How Brains Make Mental Models

Paul Thagard

Abstract. Many psychologists, philosophers, and computer scientist have written about mental models, but have remained vague about the nature of such models. Do they consist of propositions, concepts, rules, images, or some other kind of mental representation? This paper will argue that a unified account can be achieved by understanding mental models as representations consisting of patterns of activation in populations of neurons. The fertility of this account will be illustrated by showing its applicability to causal reasoning and the generation of novel concepts in scientific discovery and technological innovation. I will also discuss the implications of this view of mental models for evaluating claims that cognition is embodied.

1 Introduction

Mental models are psychological representations that have the same relational structure as what they represent. They have been invoked to explain many important aspects of human reasoning, including deduction, induction, problem solving, language understanding, and human-machine interaction. But the nature of mental models and the processes that operate on them has not always been clear from the psychological discussions. The main aim of this paper is to provide a neural account of mental models by describing some of the brain mechanisms that produce them.

The neural representations required to understand mental models are also valuable for providing new understanding of how minds perform abduction, a kind of inference that generates and/or evaluates explanatory hypotheses. Considering the neural mechanisms that support abductive inference makes it possible to address several aspects of abduction, some first

Paul Thagard
Department of Philosophy, University of Waterloo, Waterloo, Canada
e-mail: pthagard@uwaterloo.ca

L. Magnani et al. (Eds.): Model-Based Reasoning in Science & Technology, SCI 314, pp. 447–461.
springerlink.com

proposed by Charles Peirce, that have largely been neglected in subsequent research. These aspects include the generation of new ideas, the role of emotions such as surprise, the use of multimodal representations to produce "embodied abduction", and the nature of the causal relations that are required for explanations.

The suggestion that abductive inference is embodied raises issues that have been very controversial in recent discussions in psychology, philosophy, and artificial intelligence. This paper argues that the role of emotions and multimodal representations in abduction supports a moderate thesis about the role of embodiment in human thinking, but not an extreme thesis that proposes embodied action as an alternative to the computational-representational understanding of mind.

2 Mental Models

How do you solve the following reasoning problem? Adam is taller than Bob, and Bob is taller than Dan; so what do you know about Adam and Dan? Readers proficient in formal logic may translate the given information into predicate calculus and use their encoding of the transitivity of "taller than" to infer that Adam is taller than Dan, via applications of the logical rules of universal instantiation, and-introduction, and modus ponens. Most people, however, report using a kind of image or model of the world in which they visualize Adam as taller than Bob and Bob as taller than Dan, from which they can simply read off the fact that Adam is taller than Dan.

The first modern statement of the hypothesis that minds use mechanical processes to model the world was by Kenneth Craik, who in 1943 proposed that human thought provides a convenient small-scale model of a process such as designing a bridge [3, p. 59]. The current popularity of the idea of mental models in cognitive science is largely due to Philip Johnson-Laird, who has used it extensively in explanations of deductive and other kinds of inference as well as many aspects of language understanding (e.g. [16, 18, 19]. In his history of mental models, Johnson-Laird cites as an important precursor the ideas of Charles Peirce about the class of signs he called "likenesses" or "icons", which stand for things by virtue of a relation of similarity [17]. Earlier precursors may have been Locke and Hume with their idea that ideas are copies of images. Many recent researchers have used mental models to explain aspects of thinking including problem solving [13], inductive learning [15], and human-machine interaction (e.g. [39]). Hundreds of psychological articles have been published on mental models[1].

Nevertheless, the nature of mental models has remained rather fuzzy. Nersessian [26, p. 93] describes a mental model as a "structural, behavioral, or functional analog representation of a real-world or imaginary situation, event or process. It is analog in that it preserves constraints inherent in what is

[1] http://www.tcd.ie/Psychology/other/Ruth_Byrne/mental_models/

represented." But what is the exact nature of the psychological representations that can preserve constraints in the required way? One critic of mental model explanations of deduction dismisses them as "mental muddles" [34].

This paper takes a new approach to developing the vague but fertile notion of mental models by characterizing them in terms of neural processes. A neural approach runs counter to the assumption of mental modelers such as Johnson-Laird and Craik that psychological explanation can proceed at an abstract functional and computational level, but I will try to display the advantages of operating at the neural as well as the psychological level. One advantage of a neural account of mental models is that it can shed new light on aspects of abductive inference.

