Models of Scientific Explanation

Paul Thagard and Abninder Litt

Thagard, P., & Litt, A. (forthcoming). Models of scientific explanation. In R. Sun (Ed.), *The Cambridge handbook of computational cognitive modeling*. Cambridge: Cambridge University Press.

EXPLANATION

Explanation of why things happen is one of humans' most important cognitive operations. In everyday life, people are continually generating explanations of why other people behave the way they do, why they get sick, why computers or cars are not working properly, and of many other puzzling occurrences. More systematically, scientists develop theories to provide general explanations of physical phenomena such as why objects fall to earth, chemical phenomena such as why elements combine, biological phenomena such as why species evolve, medical phenomena such as why organisms develop diseases, and psychological phenomena such as why people sometimes make mental errors.

This chapter reviews computational models of the cognitive processes that underlie these kinds of explanations of *why* events happen. It is not concerned with another sense of explanation that just means clarification, as when someone explains the U. S. constitution. The focus will be on scientific explanations, but more mundane examples will occasionally be used, on the grounds that the cognitive processes for explaining why events happen are much the same in everyday life and in science, although scientific explanations tend tobe more systematic and rigorous than everyday ones. In addition to providing a concise review of previous computational models of explanation, this chapter describes a new neural network model that shows how explanations can be performed by multimodal distributed representations.

Before proceeding with accounts of particular computational models of explanation, let us characterize more generally the three major processes involved in explanation and the four major theoretical approaches that have been taken in computational models of it. The three major processes are: providing an explanation from available information, generating new hypotheses that provide explanations, and evaluating competing explanations. The four major theoretical approaches are: deductive, using logic or rule-based systems; schematic, using explanation patterns or analogies; probabilistic, using Bayesian networks; and neural, using networks of artificial neurons. For each of these theoretical approaches, it is possible to characterize the different ways in which the provision, generation, and evaluation of explanations are understood computationally.

The processes of providing, generating and evaluating explanations can be illustrated with a simple medical example. Suppose you arrive at your doctor's office with a high fever, headache, extreme fatigue, a bad cough, and major muscle aches. Your doctor will probably tell you that you have been infected by the influenza virus, with an explanation like:

> People infected by the flu virus often have the symptoms you describe.
>
> You have been exposed to and infected by the flu virus.
>
> So, you have these symptoms.

If influenza is widespread in your community and your doctor has been seeing many patients with similar symptoms, it will not require much reasoning to provide this explanation by stating the flu virus as the likely cause of your symptoms.

Sometimes, however, a larger inferential leap is required to provide an explanation. If your symptoms also include a stiff neck and confusion, your doctor may make the less common and more serious diagnosis of meningitis. This diagnosis requires generating the hypothesis that you have been exposed to bacteria or viruses that have infected the lining surrounding the brain. In this case, the doctor is not simply applying knowledge already available to provide an explanation, but generating a hypothesis about you that makes it possible to provide an explanation. This hypothesis presupposes a history of medical research that led to the identification of meningitis as a disease caused by particular kinds of bacteria and viruses, research that required the generation of new general hypotheses that made explanation of particular cases of the disease possible.

In addition to providing and generating explanations, scientists and ordinary people sometimes need to evaluate competing explanations. If your symptoms are ambiguous, your doctor may be unsure whether you have influenza or meningitis, and therefore consider them as competing explanations of your symptoms. The doctor's task is then to figure out which hypothesis, that you have influenza or meningitis, is the *best* explanation of your disease. Similarly, at a more general level, scientific researchers had to consider alternative explanations of the causes of meningitis and select the best one. This selection presupposed the generation and provision of candidate explanations and involved the additional cognitive processes of comparing the candidates in order to decide which was most plausible.

Provision, generation, and evaluation of explanations can all be modeled computationally, but the forms these models take depends on background theories about what constitutes an explanation. One view, prominent in both philosophy of science and artificial intelligence, is that explanations are deductive arguments. An explanation consists of a deduction in which the explanatory target, to be explained, follows logically from the explaining set of propositions. Here is a simple example:

Anyone with influenza has fever, aches, and cough.

You have influenza.

