Chapter 18
Scientific Concepts as Semantic Pointers

Paul Thagard
University of Waterloo

Draft 4, February, 2011

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

What are Concepts?

The discussion of conceptual change in part IV assumed that concepts are an important part of scientific knowledge, but largely ignored the crucial question of what concepts are. Previous chapters mentioned many important science concepts such as *life, mind,* and *disease*. I have taken for granted the standard cognitive science assumption that concepts are mental representations, but have not provided a theory of what concepts are. This chapter draws on ideas of Chris Eliasmith (forthcoming) to argue that concepts, including scientific ones, are semantic pointers - neural processes with powerful semantic, syntactic, and pragmatic capabilities.

Most generally, concepts are representations corresponding to individual words that stand for classes of things; for example the concept *car* corresponding to the word "car" that stands for the class of cars. The history of philosophy and cognitive science has witnessed many different interpretations of concepts, including:

- Concepts are abstract entities, for example the forms of Plato.
- Concepts are copies of sense impressions, for example the ideas of Hume (1888).
- Concepts are data structures that depict prototypes, for example the frames of Minsky (1975).
- Concepts are distributed representations in neural networks, for example the schemas of Rumelhart and McClelland (1986).

Another possibility is that concepts do not exist and should be eliminated from scientific discussions (Machery 2009).

Rather than eliminating concepts, however, I think cognitive science needs to develop a robust theory of them that can contribute to the aims of all the constituent disciplines, including:

- Psychology: explain behavioral experiments about the use of words and address questions about innate vs. learned knowledge.
- Philosophy: provide answers to epistemological questions about the nature of knowledge and meaning, and answers to metaphysical questions about what exists.
- Linguistics: explain how people produce and understand language.
- Neuroscience: explain neurological observations such as the loss of verbal ability in kinds of agnosia.
- Anthropology: explain cultural differences in language and categorization.
- Artificial intelligence: generate new data structures and algorithms for knowledge representation and language understanding.

More specifically, the cognitive science of science needs a theory of concepts for characterizing the structure and growth of scientific knowledge.

This chapter proposes a new approach to scientific concepts using the remarkable new theory of semantic pointers developed by Chris Eliasmith (forthcoming). A semantic pointer is a kind of neural representation whose nature and function is highly compatible with what is currently known about how brains process information. Semantic pointers are neural processes that (1) provide shallow semantics through relations to the world and other representations, (2) can be expanded to provide deeper semantics with relations to perceptual, motor, and emotional information, and (3) support complex syntactic operations. Semantic pointers are naturally represented mathematically by vectors in high-dimensional spaces, just as forces are physical processes that are naturally represented mathematically in 2-dimensional spaces where the dimensions indicate magnitude and direction.

To take a simple example, consider the concept *car*. The concept of car has important semantic functions such as referring to cars in the world, important syntactic functions such as the generation of sentences like "Electric cars help the environment", and important pragmatic functions such as helping to capture the intentions of speakers in particular contexts. Whereas traditional approaches to formal logic and linguistics take syntax as central and treat semantics and pragmatics as add-ons, Eliasmith's semantic pointer hypothesis shows how concepts construed as neural processes can simultaneously have syntactic, semantic, and pragmatic functions.

Neurons have synaptic connections that enable them to fire in patterns in response to inputs from other neurons. These patterns of firing can function syntactically as a result of the binding operations described in chapter 8: two concepts can be bound into a combined concept by convolution, and the same process can produce syntactic entities of great complexity (Eliasmith forthcoming). Such neural representations carry only partial, compressed meaning, but they can point to patterns of firing in neural populations that are much semantically richer by virtue of their relations to perceptual, motor, and emotional representations. The emotional associations of concepts contribute to the pragmatics of concept use. For example, if your past experiences with cars have led you to like them, then your future actions will be disposed to activities that use them. The concept of car as a semantic pointer has an emotional component by virtue of connections between neural populations that carry the pointer and neural populations that encode goals. This component allows concepts to represent values as described in chapter 17. Hence concepts construed as semantic pointers participate in processes where brains are sensitive to context and goals, enabling concepts to contribute to the pragmatics of language and decision.

In computer science, a pointer is a special kind of data type, with properties different from more familiar data types such as bits (0, 1), integers, and strings of characters. A pointer has a value that directs a program to a place in computer memory that stores another value, roughly the way that a street address directs a person to a place in a town where a house is located. Eliasmith's semantic pointers are a generalization of this idea, in that they can refer to multiple locations in a neural memory in a way that endows them with multifarious meanings. Semantic pointers have both a partial meaning that suffices for them to participate in syntactic operations such as forming sentences and entering into inferences, but also a deep meaning that can be accessed by reference to multiple memory locations with perceptual, motor, and emotional information.