3 Abduction

Magnani [22] made the explicit connection between model-based reasoning and abduction, arguing that purely sentential accounts of the generation and evaluation of explanatory hypotheses are inadequate (see also [23, 5, 45]). A broader account of abduction, more in keeping with the expansive ideas of Peirce [29, 30], can be achieved by considering how mental models such as ones involving visual representations can contribute to explanatory reasoning. Sententially, abduction might be taken to be just "If p then q; why q? Maybe p". But much can be gained by allowing the p and q in the abductive schema to exceed the limitations of verbal information and include visual, olfactory, tactile, auditory, gustatory, and even kinesthetic representations. To take an extreme example, abduction can be prompted by a cry of "What's that awful smell?" that generates an explanation that combines verbal, visual, auditory, and motor representations into the answer that "Joe was trying to grate cheese onto the omelet but he slipped, cursed, and got some cheese onto the burner".

Moreover, there are aspects of Peirce's original descriptions of abduction that cannot be accommodated without taking a broader representational perspective. Peirce said that abduction is prompted by surprise, which is an emotion, but how can surprise be fitted into a sentential framework? Similarly, Peirce said that abduction introduces new ideas, but how could that happen in sentential schemas? Such ideas can generate flashes of insight, but both insight and their flashes seem indescribable in a sentential framework. Another problem concerns the nature of the "if-then" relation in the sentential abductive schema. Presumably it must be more than material implication, but what more is required? Logic-based approaches to abduction tend to assume that explanation is a matter of deduction, but philosophical discussions show that deduction is neither necessary nor sufficient for explanation (e.g. [35]). I think that good explanations exploit causal mechanisms, but what constitutes the causal relation between what is explained and what gets explained? I aim to show that all of these difficult aspects of abduction

– the role of surprise and insight, the generation of new ideas, and the nature of causality – can be illuminated by consideration of neural mechanisms.

Terminological note: Magnani [24] writes of "non-explanatory abduction", which strikes me as self-contradictory. Perhaps there is a need for a new term describing a kind of generalization of abduction to cover other kinds of backward or inverse reasoning such as generating axioms from desired theorems, but let me propose to call this generalized abduction *gabduction* and retain abduction for Peirce's idea of the generation and evaluation of explanatory hypotheses.

4 Neural Representation and Processing

A full and rigorous description of current understanding of the nature of neural representation and processing is beyond the scope of this paper, but I will provide an introductory sketch (for fuller accounts, see such sources as [2, 5, 10, 27, 42].

The human brain contains around 100,000,000,000 neurons, each of which has many thousands of connections with other neurons. These connections are either excitatory (the firing of one neuron increases the firing of the one it is connected to) or inhibitory (the firing of one neuron decreases the firing of the one it is connected). A collection of neurons that are richly interconnected is called a neural population (or group, or ensemble). A neuron fires when it has accumulated sufficient voltage as the result of the firing of the neurons that have excitatory connections to it. Typical neurons fire around 100 times per second, making them vastly slower than current computers that operate at speeds of billions of times per second, but the massive parallel processing of the intricately connected brain enables it to perform feats of inference that are still far beyond the capabilities of computers.

A neural representation is not a static object like a word on paper or a street sign, but is rather a dynamic process involving ongoing change in many neurons and their interconnections. A population of neurons represents something by its pattern of firing. The brain is capable of a vast number of patterns: assuming that each neuron can fire 100 times per second, then the number of firing patterns of that duration is $(2^{(100)})^{100000000000}$, a number far larger that the number of elementary particles in the universe, which is only about 10^{80}. I call this "Dickenson's theorem", after Emily Dickenson's beautiful poem "The brain is wider than the sky". A pattern of activation in the brain constitutes a representation of something when there is a stable causal correlation between the firing of neurons in a population and the thing that is represented, such as an object or group of objects in the world [9, 28]. The claim that mental representations are patterns of firing in neural populations is a radical departure from everyday concepts and even from cognitive psychology until recently, but is increasingly supported by data

acquired through experimental techniques such as brain scans and by rapidly developing theories about how brains work (e.g. [1, 37, 42]).

5 Neural Mental Models

Demonstrating that neural representations can constitute mental models requires showing how they can have the same relational structure as what they represent, both statically and dynamically. Static mental models have spatial structure similar to what they represent, whereas dynamic mental models have similar temporal structure. Combined mental models capture both spatial and temporal structure, as when a person runs a mental movie that represents what happens in some complex visual situation such as two cars colliding.