So, you have fever, aches, and cough.

In this oversimplified case, it is plausible that the explanatory target follows deductively from the explaining propositions.

Often, however, the relation between explainers and explanatory targets is looser than logical deduction, and an explanation can be characterized as a causal schema rather than a deductive argument. A schema is a conceptual pattern that specifies a typical situation, as in the following example:

Explanatory pattern: Typically, influenza causes fever, aches, and cough.

Explanatory target: You have fever, aches, and cough.

Schema instantiation: Maybe you have influenza.

In medical research, the explanatory pattern is much more complex, as scientists can provide a much richer description of the genetic, biological, and immunological causes of infection. Like deductive explanations, schematic ones can be viewed as providing causes, but with a more flexible relation between explainers and what is explained.

Probability theory can also be used to provide a less rigid conception of explanation than logical deducibility. A target can be explained by specifying that it is probable given the state of affairs described by the explainers. In the flu case, the explanation has this kind of structure:

The probability of having fever, aches, and coughs given influenza is high.

So influenza explains why you have fever, aches, and cough.

On this view, explanation is a matter of conditional probability rather than logical deducibility or schematic fit. Like deduction and schema views, the probabilistic view of explanation has inspired interesting computational models, particularly ones involving Bayesian networks that will be described below.

A fourth computational way of modeling explanation derives from artificial neural networks which attempt to approximate how brains use large groups of neurons, operating in parallel to accomplish complex cognitive tasks. The neural approach to explanation is not in itself a theory of explanation in the way that the deductive, schema, and probabilistic views are, but it offers new ways of thinking about the nature of the provision, generation, and evaluation of explanations. This quick overview sets the stage for the more detailed analysis of computational models of scientific explanation that follows. For a concise review of philosophical theories of explanation, see Woodward (2003); for more detail, see Kitcher and Salmon (1989).

DEDUCTIVE MODELS

The view that explanations are deductive arguments has been prominent in the philosophy of science. According to Hempel (1965, p. 336) an explanation is an argument of the form:

$C_1, C_2, \dots C_k$

$L_1, L_2, \dots, L_r$

E

Here Cs are sentences describing particular facts, the Ls are general laws, and E is the sentence explained by virtue of being a logical consequence of the other sentences. This sort of explanation does occur in some areas of science such as physics, where laws stated as mathematical formulas enable deductive predictions.

Many computational models in artificial intelligence have presupposed that explanation is deductive, including ones found in logic programming, truth maintenance systems, explanation-based learning, qualitative reasoning, and in some approaches to abduction (a form of inference that involves the generation and evaluation of explanatory hypotheses). See, for example, Russell and Norvig (2003), Bylander et al. (1991), and Konolige (1992). These AI approaches are not intended as models of human cognition, but see the chapter by Bringsjord in this volume for discussion of use of formal logic in cognitive modeling.

Deductive explanation also operates in rule-based models which have been proposed for many kinds of human thinking (Anderson, 1983, 1993; Holland, Holyoak, Nisbett, and Thagard, 1986; Newell and Simon, 1972; Newell, 1990; see also the chapter by Taatgen and Anderson in this volume). A rule-based system is a set of rules with an IF part consisting of conditions (antecedents) and a THEN part consisting of actions (consequents). Rule-based systems have often been used to model human problem solving in which people need to figure out how to get from a starting state to a goal state by applying a series of rules. This is a kind of deduction, in that the application of rules in a series of if-then inferences amounts to a series of applications of the rule of deductive inference, modus ponens, which licenses inferences from *p* and *if p then q* to *q*. Most rule-based systems, however, do not always proceed just from starting states to goal states, but can also work backward from a goal state to find a series of rules that can be used to get from the starting state to the goal state.

Explanation can be understood as a special kind of problem solving, in which the goal state is a target to be explained. Rule-based systems do not have the full logical complexity to express the laws required for Hempel's model of explanation, but they can perform a useful approximation. For instance, the medical example used in the introduction can be expressed by a rule like:

IF *X* has influenza, THEN *X* has fever, cough, and aches.

Paul has influenza.

Paul has fever, cough, and aches.