Semantic pointers are analogous to compressed computer files that throw away much information but are still adequate for their intended purpose. For example, the song files that store music in applications such as iTunes and mp3 players drop much of the information contained in an analog recording, but still contain the digital data needed for a speaker to reproduce music satisfying to almost all listeners. Similarly, semantic pointers drop much sensory detail about the external world, but are still able to participate in the many kinds of inference involved in perception, motor control, and even high-level reasoning. Like compressed audio files, semantic pointers are more efficient to transport and manipulate than uncompressed information, which pointers can regenerate when it is needed for deeper processing.

Mathematically, semantic pointers can be described as vectors in a high-dimensional space. A vector is a mathematical structure that represents 2 or more quantities. A simple example is a vector that represents a car travelling at 100 kilometers per hour headed east, which is a 2-dimensional vector that can be expressed as a structure (100, 90) where 100 is the speed in kilometers and 90 is the angle in degrees. The space in which semantic pointers operates includes hundreds of dimensions derived from information relevant to their verbal, perceptual, or motor uses. The mathematical representation of semantic pointers is very useful for exploring their syntactic and semantic functions; but such explorations are highly technical, so the rest of this chapter will stick to the mechanistic terminology of patterns of firing in neural populations. A more general defense of the desirability of thinking of concepts and other mental representations as neural processes can be found in Thagard (2010a).

For the semantic pointer interpretation of concepts to contribute to the cognitive science of science, it should help to provide answers to questions such as the following:

1. How are scientific concepts meaningful, especially theoretical ones that are distant from sense experience?
2. How can scientific concepts contribute to scientific explanations?
3. How can scientific concepts contribute to discoveries and conceptual change?

4. How can scientific concepts contribute to the practical goals of science?

I will now show the relevance of semantic pointers to these questions by considering the nature of three fundamental scientific concepts from physics, chemistry, and biology: *force, water,* and *cell*.

Force

The concepts *force, water,* and *cell* are central in their respective sciences, and all have multimodal aspects that fit well with the semantic pointer view of concepts. By "multimodal" I mean that these concepts are mental representations that are not only verbal but also contain information tied to sensory, motor, and emotional modalities.

Consider first the concept of force, which is fundamental in modern physics (Jammer 1957). Physics students are taught gravitational force and move on to learn about other forces such as electromagnetic, frictional, viscous, adhesive, chemical, molecular, and nuclear. According to current physical theory, there are four fundamental forces: gravitational, electromagnetic, and the strong and weak forces that operate at the atomic level. Commonly, a force is defined as an influence that causes a body to undergo a change in speed, direction, or shape, but this definition is not very helpful, since *influence* seems to be much the same concept as *force,* and the concept of *cause* is notoriously difficult to define. Introductory physics texts often define a force as a push or a pull, which seems anthropomorphic rather than scientific.

Positivists such as Ernst Mach worried that there was something illegitimate about the concept of force, despite its centrality in Newtonian physics. Following the empiricist philosophy of Locke, Berkeley, and Hume, they assumed that the meaning of concepts derives from sense experience, which made Newton's idea of force suspect, since force as characterized by Newton's laws is not observable. Force equals mass times acceleration, but unlike mass and speed, force is not directly measurable. Science educators remain puzzled about how to teach students about the nature of force (Coelho 2010).

The mysteriousness of the meaning of *force* dissipates if one considers both the history of the concept and the semantic pointer interpretation of concepts. According to Jammer (1957, p. 7), the concept of force originated in familiarity with human will power and muscular effort and became projected onto inanimate objects as a power dwelling in them. Initially, this power was construed mentally, so that the force in objects was viewed as depending on spirits or divine action. By the seventeenth century, however, the mental construal of force had dropped out through the influence of Kepler, Galileo, and others.

In contrast to the association with mental activity, the muscular effort association of force survives in the common idea that a force is a push or a pull. According to Jammer (1957, p. 17):

> The idea of force, in the prescientific stage, was formed most probably by the consciousness of our effort, spent in voluntary actions, as in the immediate experience of moving our limbs, or by the consciousness of the feeling of a resistance to be overcome in lifting a heavy object from the ground and carrying it from one place to another.

How do we know what pushes and pulls are? They are not easily definable in terms of other words, but rather involve visual, tactile, and kinesthetic sensations. When I push a box, I can see and feel myself in contact with the box, and just as important I have the kinesthetic sense of my arms and the rest of my body moving. Pushing is not captured by a single kind of sense experience, but rather by a combination of visual, tactile, and kinesthetic sensations. Pulling has a different combination involving another set of motions of muscles and body parts. People's concepts of pushes and pulls are not abstract or purely verbal entities, but rather amount to representations involving several sensory modalities. The fact that pushes and pulls require a combination of modalities shows that they are not simply copies of sense impressions as Hume assumed.