The most straightforward kind of neural mental models are topographical sensory maps, for which Knudsen, du Lac, and Esterly [21, p. 61] provide the following summary:

> The nervous system performs computations to process information that is biologically important. Some of these computations occur in maps – arrays of neurons in which the tuning of neighboring neurons for a particular parameter value varies systematically. Computational maps transform the representation into a place-coded probability distribution that represents the computed values of parameters by sites of maximum relative activity. Numerous computational maps have been discovered, including visual maps of line orientation and direction of motion, auditory maps of amplitude spectrum and time interval, and motor maps of orienting movements.

The simplest example is the primary visual cortex, in which neighboring columns of neurons process information from neighboring small regions of visual space [21, 20]. In this case, the spatial organization of the neurons corresponds systematically to the spatial organization of the world, in the same way that the location of major cities on a map of Brazil corresponds to the actual location of those cities.

Such topographic neural models are useful for basic perception, but they are not rich enough to support high level kinds of reasoning such as my "taller than" example. How populations of neurons can support such reasoning is still unknown, as brain scanning technologies do not have sufficient resolution to pin down neural activity in enough detail to inspire theoretical models of how high-level mental modeling can work. But let me try to extrapolate from current views on neural representation, particularly those of Eliasmith and Anderson [10], to suggest how the brain might be able to make extra-topographic models of the world (see also [9]).

Neural populations can acquire the ability to encode features of the world as their firing activity becomes causally correlated with those features. (A and B are causally correlated if they are statistically correlated as the result of causal interactions between A and B). Neural populations are also capable

of encoding the activity of other neural populations, as the firing patterns of one population becomes causally correlated with the firing patterns of another population that feeds into it. If the input population is a topographic map, then the output population can become a more abstract representation of the features of the world, in two ways. The most basic retains some of the topographic structure of the input population, so that the output population is still a mental model of the world in that it shares some (but not all) relational structure with it. An even more abstract encoding is performed by an output neural population that captures key aspects of the encoding performed by the input population, but does so in a manner analogous to the way that language produces arbitrary, non-iconic representations. Just as there is no similarity between the word "cat" and cats, so the output neural population may have lost the similarity with the original stimulus: not all thinking uses mental models. Nevertheless, in some cases the output population provides sufficient information to enable decoding that generates an inference fairly directly, as in the "taller-than" example. The encodings of Adam, Bob, and Dan that include their heights makes it possible to just "see" that Adam is taller than Dan.

A further level of representation is required for consciousness, such as the experienced awareness that Adam is taller than Dan. Many philosophers and scientists have suggested that consciousness requires representation of representation (for references see [43]), but mental models seem to require several layers: representation of representation of representation. The conscious experience of an answer to a problem comes about because of activity in top-level neural populations that encode activity of medium-level modeling populations, that encode activity of low-level populations, that topographically represent features of the world. To put it another way, conscious models represent mid-level models that represent low-level topographic models that represent features of the world. The relation *representation* need not be transitive, but in this case it carries through, so that the conscious mental model represents the world and generates inferences about it.

So far, I have been focusing on mental models where the similar relation-structure is spatial, but temporal relations are just as important. When you imagine your national anthem sung by Michael Jackson, you are creating a mental model not only of the individual notes and tones but also of their sequence in time. Similarly, a mental model of a working device such as a windmill requires both visual/spatial representations of the blades and base of the windmill and also temporal representations of the blades of the mill. Not a lot of research has been done on how neurons can encode temporal relations, but I will explore two possibilities.

Elman [11] and other researchers have shown how simple recurrent networks can encode temporal relationships needed for understanding language. A recurrent network is one in which output neurons feed back to provide input to the input neurons, producing a kind of temporal cycle that can retain information. Much more complex neural structures, however, would be

needed to encode a song or running machine, perhaps something like the neural representation of a rule-based inference system being developed by Terry Stewart using Chris Eliasmith's neural engineering framework [38]. On this approach, a pattern of neural activation encodes a state of affairs that can be matched by a collection of rules capturing if-then relations. Then running a temporal pattern is a matter of firing off a sequence of rules, not by the usual verbal matching employed by rule-based cognitive architectures such as Anderson's [1] ACT, but by purely neural network operations. If the neural populations representing the states of affairs are mental models of either the direct, topographic kinds or the abstracted, structure-preserving kinds, then the running of the rule-based system would constitute a temporal *and* spatial mental model of the world.