Modus ponens provides the connection between the rule and what is to be explained. In more complex cases, the connection would come from a sequence of applications of modus ponens as multiple rules get applied. In contrast to Hempel's account in which an explanation is a static argument, rule-based explanation is usually a dynamic process involving application of multiple rules. For a concrete example of a running program that accomplishes explanations in this way, see the PI cognitive model of Thagard (1988; code is available at http://cogsci.uwaterloo.ca/). The main scientific example to which PI has been applied is the discovery of the wave theory of sound, which occurs in the context of an attempt to explain why sounds propagate and reflect.

Thus rule-based systems can model the provisions of explanations construed deductively, but what about the generation and evaluation of explanations? A simple form of abductive inference that generates hypotheses can be modeled as a kind of backward chaining. Forward chaining means running rules forward in the deductive process that proceeds from the starting state toward a goal to be solved. Backward chaining occurs when a system works backward from a goal state to find rules that could produce it from the starting state. Human problem solving on tasks such as solving mathematics problems often involves a combination of forward and backward reasoning, in which a problem solver looks both at the how the problem is described and the answer that is required, attempting to make them meet. At the level of a single rule, backward chaining has the form: goal *G* is to be accomplished; there is the rule IF *A* THEN *G*, i.e. action A would accomplish *G*; so set *A* as a new subgoal to be accomplished. Analogously, people can backchain to find a possible explanation: fact *F* is to be explained; there is a rule IF *H* THEN *F*, i.e. hypothesis *H* would explain *F*; so hypothesize that *H* is true. Thus if you know that Paul has fever, aches, and a cough, and the rule that IF *X* has influenza, THEN *X* has fever, cough, and aches, then you can run the rule backward to produce the hypothesis that Paul has influenza.

The computational model PI performs this simple kind of hypothesis generation, but it also can generate other kinds of hypotheses (Thagard, 1988). For example, from the observation that the orbit of Uranus is perturbed, and the rule that IF a planet has another planet near it THEN its orbit is perturbed, PI infers that there is some planet near Uranus; this is called existential abduction. PI also performs abduction to rules that constitute the wave theory of sound: the attempt to explain why an arbitrary sound propagates generates not only the hypothesis that it consists of a wave but the general theory that all sounds are waves. PI also performs a kind of analogical abduction, a topic discussed in the next section on schemas.

Abductive inference that generates explanatory hypotheses is an inherently risky form of reasoning because of the possibility of alternative explanations. Inferring that Paul has influenza because it explains his fever, aches, and cough is risky because other diseases such as meningitis can cause the same symptoms. People should only accept an explanatory hypothesis if it is better than its competitors, a form of inference that philosophers call *inference to the best explanation* (Harman, 1973; Lipton, 2004). The PI cognitive model performs this kind of inference by taking into account 3 criteria for the best explanation: consilience, which is a measure of how much a hypothesis explains; simplicity, which is a measure of how few additional assumptions a hypothesis needs to carry out an explanation; and analogy, which favors hypotheses whose explanations are analogous to accepted ones. A more psychologically elegant way of performing inference to the best explanation, the model ECHO, is described below in the section on neural networks. Neither the PI nor the ECHO way of evaluating competing explanations requires that explanations be deductive.

In artificial intelligence, the term "abduction" is often used to describe inference to the best explanation as well as the generation of hypotheses. In actual systems, these two processes can be continuous, for example in the PEIRCE tool for abductive inference described by Josephson and Josephson (1994 , p. 95). This is primarily an engineering tool rather than a cognitive model, but is mentioned here as another approach to generating and evaluating scientific explanations, in particular medical ones involving interpretation of blood tests. The PEIRCE system accomplishes the goal of generating the best explanatory hypothesis by achieving three subgoals:

1. generation of a set of plausible hypotheses
2. construction of a compound explanation for all the findings
3. criticism and improvement of the compound explanation.

PEIRCE employs computationally effective algorithms for each of these subgoals, but does not attempt to do so in a way that corresponds to how people accomplish them.