On the semantic pointer interpretation of concepts, mental representation of pushes and pulls is a neural process involving firing in populations of spiking neurons linked to brain areas capable of visual, tactile, and kinesthetic representations. The relevant areas are mainly in visual cortex, somatosensory cortex, and motor cortex, respectively. The highly distributed neural population that represents pushes combines these modalities together using the sort of binding discussed in chapter 8. The deep semantics of the concept of pushing comes from the multiple sensory modalities that perceive pushing by oneself or by others, although particular uses of the concept of pushing need not access the full range of sensory activations. Hence, the neural population most active when we think of pushing can provide a compressed representation whose shallow semantics involves correlations with other neural populations for kinds of objects related to pushing, such as doors and people. Kinesthetic representations of force can also arise from experiences with magnets, including the feeling of pulling that comes when a magnet attracts metal, but also pushing when two magnets repel each other.

Obviously, concepts in modern physics go far beyond muscular representations of pushes and pulls. Newton's *Principia* contains the beginning of the modern concept of force, extending ideas developed by Kepler and Galileo. Here are Newton's three famous laws of motion (Jammer 1957, p. 123).

> I. Every body continues in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by force impressed upon it.
>
> II. The change of motion is proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed.
>
> III. To every action there is always opposed an equal reaction; or the mutual actions of two bodies upon each other are always equal, and directed to contrary parts.

Here force is not restricted to the actions of human bodies, but can be attributed to other actions such as the motion of planets. Newton's concept of force, however, did not come out of the blue, but built on his sensory understanding of force derived from experiences with pushes and pulls.

Newton expressed his second law in words, but it naturally has mathematical expressions: $F=ma$ or $F=d(mv)/dt$. The reason that mathematical concepts are so useful is that they provide a compression similar to what happens with semantic pointers. It is hard to think abstractly about what forces are as causes of acceleration and as pushes or pulls, but mathematical representations such as vectors discard the many associations of deep semantics and provide symbolic representations that can be manipulated syntactically in mathematical operations such as proofs and calculations. Numbers and variables have the same effect: it is mentally easier to manipulate “7” than “7 things”, or “x” rather than “an unknown number”. Mathematical representations dramatically reduce the load on short-term working memory, just as semantic pointers do. Hence it is useful to represent forces by vectors in 2-dimensional spaces, and semantic pointers as vectors in higher-dimensional spaces.

Thus the concept of force has its origins in multimodal sensory experience, but through verbal and mathematical theories can be expanded into the amazingly rich kind of representation found in the minds of physicists. But the accretion of large amounts of verbal and mathematical information to the concept of force does not overturn its connections with sensory-motor experience.

Water

For a chemical example, consider the concept of water. Like *force, water* has a substantial prescientific meaning, and the semantic pointers that operate in ordinary people’s brains tie water to many sensory experiences, including taste and smell as well as vision and touch. Sound is also tied to the concept of water through the familiar rhythms of waves, and there is also kinesthetic experience of water familiar through wading and swimming. Thus the neural processes that encode the concept of water point to experiences in many modalities.

Like force, however, water can take on diverse theoretical roles, extensively reviewed by Grisdale (2010). The early Greek philosopher Thales proposed that water is the basic element out of which everything else is formed. Aristotle disagreed and proposed that there were four other elements besides water: earth, air, fire, and aether. Water was viewed as an element for more than two thousand years until the 1700s, when researchers such as Lavoisier provided evidence that water is actually a compound rather than an element, consisting of hydrogen and oxygen. The association of water with H_2O is now part of the conceptual structure of anyone with rudimentary instruction in science. This part of the representation of water is largely verbal, although images of water showing two hydrogen atoms joined to an oxygen atom are also available. Today, quantum theory explains how bonds form between hydrogen and oxygen atoms through electron attractions.

Despite all this theoretical progress, the concept of water for scientists retains the multimodal associations it has for ordinary people. The concept of water has changed enormously from the ancient mythological views that tied different forms of water to various gods. The view of water as element as been replaced by a far more explanatorily successful view of water as a specific kind of compound. Nevertheless, it is legitimate to say that the concept of water was changed rather than abandoned because of the retention of multimodal information carried by the semantic pointers of people from the ancient Greeks to today.

Grisdale's (2010) discussion of modern conceptions of water refutes a highly influential thought experiment that the meaning of water is largely a matter of reference to the world rather than mental representation. Putnam (1975) invited people to consider a planet, Twin Earth, that is a near duplicate of our own. The only difference is that on Twin Earth water is a more complicated substance XYZ rather than H_2O. Water on Twin Earth is imagined to be indistinguishable from H_2O, so people have the same mental representation of it. Nevertheless, according to Putnam, the meaning of the concept of water on Twin Earth is different because it refers to XYZ rather than H_2O. Putnam's famous conclusion is that meaning just ain't in the head.