6 Generating New Ideas

Peirce claimed that abduction could generate new ideas, but he did not specify how this could occur. If abduction is analyzed as a logical schema, then it is utterly mysterious how any new ideas could arise. The schema might be something like: "q is puzzling, p explains q, so maybe p." But this schema already includes the proposition p, so nothing new is generated. Hence logic-based approaches to abduction seem impotent to address what Peirce took to be a major feature of this kind of inference (cf. [45, 4]). Thagard [4] gave an account of how new concepts can be generated in the context of explanatory reasoning, but this account only applied to verbal concepts represented as frames with slots and values.

In contrast, the view of representations as patterns of activity in neural populations can be used to describe the generation of new multimodal concepts. Here I give only a quick sketch, as full details including mathematical analysis and computer simulations are provided in [46].

Assuming that two concepts are represented by patterns of activity in neural populations, which may be disjoint or overlapping, then a new combined concept can be represented by a new pattern of activity in a neural population which may also be disjoint or overlapping. A mathematical operation that combines patterns of neural activity is called convolution, which was originally a method for combining waves in signal processing theory. Tony Plate [31] adapted convolution to apply to vectors that stand for high-level symbolic representations, and Chris Eliasmith [8] developed a method for using biologically realistic neural networks to perform convolution. Thagard and Stewart [46] describe how many kinds of creativity and innovation, including scientific discovery, technological invention, social innovation, and artistic imagination, can be understood in terms of mechanisms of representation combination.

The convolution model of creative conceptual combination is fully multimodal, applying to whatever neural populations can represent, including information that is visual, auditory, olfactory, gustatory, tactile, kinesthetic,

or pain-related. Moreover, the Thagard and Stewart [46] account of creativity also applies to emotions, which can also be understood as patterns of activity in neural populations involving multiple brain areas involved in both cognitive appraisal and physiological perception [43]. In particular, the wonderful AHA! experience that attends creative breakthroughs can be understood as a neural process that involves a triple convolution:

1. Two representations are convolved to produce a novel one.
2. An emotional reaction to the new representation requires a convolution of cognitive appraisal and physiological perception.
3. The AHA or EUREKA reaction is a convolution of the new representation and the emotional reaction to it.

Thus the mechanism of convolution in neural networks is capable of modeling not only the combination of representations but also the emotional reaction that successful combinations generates.

Creative conceptual combination does not occur randomly, but rather in the directed context of attempts to solve problems, including ones that require generation of new explanations. Let us now consider how abductive inference can operate with neural populations.

7 Neural Abduction and Causality

Following ideas suggested in [41, 44] presented a neural network model of abductive reasoning based on the account of reasoning with conditionals developed by Eliasmith [8]. At one level, our neural model of abduction is very simple, using thousands of neurons to model a transition from q and p causes q to p. The advantage in taking a neural approach to modeling, as I have already described, is that p and q need not be linguistic representations, but can operate in any modality. To take a novel example, q could be a neural encoding of pain that I feel in my finger, and p could be neural encoding of a picture of splinter in my finger. Then my abductive inference goes from the experience of pain to the adoption of the visual representation that there is a splinter.

Moreover, the neural model of abduction tracks the relevant emotions. Initially, the puzzling q is associated with motivating emotions such as surprise and irritation. But as the hypothesis p is abductively adopted, the emotional reaction changes to relief and pleasure. Thus neural modeling can capture emotional aspect of abductive reasoning.

But how can we understand the causal relation in "p causes q"? Thagard and Litt ([44], see also [41]) argue that causality should not be construed formally as a deductive or probabilistic relation, but as a schema that derives from patterns of visual-motor experience. For example, when a baby discovers that moving its hand can move a rattle, it is forming an association that combines an initial visual state with a subsequent motor and tactile

state (pushing the rattle and feeling it) with a subsequent visual-auditory state (seeing the rattle move and make noise). I do not know whether such sensory-motor-sensory schemas are innate, having been acquired by natural selection in the form of neural connections that everyone is born with; alternatively they may be acquired very quickly by infants thanks to innate learning mechanisms. But on the basis of perceptual experiments in both adults and children, there is evidence that understanding of causality is tied to such multimodal representations (e.g. [25]). Moreover, the concept of force that figures centrally in many accounts of physical causality has its cognitive roots in body-based experiences of pushes and pulls. Hence it seems appropriate to speak of "embodied abduction", since both the causal relation itself and the multimodal representations of many hypotheses and facts to be explained are tied to sensory operations of the human body. However, the topic of embodiment is highly controversial, so I now discuss how I think the embodiment of abduction and mental models needs to be construed.