SCHEMA AND ANALOGY MODELS

In ordinary life and in many areas of science less mathematical than physics, the relation between what is explained and what does the explaining is usually looser than deduction. An alternative conception of this relation is provided by understanding an explanation as the application of a causal schema, which is a pattern that describes the relation between causes and effects. For example, cognitive science uses a general explanation schema that has the following structure (Thagard, 2005):

Explanation target: Why do people have a particular kind of **intelligent behavior**?

Explanatory pattern:

People have mental **representations.**

People have algorithmic **processes** that operate on those **representations.**

The **processes**, applied to the **representations**, produce the **behavior**.

This schema provides explanations when the terms shown in boldface are filled in with specifics, and subsumes schemas that describe particular kinds of mental representations such as concepts, rules, and neural networks. Philosophers of science have discussed the importance of explanation schemas or patterns (Kitcher, 1993; Thagard, 1999).

A computational cognitive model of explanation schemas was developed in the SWALE project (Schank, 1986; Leake, 1992). This project modeled people's attempts to explain the unexpected 1984 death of a racehorse, Swale. Given an occurrence, the program SWALE attempts to fit it into memory. If a problem arises indicating an anomaly, then the program attempts to find an explanation pattern stored in memory. The explanation patterns are derived from previous cases, such as other unexpected deaths. If SWALE finds more than one relevant explanation pattern, it evaluates them to determine which is most relevant to the intellectual goals of the person seeking understanding. If the best explanation pattern does not quite fit the case to be explained, it can be tweaked (adapted) to provide a better fit, and the tweaked version is stored in memory for future use. The explanation patterns in SWALE's data base included both general schemas such as *exertion + heart defect causes fatal heart attack* and particular examples, which are used for case-based reasoning, a kind of analogical thinking. Leake (1992) describes how competing explanation patterns can be evaluated according to various criteria, including a reasoner's pragmatic goals.

Explaining something by applying a general schema involves the same processes as explaining using analogies. In both cases, reasoning proceeds as follows:

Identify the case to be explained.

Search memory for a matching schema or case.

Adapt the found schema or case to provide an explanation of the case to be explained.

In deductive explanation, there is a tight logical relation between what is explained and the sentences that imply it, but in schematic or analogical explanation there need only be a roughly specified causal relation.

Falkenhainer (1990) describes a program PHINEAS that provides analogical explanations of scientific phenomena. The program uses Forbus' (1984) qualitative process theory to represent and reason about physical change, and is provided with knowledge about liquid flow. When presented with other phenomena to be explained such as osmosis and heat flow, it can generate new explanations analogically by computing similarities in relational structure, using the Structure Mapping Engine (Falkenhainer , Forbus, and Gentner, 1989). PHINEAS operates in four stages: access, mapping/transfer, qualitative simulation, and revision. For example, it can generate an explanation of the behavior of a hot brick in cold water by analogy to what happens when liquid flows between two containers. Another computational model that generates analogical explanations is the PI system (Thagard, 1988), which simulates the discovery of the wave theory of sound by analogy to water waves.

Thus computational models of explanation that rely on matching schematic or analogical structures based on causal fit provide an alternative to models of deductive explanation. These two approaches are not competing theories of explanation, because explanation can take different forms in different areas of science. In areas such as physics that are rich in mathematically expressed knowledge, deductive explanations may be available. But in more qualitative areas of science and everyday life, explanations are usually less exact and may be better modeled by application of causal schemas or as a kind of analogical inference.

PROBABILISTIC MODELS

Another, more quantitative way of establishing a looser relation than deduction between explainers and their targets is to use probability theory. Salmon (1970) proposed that the key to explanation is *statistical relevance,* where a property *B* in a population *A* is relevant to a property *C* if the probability of *B* given *A* and *C* is different from the probability of *B* given *A* alone: $P(B|A\&C) \neq P(B|A)$. Salmon later moved away from a statistical understanding of explanation toward a causal mechanism account (Salmon, 1984), but other philosophers and artificial intelligence researchers have focused on probabilistic accounts of causality and explanation. The core idea here is that people explain why something happened by citing the factors that made it more probable than it would have been otherwise.