The apparent conceivability of Twin Earth as identical to Earth except for the different constitution of water depends on ignorance of chemistry. As Grisdale (2010) documents, even a slight change in the chemical constitution of water produces dramatic changes in its effects. If normal hydrogen is replaced by different isotopes, deuterium or tritium, the water molecule markedly changes its chemical properties. Life would be impossible if H_2O were replaced by heavy water, D_20 or T_20; and compounds made of elements different from hydrogen and oxygen would be even more different in their properties. Hence Putnam's thought experiment is scientifically incoherent: if water were not H_2O, Twin Earth would not be at all like Earth.

This incoherence should serve as a warning to philosophers who try to base theories on thought experiments, a practice I have criticized in relation to concepts of mind (Thagard 2010a, ch. 2). Some philosophers have thought that the non-material nature of consciousness is shown by their ability to imagine beings (zombies) who are physically just like people but who lack consciousness. It is entirely likely, however, that once the brain mechanisms that produce consciousness are better understood, it will become clear that zombies are as fanciful as Putnam's XYZ. Just as imagining that water is XYZ is a sign only of ignorance of chemistry, imagining that consciousness is non-biological may well turn out to reveal ignorance rather than some profound conceptual truth about the nature of mind. Of course, the hypothesis that consciousness is a brain process is not part of most people's everyday concept of consciousness, but psychological concepts can progress just like ones in physics and chemistry (Thagard, forthcoming-c).

Zombies aside, we should draw the lesson that a theory of scientific concepts should be based on scientific evidence, not thought experiments. The semantic pointer interpretation of concepts shows how to integrate the pre-scientific, multimodal understanding of concepts based on vision and other sensory modalities with the theoretical developments about the constitution of elements provided by chemistry and quantum theory. Hence concepts can undergo dramatic change while retaining sensory continuity.

Cell

Biology also undergoes conceptual change, and no concept in biology is more central than *cell,* representing the smallest unit of life. This concept is multimodal in ways that make sense from the perspective of the semantic pointer view of concepts. Unlike force and water, people have no direct sensory experience of cells, which were first observed around the 1660s when Robert Hooke (1665) used a new instrument, the microscope, to examine a piece of cork. He identified small bounded areas in cork that he variously called pores, boxes, and cells, adapting the latter term from the Latin word for a small room. It was only in the 1830s that theories of the formation and function of cells were developed, thanks to improved observations using microscopes with achromatic lenses (Bechtel 2006). Detailed observations of the structure of the cells only became possible in the 1940s through the development of the electron microscope that could identify small structures such as mitochondria.

In all three main stages of development, the concept of a cell had a substantial visual component, even though plant cells (unlike rooms) are not directly observable. Hooke's microscope enabled him to see similarities between cells and objects (pores, boxes, room) that are observable and for which people already had concepts. The new concept of cell arose by a process of visual analogy, combining the observation of new structures within plants with past observations of familiar objects. Hence Hooke's concept of a cell as a component of organic material such as corks and carrots was in part a pointer to the visual representation that he acquired through his own observations and that others could acquire through looking at the elegant illustrations in his book *Micrographia*. Much richer visual representations arose with better microscopes leading to the beautiful color illustrations of cells in modern textbooks and videos of mitosis that can easily be found on the Web. Neurons are nerve cells that became observable with optical microscopes in the 1870s thanks to improved staining techniques; the fine structure of neurons structure as synaptic connections became observable in the 1940s via electron microscopes. In this way, the visual aspect of cells carried in the deep semantics of the semantic pointer for *cell* allowed the concept to retain some continuity despite major theoretical changes.

I have argued that three central scientific concepts – *force, water*, and *cell* – can be understood as semantic pointers. All three have a substantial multimodal aspect accommodated by the expansion capabilities of semantic pointers, while retaining the capacity to serve as symbols in theories in physics, chemistry, and biology. After further discussion of meaning, I will argue that the semantic pointer hypothesis applies to concepts generally.

Meaning

Philosophers have long debated about the source of the meaning of concepts. Here are five possible answers to the question why concepts are meaningful:

1. Concepts are meaningful because of sense experience (Hume 1988).
2. Concepts are meaningful because they are innate.
3. Concepts are meaningful because they refer to things in the world (Putnam 1975).
4. Concepts are meaningful because they have functional, procedural roles in relation to other concepts (Harman 1987, Miller and Johnson-Laird 1976).
5. Concepts are meaningful because they are used socially in communication (Wittgenstein 1968).