8 Embodiment: Moderate and Extreme

I emphatically reject the extreme embodiment thesis that thinking is just embodied action and therefore incompatible with computational-representational approaches to how brains work [6]. I argue below that even motor control requires a high degree of representation and computation. Much more plausible is the moderate embodiment thesis that language and thought are inextricably shaped by embodied action, a view that is maintained by Gibbs [14], Magnani [24] and others. On this view, thinking still requires representations and computations, but the particular nature of these depends in part on the kind of bodies that people have, including their sensory and motor capabilities. My remarks about multimodal representations and the sensory-motor-sensory schemas that underlie causal reasoning provide support for the moderate embodiment thesis.

However, there are two main reasons for not endorsing the extreme embodiment thesis. First, many kinds of thinking including causal reasoning, emotion, and scientific theorizing take us well beyond sensorimotor processes, so explaining our cognitive capacities requires recognizing representational/computational abilities that outstrip embodied action. Second, even the central case of embodied action – motor control – requires substantial representational/computational capabilities.

I owe to Lloyd Elliott the following summary of why motor control is much harder than you might think. Merely reaching to pick up a book requires solutions to many difficult problems for the brain to direct an arm and hand to reach out and pick up the book. First, the signals that pass between the brain and its sensors and muscles are very noisy. Information about the size, shape, and location of the book is transmitted to be brain via the eyes, but the process of translating retinal signals into judgments about the book

involved multiple stages of neural transformations [37, ch. 2]. Moreover, when the brain directs muscles to move the arm and hand in order to grasp the book, the signals sent involve noisy activity in millions of nerve cells.

Second, motor control is also made difficult by the fact that the context is constantly changing. You may need to pick up a book despite the fact that there are numerous changes taking place, not only in the orientation of your body, but also in visual information such as light intensity and the presence of other objects in the area. A person can pick up a book even though another person has reached across to pick up another book. Third, there are unavoidable time delays as the brain plans and attempts to move the arm to pick up the book.

Fourth, motor control is not an automatic process that occurs instantly to people, but usually requires large amounts of learning. It takes years for babies to become adept at handling physical objects, and even adults require months or years to become proficient at difficult motor tasks such as playing sports. Fifth, motor control is not a simple linear process of the brain just telling a muscle what to do, but requires non-linear integrations of the movements of multiple muscles and joints, which operate with many degrees of freedom. Picking up a book requires the coordination of all the muscles that move different parts of fingers, wrists, elbows, and shoulders.

Hence grasping and moving objects is a highly complex task that has been found highly challenging by people attempting to build robots. Fortunately for humans, millions of years of animal evolution have provided humans with the capacity to learn how to manipulate objects. Recent theoretical explanations of this capacity understand motor control as representational and computational, requiring mental models (see e.g. [4, 47]). What follows is a concise, simplified, synthesis of their accounts.

The brain is able to manipulate objects because its learning mechanisms, both supervised and unsupervised, enable it to build powerful internal models of connections among sensors, brain, and world. A brain needs a *forward model* from movements to sensory results, which enables it to predict what will be perceived as the result of particular movements. It also needs an *inverse model* from sensory results to movements, which enables it to predict what movement will produce the desired perceived result. Forward and inverse models are both dynamic mental models in the sense I discussed earlier: the relational structure they share with what they represent is both spatial and temporal, concerning the location and movement of limbs to produce changes in the world. Motor control in general requires a high-level control process in which the brain enables the body to interact productively with the world through a combination of representations of situations and goals, forward and inverse models, perceptual filters, and muscle control processes. The overall process is highly complex and not all like the kinds of manipulations of verbal symbols that some philosophers still take as the hallmark of representation and computation. But the brain's neural populations still stand for muscle movement and visual changes, with which their activity is

causally correlated, so it is legitimate to describe the activities of such populations as representational. Moreover, the mental modeling, both forward and inverse, is carried out by systematic changes in the neural populations, which hence qualifies as "principled manipulation of representations" [7, p. 29].