The main computational method for modeling explanation probabilistically is Bayesian networks, developed by Pearl (1988, 2000) and other researchers in philosophy and computer science (e.g Spirtes, Glymour, and Scheines, 1993; Neapolitain, 1990; Glymour, 2001; see also the chapter by Tenenbaum and Griffith in this volume). A Bayesian network is a directed graph in which the nodes are statistical variables, the edges between them represent conditional probabilities, and no cycles are allowed: you cannot have *A* influencing *B* which influences *A*. Causal structure and probability are connected by the Markov assumption, which says that a variable A in a causal graph is independent of all other variables that are not its effects, conditional on its direct causes in the graph (Glymour, 2003).

Bayesian networks are convenient ways for representing causal relationships, as in figure 1. Powerful algorithms have been developed for making probabilistic inferences in Bayesian networks and for learning causal relationships in these networks. Applications have included scientific examples, such as developing models in the social sciences (Spirtes, Glymour, and Scheines, 1993). Bayesian networks provide an excellent tool for computational and normative philosophical applications, but the relevant question for this chapter is how they might contribute to cognitive modeling of scientific explanation.

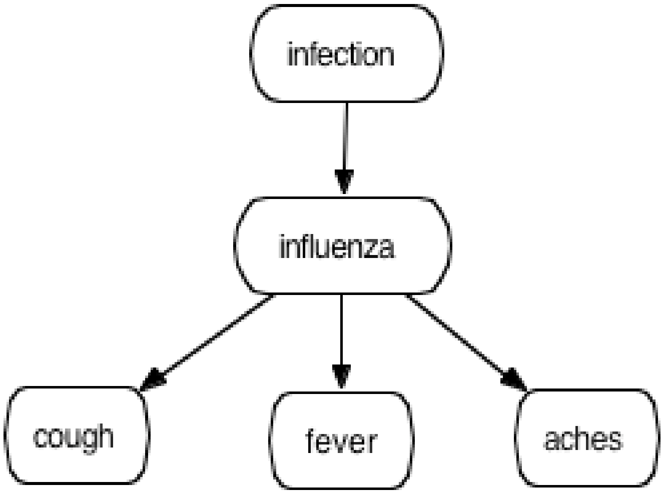

Figure 1. Causal map of a disease.   In a Bayesian network, each node is a variable and the arrow indicates causality represented by conditional probability.

The psychological plausibility of Bayesian networks has been advocated by Glymour (2001) and Gopnik et al. (2004).   They show the potential for using Bayesian networks to explain a variety of kinds of reasoning and learning studied by cognitive and developmental psychologists.   Gopnik et al. (2004) argue that children's causal learning and inference may involve computations similar to those for learning Bayesian networks and for predicting with them.   If they are right about children, it would be plausible that the causal inferences of scientists are also well modeled by Bayesian networks.   From this perspective explaining something consists of instantiating it in a causal network and using probabilistic inference to indicate how it depends causally on other factors. Generating an explanation consists of producing a Bayesian network, and evaluating competing explanations consists of calculating the comparative probability of different causes.

Despite their computational and philosophical power, there are reasons to doubt the psychological relevance of Bayesian networks. Although it is plausible that people's mental representations contain something like rough causal maps depicted in figure 1, it is much less plausible that these maps have all the properties of Bayesian networks. First, there is abundant experimental evidence that reasoning with probabilities is not a natural part of people's inferential practice (Kahneman, Slovic, Tversky, 1982; Gilovich, Griffin, and Kahneman, 2002). Computing with Bayesian networks requires a very large number of conditional probabilities that people not working in statistics have had no chance to acquire. Second, there is no reason to believe that people have the sort of information about independence that is required to satisfy the Markov condition and to make inference in Bayesian networks computationally tractable. Third, although it is natural to represent causal knowledge as directed graphs, there are many scientific and everyday contexts in which such graphs should have cycles because of feedback loops. For example, marriage breakdown often occurs because of escalating negative affect, in which the negative emotions of one partner produces behaviors that increase negative emotions of the other which then produces behavior that increases the negative emotions of the first partner (Gottman et al., 2003). Such feedback loops are also common in biochemical pathways needed to explain disease (Thagard, 2003). Fourth, probability by itself is not adequate to capture people's understanding of causality, as argued in the last section of this chapter. Hence it is not at all obvious that Bayesian networks are the best way to model explanation by human scientists. Even in statistically rich fields such as the social sciences, scientists rely on an intuitive, non-probabilistic sense of causality of the sort discussed below.