The semantic pointer view of concepts does not force a choice among these answers construed as alternative theories, but rather shows how multiple processes can contribute to the meaning of concepts. The five sources of meaning are not completely distinct, because sensory experience, innateness, reference, and social use can all contribute to functional roles, and reference through interaction with the world can contribute to sense experience. Nevertheless, it would be futile to try to reduce the five sources of meaning to a smaller, fundamental set, ignoring the other kinds of interactions that provide the full relational range of meaning.

I will call this comprehensive view the *multirelational* theory of concept meaning because it proposes that the meaning of a concept derives from many processes that can affect patterns of neural firing that constitute semantic pointers. I avoid talking of the *content* of concepts because that term misleadingly suggests that meaning is a thing rather than a relational process. Similarly, we should not talk of the meaning of a sentence as a content or a proposition, which philosophers often construe as an abstract entity for whose existence there is no evidence. The conceptual change of thinking of meaning as a process rather than a thing is analogous to the important historical shift of thinking of mass as a quantity like weight to thinking it as the result of a process relating multiple objects.

In order to connect the multirelational view of meaning with the theory of concepts as semantic pointers, I need to show how semantic pointers relate to sense experience, innateness, reference, functional role, and social uses. The neural processes that constitute concepts are often causally correlated with sense experience, most obviously in sensory concepts such as *heavy* and *blue*. Neural populations can encode sensory phenomena because particular neurons become tuned to different aspects of sensory inputs. However, the semantic pointer view of concepts is not restricted to a narrowly empiricist view because the synaptic connections that generate patterns of firing may be the result of internal interactions with other neural populations, not directly with sensory inputs.

Moreover, some of the synaptic connections that generate patterns of firing may be innate, as is perhaps the case with a few core concepts like *object* or *face*. The issue of the extent of innateness of concepts is highly controversial, with current views ranging from ones that minimize innateness in favor of powerful learning capabilities (e. g. Quartz and Sejnowsky 1996, Elman et al 1996) to ones more inclined to posit innate concepts (e.g. Carey 2009). The semantic pointer hypothesis is neutral on the question of innateness, and is compatible both with meaning arising from sense experience and with meaning arising from innate connections instilled by natural selection.

Perhaps the concept of *face* is innate in humans, as suggested by the finding that infants orient to and respond to faces shortly after birth (Slater and Quinn 2001). This concept is clearly not just verbal, as the language resources of newborns are at best limited, but involves visual representations that indicate the appropriate structure of eyes, nose, and mouth. Face representation is dynamic, as infants respond to changes such as smiles. Supposing that the concept of face is innate requires infants to be born with neural populations that detect features of faces and the overall configuration that signifies a face. But the concept, construed as a semantic pointer, can operate also at a symbolic level that can generate rule-like behaviors such as: If someone makes a face, then make a face back. To produce these results, all that is needed is to have brain development put in place before birth a set of neurons and synapses that will generate the appropriate firing patterns when the infant receives sensory stimulation. Because such structures are no different from the ones that the semantic interpretation of concepts assumes can be learned from experience, this interpretation can accommodate meaning arising from both sense experience and innateness.

The semantic pointer interpretation makes sense of the origins of sensory concepts, but it can also recognize that for many concepts the correlations with sense experience are highly remote. The remoteness is particularly acute with theoretical concepts in science such as *electron, virus,* and *black hole*. Positivist philosophers worried about how such concepts could be meaningful when they are not translatable into sense experience, but they never justified their contention that all meaning must derive just from the senses. Concepts can also gain their meaning from interactions with other concepts, which need not be definitional: very few terms outside mathematics have strict definitions, as the above examples of force, water, and cell confirm.

It is easy to see how semantic pointers can get some of their meaning from other semantic pointers, both through low level processes like spreading activation and stronger more rule-like associations. For example, the meaning of the concept *car* derives in part from its association with kinds like *vehicle* and with parts like *engine*. However, whereas the semantic pointer for *car* can be expanded into pictures, sounds, and smells of cars, there is no similar sensory expansion of *electron* and *black hole,* because we have no senses or instruments for observing them. Nevertheless, such concepts can be meaningful because the neural processes that constitute them have systematic causal relations to other concepts, including ones tied to sensory experience. For example, we have reason to believe that electrons exist because of their causal effects on many observable results such as lamps going on. The concept *electron* is then associated neurally with the observable concept *lamp*, allowing semantic pointers to get part of their meaning from relations to other concepts.