Let me summarize the argument in this section. Embodied action requires motor control. Motor control requires mental models, both forward and inverse, to identify dynamic relations among sensory information and muscle activity. Mental models are representational and computational. Hence embodied action requires representation and computation, so that it cannot provide an alternative to the representational/computational view of mind. Therefore considerations of multimodal representations, embodied abduction, and sensory-motor conceptions of causality only support the moderate embodiment thesis, and in fact require rejection of the extreme version.

Proponents of representation-free intelligence like to say that "the world is its own best model". As an advisory that a robot or other intelligent system should not need to represent everything to solve problems, this remark is useful; but literally it is clearly false. For imagining, planning, explaining, and many other important cognitive activities, the world is a very inadequate model of itself: far too complex and limited in its manipulability. In contrast, mental models operating at various degrees of abstraction are invaluable for high-level reasoning. The world might be its own best model if you're a cockroach, with very limited modeling abilities. But if you have the representational power of a human or powerful robot, then you can build simplified but immensely useful models of past and future events, as well as of events that your senses do not enable you observe. Hence science uses abductive inference and conceptual combination to generate representations of theoretical (i.e. non-observable) entities such as electrons, viruses, genes, and mental representations.

Cockroaches and many other animals are as embodied, embedded, and situated in the world as human beings, but they are far less effective than people at building science, technology, and other cultural developments. One of the many advantages that people gain from our much larger brains is the ability to work with mental models, including ones used for abduction and the generation of new ideas.

9 Conclusion

I finish with a reassessment of Peirce's ideas about abduction from the neural perspective that I have been developing. Peirce did most of his work on inference in the nineteenth century, well before the emergence of ideas about computation and neural processes. He was a scientist as well as a philosopher of science, and undoubtedly would have revised his views on the nature of inference in line with subsequent scientific developments.

On the positive side, Peirce was undoubtedly right about the importance of abduction as a kind of inference. The evaluative aspect of abduction is recognized in philosophy of science under the headings of inference to the best explanation and explanatory coherence, and the creative aspect is recognized in philosophy and artificial intelligence through work on how hypotheses are formed. Second, Peirce was prescient in noticing the emotional instigation of abduction as the result of surprise, although I do not know if he also noticed that achieving abduction generates the emotional response of relief. Third, Peirce was right in suggesting that the creation of new ideas often occurs in the context of abductive inference, even if abduction itself is not the generating process.

On the other hand, there are several suggestions that Peirce made about abduction that do not fit well with current psychological and neural understanding of abduction. I do not think that emotion is well described as a kind of abduction, as it involves an extremely complex process that combines cognitive appraisal of a situation with respect to ones goals and perception of bodily states [43, 42]. At best, abductive inference is only a part of the broader parallel process of emotional reactions. Similarly, perception is not a kind of abduction, as it involves many more basic neuropsychological processes that are not well described as generation and evaluation of explanatory hypotheses (see e.g. [37, ch. 2]).

Finally, Peirce's suggestion that abduction requires a special instinct for guessing right is not well supported by current neuropsychological findings. Perhaps evolutionary psychologists would want to propose that there is an innate module for generating good hypotheses, but there is a dearth of evidence that would support this proposal. Rather, I prefer the suggestion of Quartz and Sejnowski [32, 33] that what the brain is adapted for is adaptability, through powerful learning mechanisms that humans can apply in many contexts. One of these learning mechanisms is abductive inference, which leads people to respond to surprising observations with a search for hypotheses that can explain them. Like all cognitive processes, this search must be constrained by contextual factors such as triggering conditions that cut down the number of new conceptual combinations that are performed [4]. Abduction and concept formation occur as part of the operations of a more general cognitive architecture.

I see no reason to claim that the constraints on these operations include preferences for particular kinds of hypotheses, which is how I interpret Peirce's instinct suggestion. Indeed, scientific abduction has led to the generation of many hypotheses that scientists now think are wrong (e.g. humoral causes of disease, phlogiston, caloric, and the luminiferous ether) and to many hypotheses that go against popular inclinations (e.g. Newton's force at a distance, Darwin's evolution by natural selection, Einstein's relativistic account of space-time, quantum mechanics, and the mind-brain identity theory). Although it is reasonable to suggest that the battery of innate human learning mechanisms includes ones for generating hypotheses to explain

surprising events, there is no support for Peirce's contention that people must have an instinct for guessing *right*. Evolutionary psychologists like to compare the brain to a Swiss army knife that has many specific built-in capacities; but a more fertile comparison is the human hand, which evolved to be capable of many different operations from grasping to signaling to thumb typing on smartphones. Peirce's view of abduction as requiring innate insight is thus as unsupported by current research as the view of Fodor [12] that cognitive science cannot possibly explain abduction: many effective techniques have been developed by philosophers and AI researchers to explain complex causal reasoning.