NEURAL NETWORK MODELS

The most important approach to cognitive modeling not yet discussed employs artificial neural networks. Applying this approach to high-level reasoning faces many challenges, particularly in representing the complex kinds of information contained in scientific hypotheses and causal relations. Thagard (1989) provided a neural network model of how competing scientific explanations can be evaluated, but did so using a localist network in which entire propositions were represented by single artificial neurons and in which relations between propositions are represented by excitatory and inhibitory links between the neurons. Although this model provides an extensive account of explanation evaluation that is reviewed below, it reveals nothing about what an explanation is or how explanations are generated. Neural network modelers have been concerned mostly with applications to low-level psychological phenomena such as perception, categorization, and memory, rather than high-level ones such as problem solving and inference (O'Reilly and Munakata, 2000). However, this section shows how a neurologically complex model of explanation and abductive inference can be constructed. For a review of neural network approaches to cognitive modeling, see the chapter by McClelland in this volume.

One benefit of attempting neural analyses of explanation is that it becomes possible to incorporate multimodal aspects of cognitive processing that tend to be ignored from deductive, schematic, and probabilistic perspectives. Thagard (forthcoming) describes how both explainers and explanation targets are sometimes represented non-verbally. In medicine, for example, doctors and researchers may employ visual hypotheses (say about the shape and location of a tumor) to explain observations that can be represented using sight, touch, and smell as well as words. Moreover, the process of abductive inference has emotional inputs and outputs, because it is usually initiated when an observation is found to be surprising or puzzling, and it often results in a sense of pleasure or satisfaction when a satisfactory hypothesis is used to generate an explanation. Figure 2 provides an outline of this process. Let us now look at an implementation of a neural network model of this sketch.

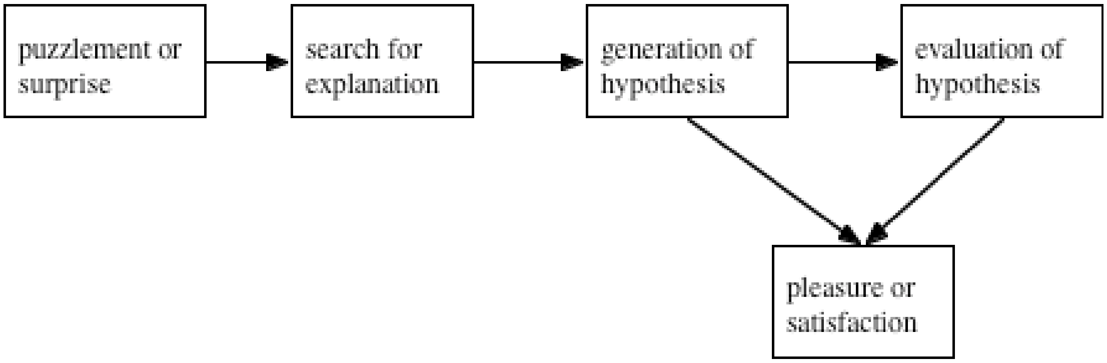

Figure 2. The process of abductive inference. From Thagard (forthcoming).

The model of abduction described here follows the Neural Engineering Framework (NEF) outlined in Eliasmith & Anderson (2003), and is implemented using the MATLAB-based NEF simulation software *NESim*. The NEF proposes three basic principles of neural computation (Eliasmith and Anderson, p. 15):

1. Neural representations are defined by a combination of nonlinear encoding and linear decoding.

2. Transformations of neural representations are linearly decoded functions of variables that are represented by a neural population.

3. Neural dynamics are characterized by considering neural representations as control theoretic state variables.

These principles are applied to a particular neural system by identifying the interconnectivity of its subsystems, neuron response functions, neuron tuning curves, subsystem functional relations, and overall system behavior. For cognitive modeling, the NEF is useful because it provides a mathematically rigorous way of building more realistic neural models of cognitive functions.