Whereas empiricist philosophers maintained that concepts get meaning from sense experience, realists see meaning as emanating from relations to the world. From the perspective of a neuropsychological view of how perception works, the sensory view of meaning fits well with the referential view of meaning, because interactions with the world are via the senses. My concept of *car* gets a lot of its meaning from my sense experiences of cars, but the physiology of perception tells us that this sense experience is often the result of causal interactions with cars. For example, light reflects off a car into my eyes and produces my visual experience of a car. The motions of a car engine generate sounds waves that stimulate my ears to produce auditory experiences of a car. Hence reference can be seen to be part of the meaning of *car,* construed as a semantic pointer, via the causal interactions between people's brains and objects in the world. People need not be passive recipients in such interactions, but can use their bodies to manipulate the world in ways that generate new sensory experiences, as when I turn a key to start a car and produce new sounds and sights. For more on neurosemantics, see Eliasmith (2005) and Parisien and Thagard (2008).

There is one more source of the meaning of concepts that must be mentioned. The activation of concepts is not just the result of interactions with the world, but for language-using beings is often also the result of interactions with other people. The associations between semantic pointers described in relation to functional interactions often arise because of communication that people have with others. I need never have seen a penguin to have a concept of penguin, not just because I have seen pictures of penguins, but also because I have heard other people talk of penguins. Conversations generate neural activity, sometimes leading to the acquisition of new concepts, as in the rapid vocabulary increase found in children and other learners. The semantic pointers of one person make possible uses of language that are perceived by other people, and thereby can lead to altered neural processes in the listener, including new concepts.

Thus the semantic pointer interpretation of concepts is consistent with the multirelational view of meaning as potentially deriving from *all* of sense experience, innateness, functional roles, reference, and social use. It should be clear that this view has a ready answer to various conundrums that have plagued the view that meaning belongs to abstract symbols. The symbol-grounding problem that arises with purely computational systems with no relation to the world is clearly not a problem for semantic pointers, which have causal links with senseexperience and reference. Such links need not be special to brains, as robots can have them too (Parisien and Thagard 2008). Semantic pointers are also consistent with the moderate embodiment thesis discussed in chapter 4, because their multimodal expansions operate via the full range of bodily senses, including kinesthesia that affects concepts like *force* and *water*.

However, the semantic pointer view is incompatible with the extreme embodiment thesis that throws out ideas about representation altogether. I see as one of the major virtues of the semantic pointer hypothesis that it shows how cognition can be grounded, embodied, *and* representational. Looking back to chapter 4, it should be clear how semantic pointers can support mental models that have syntactic as well as semantic properties.

What Concepts Are

The sections on force, water, and cells showed that three central scientific concepts can naturally be understood as semantic pointers, but it would be hasty to generalize from a few examples to all scientific concepts, let alone all concepts in other domains. The best way to establish generally that concepts are semantic pointers, i.e. neural structures and processes with the syntactic, semantic, and pragmatic properties described above, would be to show that this supposition explains many results of psychological experiments. There are thousands of experiments about concepts, and ideally semantic pointers could be used to simulate the full range of results. In place of that overwhelming task, it would be desirable to show that semantic pointers can display the kinds of behaviors that have been used by proponents of various psychological theories to explain experimental findings.

Currently in cognitive science, there are three main competing interpretations of the nature of concepts, construing them as prototypes, sets of exemplars, or parts of explanatory theories (Machery 2009, Murphy 2002, Thagard 2005). A unified theory of concepts would have to show that the phenomena supporting all three interpretations can be explained by the same mechanism, in this case semantic pointers. My colleagues and I plan to use computer simulations of neural networks to show that this unification is feasible, but in advance of such results I can speculate about how semantic pointers might be relevant to understanding the full range of properties of concepts described by available theories.

Since the 1970s, many philosophers, psychologists, and computer scientists have advocated a view of concepts as *prototypes*, which are mental representations that specify typical rather than defining properties. Prototypes are more flexible than definitions and there are experimental reasons to think that they give a better account of the psychology of concepts. However, they may not be flexible enough, so some psychologists have claimed that people do not actually store concepts as prototypes, but rather as sets of examples. This claim is called the *exemplar* theory of concepts. The third major account of concepts currently discussed by psychologists is called the *knowledge* view or sometimes the theory-theory. This view points to the large role that concepts play in providing explanations.

Depending on which theory of concepts one adopts, the concept of force, for example, would be variously construed as a representations of typical properties of forces, as set of examples of forces, or as a theoretical structure providing explanations in terms of forces. Thagard (2010a) contends that these are not competing theories of concepts but merely point to different aspects of them that can be unified using sufficiently rich neurocomputational ideas.

Semantic pointers seem to have the desired richness. A neural network such as those trained by back propagation can acquire its connection weights by exposure to a large number of examples and thereby have the ability to access an approximate representation of them. The semantic pointer, however, need not have the full representation of those examples, which would make it unwieldy and incapable of managing the prototype and explanatory functions of concepts. But the semantic pointer can point (through neural connections) to locations in the brain where information about multiple examples is stored multimodally.