I have tried to show in this paper how Peirce's abduction is, from a neural perspective, highly consonant with psychological theories of mental models, which can also productively be construed as neural processes. Brains make mental models through complex patterns of neural firing and use them in many kinds of inference, from planning actions to the most creative kinds of abductive reasoning. I have endorsed a moderate thesis about the importance of embodiment for the kinds of representations that go into mental modeling, but critiqued the extreme view that sees embodiment as antithetical to mental models and other theories of representation. Further developments of neural theories of mental models should further clarify their roles in many important psychological phenomena.

Acknowledgements. For valuable ideas and useful discussions, I am grateful to Chris Eliasmith, Terry Stewart, Lloyd Elliott, and Phil Johnson-Laird. This research has been funded by the Natural Sciences and Engineering Research Council of Canada.

References

1. Anderson, J.R.: How Can the Mind Occur in the Physical Universe? Oxford University Press, Oxford (2007)
2. Churchland, P.S., Sejnowski, T.: The Computational Brain. MIT Press, Cambridge (1992)
3. Craik, K.: The Nature of Explanation. Cambridge University Press, Cambridge (1943)
4. Davidson, P.R., Wolpert, D.M.: Widespread access to predictive models in the motor system: A short review. Journal of Neural Engineering 2, S313–S319 (2005)
5. Dayan, P., Abbott, L.F.: Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. MIT Press, Cambridge (2001)
6. Dreyfus, H.L.: Why Heideggerian AI failed and how fixing it would require making it more Heidegerian. Philosophical Psychology 20, 247–268 (2007)
7. Edelman, S.: Computing the Mind: How the Mind Really Works. Oxford University Press, Oxford (2008)

8. Eliasmith, C.: Cognition with neurons: A large-scale, biologically realistic model of the Wason task. In: Bara, B., Barasalou, L., Bucciarelli, M. (eds.) Proceedings of the XXVII Annual Conference of the Cognitive Science Society, pp. 624–629. Lawrence Erlbaum Associates, Mahwah (2005)
9. Eliasmith, C.: Neurosemantics and categories. In: Cohen, H., Lefebvre, C. (eds.) Handbook of Categorization in Cognitive Science, pp. 1035–1054. Elsevier, Amsterdam (2005)
10. Eliasmith, C., Anderson, C.H.: Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems. MIT Press, Cambridge (2003)
11. Elman, J.L.: Finding structure in time. Cognitive Science 14, 179–211 (1990)
12. Fodor, J.: The Mind Doesn't Work that Way. MIT Press, Cambridge (2000)
13. Gentner, D., Stevens, A.L. (eds.): Mental Models. Lawrence Erlbaum, Hillsdale (1983)
14. Gibbs, R.W.: Embodiment and Cognitive Science. Cambridge University Press, Cambridge (2006)
15. Holland, J.H., Holyoak, K.J., Nisbett, R.E., Thagard, P.R.: Induction: Processes of Inference, Learning, and Discovery. MIT Press/Bradford Books, Cambridge (1986)
16. Johnson-Laird, P.N.: Mental Models. Harvard University Press, Cambridge (1983)
17. Johnson-Laird, P.N.: The history of mental models. In: Manktelow, K., Chung, M.C. (eds.) Psychology of Reasoning: Theoretical and Historical Perspectives, pp. 179–212. Psychology Press, New York (2004)
18. Johnson-Laird, P.N.: How We Reason. Oxford University Press, Oxford (2006)
19. Johnson-Laird, P.N., Byrne, R.M.: Deduction. Lawrence Erlbaum Associates, Hillsdale (1991)
20. Kaas, J.H.: Topographic maps are fundamental to sensory processing. Brain research bulletin 44, 107–112 (1997)
21. Knudsen, E.I., du Lac, S., Esterly, S.D.: Computational maps in the brain. Annual Review of Neuroscience 10, 41–65 (1987)
22. Magnani, L.: Model-based creative abduction. In: Magnani, L., Nersessian, N.J., Thagard, P. (eds.) Model-Based Reasoning in Scientific Discovery, pp. 219–238. Kluwer/Plenum, New York (1999)
23. Magnani, L.: Abduction, Reason, and Science: Processes of Discovery and Explanation. Kluwer/Plenum, New York (2001)
24. Magnani, L.: Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning. Springer, Berlin (2009)
25. Michotte, A.: The Perception of Causality. Miles, T.R., Miles, E. (trans.), Methuen, London (1963)
26. Nersessian, N.J.: Creating Scientific Concepts. MIT Press, Cambridge (2008)
27. O'Reilly, R.C., Munakata, Y.: Computational Explorations in Cognitive Neuroscience. MIT Press, Cambridge (2000)
28. Parisien, C., Thagard, P.: Robosemantics: How Stanley the Volkswagen represents the world. Minds and Machines 18, 169–178 (2008)
29. Peirce, C.S.: Collected Papers. Harvard University Press, Cambridge (1931-1958)
30. Peirce, C.S.: Reasoning and the Logic of Things. Harvard University Press, Cambridge (1992)
31. Plate, T.: Holographic Reduced Representations. CSLI, Stanford (2003)

