding.

REFERENCES

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

Bennett, M., Dennett, D., Hacker, P., & Searle, J. (2007). *Neuroscience and philosophy*. New York: Columbia University Press.

Blouw, P., Solodkin, E., Eliasmith, C., & Thagard, P. (2012). Concepts as semantic pointers: A theory and computational model. *unpublished manuscript, University of Waterloo*. .

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

Braun, J. (2009). Attention and awareness. In T. Bayne, A. Cleeremans & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 72-77). Oxford: Oxford University Press.

Brown, J. R. (1991). *The laboratory of the mind*. London: Routledge.

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

Carruthers, P. (2011). Higher-order theories of consciousness. *Stanford Encyclopedia of Philosophy*. Retrieved from http://plato.stanford.edu/entries/consciousness-higher/

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

Darwin, C. (1987). *Charles Darwin's notebooks, 1836-1844*. Ithaca: Cornell University Press.

Dehaene, S. (2009). Neurognal global workspace. In T. Bayne, A. Cleeremans & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 466-470). Oxford: Oxford University Press.

Descartes, R. (1980). *Discourse on method and Meditations on first philosophy* (D. Cress, Trans.). Indianapolis: Hackett.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193-222.

Driver, J., & Noesselt, T. (2008). Multisensor interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron, 57*, 11-23.

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. (forthcoming). *How to build a brain*. Oxford: Oxford University Press.

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.

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

Gerstner, W., & Kistler, W. (2002). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge: Cambridge University Press.

Glymour, C., & Danks, D. (2007). Reasons as causes in Bayesian epistemology. *Journal of Philosophy, 104*, 464-474.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*, 295-355.

Hummel, J. E., & Holyoak, K., J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*, 220-264.

Kaufman, J. C., & Sternberg, R. J. (Eds.). (2010). *The Cambridge handbook of creativity*. Cambridge: Cambridge University Press.

Koestler, A. (1967). *The act of creation*. New York: Dell.

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2003). Mental imagery: Against the nihilistic hypothesis. *Trends in Cognitive Sciences, 7*, 109-111.

Kounios, J., & Beeman, M. (2009). The *Aha!* moment: The cognitive neuroscience of insight. *Current directions in psychological science, 18*, 210-216.

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

Maia, T. V., & Cleeremans, A. (2005). Consciousness: Converging insights from connectionism modeling and neuroscience. *Trends in Cognitive Neuroscience, 9*, 397-404.

Markus, H., & Kunda, Z. (1986). Stability and malleability of the self-concept. *Journal of personality and social psychology, 51*(4), 858-866.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review, 69*(220-232).

Parisien, C., & Thagard, P. (2008). Robosemantics: How Stanley the Volkswagen represents the world. *Minds and Machines, 18*, 169-178.

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

Revonsuo, A. (2009). Binding problem. In T. Bayne, A. Cleeremans & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 101-105). Oxford: Oxford University Press.

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

Schröder, T., Stewart, T. C., & Thagard, P. (forthcoming). Intention, emotion, and action in the brain: A neurocomputational model

Schröder, T., & Thagard, P. (forthcoming). The affective meanings of automatic social behaviors: Three mechanisms that explain priming.

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, D. (1792). *Elements of the philosophy of the human mind*. London: Strahan, Cadell, and Creech.

Stewart, T. C., & Eliasmith, C. (2012). Compositionality and biologically plausible models. In W. Hinzen, E. Machery & M. Werning (Eds.), *Oxford handbook of compositionality* (pp. ADD). Oxford: Oxford University Press.

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

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

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

Thagard, P. (2010). *The brain and the meaning of life*. Princeton, NJ: Princeton University Press.

Thagard, P. (2012). *The cognitive science of science: Explanation, discovery, and conceptual change*. Cambridge, MA: MIT Press.

Thagard, P. (in press). The self as a system of multilevel interacting mechanisms. *Philosophical Psychology*.

Thagard, P. (forthcoming). Thought experiments considered harmful.

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.

Thagard, P., & Finn, T. (2011). Conscience: What is moral intuition? In C. Bagnoli (Ed.), *Morality and the emotions* (pp. 150-159). Oxford: Oxford University Press.

Thagard, P., & Nussbaum, A. D. (forthcoming). Fear-driven inference: Emotions in model-based reasoning. In L. Magnani (Ed.), *Model-based reasoning in science and technology*. Berlin: Springer.

Thagard, P., & Schröder, T. (forthcoming). Emotions as semantic pointers: Constructive neural mechanisms. In L. F. Barrett & J. A. Russell (Eds.), *The psychological construction of emotions*. New York: Guilford.

Thagard, P., & Stewart, T. C. (2011). The Aha! experience: Creativity through emergent binding in neural networks. *Cognitive Science, 35*, 1-33.

Thagard, P., & Wood, J. V. (forthcoming). Eighty phenomena to be explained by a theory of the self. *Unpublished, University of Waterloo*.

Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic robotics*. Cambridge, MA: MIT Press.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings*. Cambridge, MA: Harvard University Press.