For more efficient processing, a semantic pointer can encode a set of typical features that apply to a class of things, in just the way that typical trained neural networks can encode prototypes. A neural population that fires in a recognizable fashion when presented with particular sets of features need not treat those features as necessary or sufficient conditions for the application of a concept. Rather, the neural network responds probabilistically to the presented features and makes an approximate guess about which concept an object possessing a set of features falls under.

Finally, whereas exemplar and prototype theories of concepts have difficulty accounting for their explanatory uses, it is easy to see how semantic pointers can figure in explanations by virtue of their symbol-like ability to figure in approximate causal rules. In accord with the view of causality advocated in chapter 3, the basic human schema for causality is a sensory-motor-sensory pattern that is either innate or acquired within a few months of birth. Through learning, this schema gets generalized into an observation-manipulation-result pattern that goes beyond direct sensory experience and can be further expanded into the technical that comes with such mathematical advances as probability theory and Bayesian networks. People acquire rules of the form concept-cause-result, which could be expressed as a rule-like structure of the form IF x falls under a concept C THEN it has behavior B. For example, if we want to explain why an animal quacked, we could use the causal knowledge that ducks produce quacks.

I am assuming that all of the theoretical uses of concepts – as sets of exemplars, as prototypes, and as explanations – retain access to the multimodal, perceptual representations of concepts. That retention does not fall back into the empiricist view of concepts as simply derived from sense experience, which would not handle the prototype and explanatory uses of concepts, but it shows that even for those uses semantic pointers can retain contact with the empirical.

This section has been highly speculative and should be taken as tentative until simulations can be produced that show that semantic pointers as implemented in Eliasmith’s Neural Engineering Framework can account for all the major characteristics of concepts as revealed in psychological experiments. As chapter 1 described, the purposes of computer simulations include showing that a proposed mechanism is coherent enough to be implemented in a program and powerful enough to generate the desired behaviors. The proof is in the programming.

An alternative to a unified theory of concepts is the skeptical conclusion of Machery (2009) that exemplars, prototypes, and explanations employ very different kind of representations, so cognitive science should simply eliminate concepts from theoretical discussions. Such elimination is unnecessary, however, if semantic pointers can show how concepts, construed as neural processes, can function under different circumstances in ways prescribed by all three ways of understanding them.

According to Machery (2009), philosophers look to concepts to serve a very different purpose from the aim of psychologists to explain experimental results about how people classify things. He says that the purpose of concepts in most contemporary philosophy is to explain how people can have "propositional attitudes" such as beliefs, desires, and hopes. Such mental states are taken as relations between a person and an abstract entity called a proposition. I have argued elsewhere that the philosophical idea of propositional attitudes is a mistake and should be supplanted with ideas about mental representation taken from cognitive science (Thagard 2008). What remains of the abstract philosophical project of understanding propositional attitudes is the question of how concepts, construed naturalistically, can figure in more complex kinds of representations analogous to sentences that make claims about the world.

For this purpose, the semantic pointer interpretation of concepts is highly useful, as there is now a developed theory of how concepts can combine computationally into more complex structures of great complexity. Even embedded sentences such as "That Mary insulted John caused Mary to dislike John" can be encoded by vectors and implemented in neural networks (Eliasmith forthcoming, Eliasmith and Thagard 2001, Plate 2003). We have already seen that semantic pointers can support the many sources of meaning that philosophers have looked for at the concept level, and they can be combined in ways that support the multirelational nature of meaning.

Conclusion

In the late 1970s, I wrote a paper called "Scientific Theories as Frame Systems" that I never published, even though it is one of the best I ever wrote. The reason I never submitted it for publication it is that my ideas about applying the frame theory of Minsky (1975) to understanding scientific knowledge began to seem to me rather vague once I learned computer programming. Many of the ideas from that paper appear in more specific form in my 1988 book *Computational Philosophy of Science*. This chapter has had the same aim as that book and the unpublished paper on frames – to use the resources of cognitive science to go beyond usual approaches to understanding the growth and structure of scientific knowledge. I think that the semantic pointer view of concepts is more powerful than Minsky's frame account, in that it elegantly and precisely accommodates multimodal, exemplar, and explanation aspects of concepts, in addition to prototype aspects.

A major advantage of the semantic pointer view of scientific concepts is that the sensory aspects of force, water, and cells can be retained in their deep, expanded semantics without encumbering their many theoretical functions. Another advantage is the clarification of how concepts like force, water, and cell can change dramatically with new scientific theories while remaining identifiably continuous with earlier concepts. Thomas Kuhn's (1970) theory of scientific revolutions seemed to have the radical implication that meaning change across paradigm shifts is so dramatic that rational evaluation is impossible. Semantic pointers show how the development of concepts can display continuities in the face of dramatic theoretical change, so that scientific developments can be rational even when revolutionary (Thagard 1992). This combination of change and continuity requires a complex theory of meaning that takes into account sensory experience, innateness, reference, functional roles, and concept uses.