The NEF characterizes neural populations and activities in terms of mathematical representations and transformations. The complexity of a representation is constrained by the *dimensionality* of the neural population that represents it. In rough terms, a single dimension in such a representation can correspond to one discrete "aspect" of that representation (e.g., speed and direction are the dimensional components of the vector quantity velocity). A hierarchy of representational complexity thus follows from neural activity defined in terms of one-dimensional scalars; vectors, with a finite but arbitrarily large number of dimensions; or functions, which are essentially *continuous* indexings of vector elements, thus ranging over infinite dimensional spaces.

The Neural Engineering Framework provides for arbitrary computations to be performed in biologically realistic neural populations, and has been successfully applied to phenomena as diverse as lamprey locomotion (Eliasmith and Anderson, 2003), path integration by rats (Conklin & Eliasmith, 2005), and the Wason card selection task (Eliasmith, 2005). The Wason task model, in particular, is structured very similarly to the model of abductive inference discussed here. Both employ *holographic reduced representations*, a high-dimensional form of distributed representation.

First developed by Plate (1993, 1994, 2003), holographic reduced representations (HRRs) combine the neurological plausibility of distributed representations with the ability to maintain complex, embedded structural relations in a computationally efficient manner. This ability is common in symbolic models and is often singled out as deficient in distributed connectionist frameworks; for a comprehensive review of HRRs in the context of the distributed vs. symbolic representation debate, see Eliasmith & Thagard (2001). HRRs consist of high-dimensional vectors combined via multiplicative operations, and are similar to the tensor products used by Smolensky (1990) as the basis for a connectionist model of cognition. But HRRs have the important advantage of *fixed dimensionality*: the combination of two n-dimensional HRRs produces another n-dimensional HRR, rather than the 2n or even n^2 dimensionality one would obtain using tensor products. This avoids the explosive computational resource requirements of tensor products to represent arbitrary, complex structural relationships.

HRR representations are constructed through the multiplicative *circular convolution* (denoted by $\otimes$) and are decoded by the approximate inverse operation, *circular correlation* (denoted by #). The details of these operations are given in the appendices of Eliasmith & Thagard (2001), but in general if $C = A \otimes B$ is encoded, then $C\#A \approx B$ and $C\#B \approx A$. The approximate nature of the unbinding process introduces a degree of noise, proportional to the complexity of the HRR encoding in question and in inverse proportion to the dimensionality of the HRR (Plate, 1994). As noise tolerance is a requirement of any neurologically plausible model, this loss of representation information is acceptable, and the "cleanup" method of recognizing encoded HRR vectors using the dot product can be used to find the vector that best fits what was decoded (Eliasmith & Thagard, 2001). Note that HRRs may also be combined by simple superposition (i.e., addition): $P = Q \otimes R + X \otimes Y$, where $P \# R \approx Q$, $P \# X \approx Y$, and so on. The operations required for convolution and correlation can be implemented in a recurrent connectionist network (Plate, 1993) and in particular under the NEF (Eliasmith, 2005).

In brief, the new model of abductive inference involves several large, high-dimensional populations to represent the data stored via HRRs and learned HRR transformations (the main output of the model), and a smaller population representing emotional valence information (abduction only requires considering emotion scaling from surprise to satisfaction, and hence only needs a single dimension represented by as few as 100 neurons to represent emotional changes). The model is initialized with a base set of causal encodings consisting of 100-dimensional HRRs combined in the form

$$antecedent \otimes `a' + relation \otimes causes + consequent \otimes `b',$$

as well as HRRs that represent the successful explanation of a target '*x*' (*expl* ⊗ '*x*'). For the purposes of this model, only six different "filler" values were used, representing three such causal rules ('*a*' causes '*b*', '*c*' causes '*d*', and '*e*' causes '*f*'). The populations used have between 2000 and 3200 neurons each and are 100- or 200-dimensional, which is at the lower-end of what is required for accurate HRR cleanup (Plate, 1994). More rules and filler values would require larger and higher-dimensional neural populations, an expansion that is unnecessary for a simple demonstration of abduction using biologically plausible neurons.