32. Quartz, S.R., Sejnowski, T.J.: The neural basis of cognitive development: A constructivist manifesto. Behavioral and Brain Sciences 20, 537–556 (1997)
33. Quartz, S.R., Sejnowski, T.J.: Liars, Lovers, and Heroes: What the New Brain Science Reveals about How We Become Who We Are. Morrow, William, New York (2002)
34. Rips, L.J.: Mental muddles. In: Brand, M., Harnish, R.M. (eds.) The representation of knowledge and belief, pp. 258–286. University of Arizona Press, Tucson (1986)
35. Salmon, W.C.: Four decades of scientific explanation. In: Kitcher, P., Salmon, W.C. (eds.) Scientific Explanation. Minnesota Studies in the Philosophy of Science, vol. XIII, pp. 3–219. University of Minnesota Press, Minneapolis (1989)
36. Shelley, C.P.: Visual abductive reasoning in archaeology. Philosophy of Science 63, 278–301 (1996)
37. Smith, E.E., Kosslyn, S.M.: Cognitive Psychology: Mind and Brain. Pearson Prentice Hall, Upper Saddle River (2007)
38. Stewart, T.C., Eliasmith, C.: Spiking neurons and central executive control: The origin of the 50-millisecond cognitive cycle. In: Howes, A., Peebles, D., Cooper, R. (eds.) 9th International Conference on Cognitive Modeling ICCM 2009, Manchester, UK (2009)
39. Tauber, M.J., Ackerman, D. (eds.): Mental Models and Human-Computer Interaction, vol. 2. North-Holland, Amsterdam (1990)
40. Thagard, P.: Computational Philosophy of Science. MIT Press, Cambridge (1988)
41. Thagard, P.: Abductive inference: From philosophical analysis to neural mechanisms. In: Feeney, A., Heit, E. (eds.) Inductive Reasoning: Experimental, Developmental, and Computational Approaches, pp. 226–247. Cambridge University Press, Cambridge (2007)
42. Thagard, P.: The Brain and the Meaning of Life. Princeton University Press, Princeton (2010)
43. Thagard, P., Aubie, B.: Emotional consciousness: A neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. Consciousness and Cognition 17, 811–834 (2008)
44. Thagard, P., Litt, A.: Models of scientific explanation. In: Sun, R. (ed.) The Cambridge Handbook of Computational sychology, pp. 549–564. Cambridge University Press, Cambridge (2008)
45. Thagard, P., Shelley, C.P.: Abductive reasoning: Logic, visual thinking, and coherence. In: Dalla Chiara, M.L., Doets, K., Mundici, D., van Benthem, J. (eds.) Logic and Scientific Methods, pp. 413–427. Kluwer, Dordrecht (1997)
46. Thagard, P., Stewart, T.C.: The Aha! experience: Creativity through emergent binding in neural networks. under review (forthcoming)
47. Wolpert, D.M., Ghahramani, Z.: Computational principles of movement neuroscience. Nature Neuroscience 3, 1212–1217 (2000)