Although concepts like *force, water*, and *cell* retain some continuity based on multimodal semantics, it must also be recognized that sometimes theoretical advances do require rejection of assumptions based on sensory experience. We now recognize that, contrary to the original conception, force does not require will and can operate at a distance. Water is not just the familiar liquid and solid, but can also occur in gaseous form and be decomposed into oxygen and hydrogen. Whereas Hooke saw cells as inert walls, we now conceive of them as living entities that can reproduce. It is striking that the five elements of Aristotle have all been dramatically reclassified through theoretical advances. Water, earth, and air have all been reclassified as compounds rather than elements; fire is neither an element or a compound, but a process resulting from rapid oxidation; aether does not exist. A theory of concepts needs to be rich enough to allow for both conceptual continuity and dramatic kinds of conceptual change: the semantic pointer interpretation is up to the challenge.

Another advantage of the semantic pointer view of concepts is that it is part of a general cognitive architecture that surmounts many of the divisions that have arisen in cognitive science in the past few decades. Contrary to criticisms of the mainstream computational-representational approach in cognitive science, Eliasmith's (forthcoming) semantic pointer architecture shows how thinking can be all of the following: psychological and neural, computational and dynamical, symbolic and subsymbolic, rule-governed and statistical, abstract and grounded, disembodied and embodied, reflective and enactive (action-oriented). Achieving these reconciliations, however, requires expansion of the theoretical resources of traditional cognitive science to include the full range of syntactic, semantic, and pragmatic capabilities of semantic pointers.

The arguments in this chapter have not definitively established the semantic pointer interpretation as *the* correct view of scientific concepts or concepts in general. But they show the general power of the neural view of concepts that chapter X used to explain creative conceptual combination. I have described how important science concepts like *force, water,* and *cell* can be understood as semantic pointers. More generally, I have suggested how the exemplar, prototype, and explanatory aspects of concepts can all be understood in terms of semantic pointers. Hence the semantic pointer interpretation of concepts makes a potentially large contribution to the cognitive science of science.

REFERENCES

Bechtel, W. (2006). *Discovering cell mechanisms: The creation of modern cell biology*. New York: Cambridge University Press.

Carey, S. (2009). *The origin of concepts*. Oxford: Oxford University Press.

Coelho, R. L. (2010). On the concept of force: How understanding its history can improve physics teaching. *Science & education, 19*, 91-113.

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

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

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

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Grisdale, C. D. W. (2010). *Conceptual change: Gods, elements, and water*. Unpublished M.A. thesis, University of Waterloo, Waterloo.

Harman, G. (1987). (Nonsolopsistic) conceptual role semantics. In E. LePore (Ed.), *Semantics of natural language* (pp. 55-81). New York: Academic Press.

Hooke, R. (1665). *Micrographia: Or some physiological descriptions of minute bodies made by magnifying glasses with observations and inquiries thereupon*. London: John Martin and James Allestry.

Hume, D. (1888). *A treatise of human nature*. Oxford: Clarendon Press.

Jammer, M. (1957). *Concepts of force*. Cambridge, MA: Harvard University Press.

Machery, E. (2009). *Doing without concepts*. Oxford: Oxford University Press.

Miller, G., & Johnson-Laird, P. (1976). *Language and perception*. Cambridge, MA: Harvard University Press.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211-277). New York: McGraw-Hill.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

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.

Putnam, H. (1975). *Mind, language, and reality*. Cambridge: Cambridge University Press.

Quartz, S. R., & Sejnowski, T. J. (1997). The neural basis of cognitive development: a constructivist manifesto. *Behavioral and Brain Sciences, 20*, 537-556.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge MA: MIT Press/Bradford Books.

Slater, A., & Quinn, P. C. (2001). Face recognition in the newborn infant. *Infant and child development, 10*, 21-24.

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

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

Thagard, P. (2008). How cognition meets emotion: Beliefs, desires, and feelings as neural activity. In G. Brun, U. Doguoglu & D. Kuenzle (Eds.), *Epistemology and emotions* (pp. 167-184). Aldershot: Ashgate.

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

Thagard, P. (forthcoming-c). Conceptual change in cognitive science: The brain revolution. In W. J. Gonzalez (Ed.), *Conceptual revolutions: From cognitive science to medicine*. A Coruña, Spain: Netbiblo.

Wittgenstein, L. (1968). *Philosophical investigations* (G. E. M. Anscombe, Trans. 2nd ed.). Oxford: Blackwell.

The Cognitive Science of Science