Following detection of a surprising *'b'* , which could be an event, proposition, or any sensory or cognitive data that can be represented via neurons, the change in emotional valence spurs activity in the output population towards generating a hypothesized explanation. This process involves employing several neural populations (representing the memorized rules and HRR convolution/correlation operations) to find an antecedent involved in a causal relationship that has *'b'* as the consequent. In terms of HRRs, this means producing (*rule # antecedent*) for [(*rule # relation ≈ causes*) and (*rule # consequent ≈ 'b'*)]. This production is accomplished in the 2000-neuron, 100-dimensional output population by means of associative learning through recurrent connectivity and connection weight updating (Eliasmith, 2005). As activity in this population settles, an HRR cleanup operation is performed to obtain the result of the learned transformation. Specifically, some answer is "chosen" if the cleanup result matches one encoded value significantly more than any of the others (i.e., is above some reasonable threshold value).

After the successful generation of an explanatory hypothesis, the emotional valence signal is reversed from surprise (which drove the search for an explanation) to what can be considered pleasure or satisfaction derived from having arrived at a plausible explanation. This in turn induces the output population to produce a representation corresponding to the successful dispatch of the explanandum '*b*': namely, the HRR $expl_b = expl \otimes$ *'b'*. Upon settling, it can thus be said that the model has accepted the hypothesized cause obtained in the previous stage as a valid explanation for the target '*b*'.

Settling completes the abductive inference: emotional valence returns to a neutral level, which suspends learning in the output population and causes population firing to return to basal levels of activity.

Figure 3 shows the result of performing the process of abductive inference in the neural model, with activity in the output population changing with respect to changing emotional valence, and vice versa. The output population activity is displayed by dimension, rather than individual neuron, since the 100-dimensional HRR output of the neural ensemble as a whole is the real characterization of what is being represented. The boxed sets of numbers represent the results of HRR cleanups on the output population at different points in time; if one value reasonably dominates over the next few largest, it can be taken to be the “true” HRR represented by the population at that moment. In the first stage, the high emotional valence leads to the search for an antecedent of a causal rule" for *b*, the surprising explanandum. The result is an HRR cleanup best fitting to *a*, which is indeed the correct response. Reaching an answer with a reasonably high degree of certainty triggers an emotional valence shift (from surprise to satisfaction), which in turn causes the output population to represent the fact that *b* has been successfully explained, as represented by the HRR cleanup in the second stage of the graph. Finally, the emotional arousal shifts to a neutral state as abduction is completed, and the population returns to representing nothing particularly strongly in the final stage.

The basic process of abduction outlined previously (see Fig. 2) maps very well to the results obtained from the model. The output population generates a valid hypothesis when surprised (since “*a* causes *b*” is the best memorized rule available to handle surprising *'b'*), and reversal of emotional valence corresponds to an acceptance of the hypothesis, and hence the successful explanation of '*b*'.

In sum, the model of abduction outlined here demonstrates how emotion can influence neural activity underlying a cognitive process. Emotional valence acts as a *context gate* that determines whether the output neural ensemble must conduct a search for some explanation for surprising input, or whether some generated hypothesis needs to be evaluated as a suitable explanation for the surprising input.

The neural network model just described provides a mechanism for explanation, its emotional input and output, and a simple kind of abduction. It also does a very simple sort of explanation evaluation, in that the causal rule that it selects from memory is chosen because it is a good match for the problem at hand, namely explaining *b*. Obviously, however, this model is too simple to account for the comparative evaluation of explanatory theories as performed by the cognitive model ECHO (Thagard, 1989, 1992, 2000). In ECHO, hypotheses and pieces of evidence are represented by simple artificial neurons called units, which are connected by excitatory or inhibitory links that correspond to constraints between the propositions they represent. For example, if a hypothesis explains a piece of evidence, then there is a symmetric excitatory link between the unit that represents the hypothesis and the unit that represents the evidence. If two hypotheses contradict each other, then there is a symmetric inhibitory link between the two units that represent them. Units have activations that spread between them until the network reaches stable activation levels, which typically takes 60-100 iterations. If a unit ends up with positive activation, the proposition that it represents is accepted, whereas if a unit ends up with negative activation, the proposition that it represents is rejected.

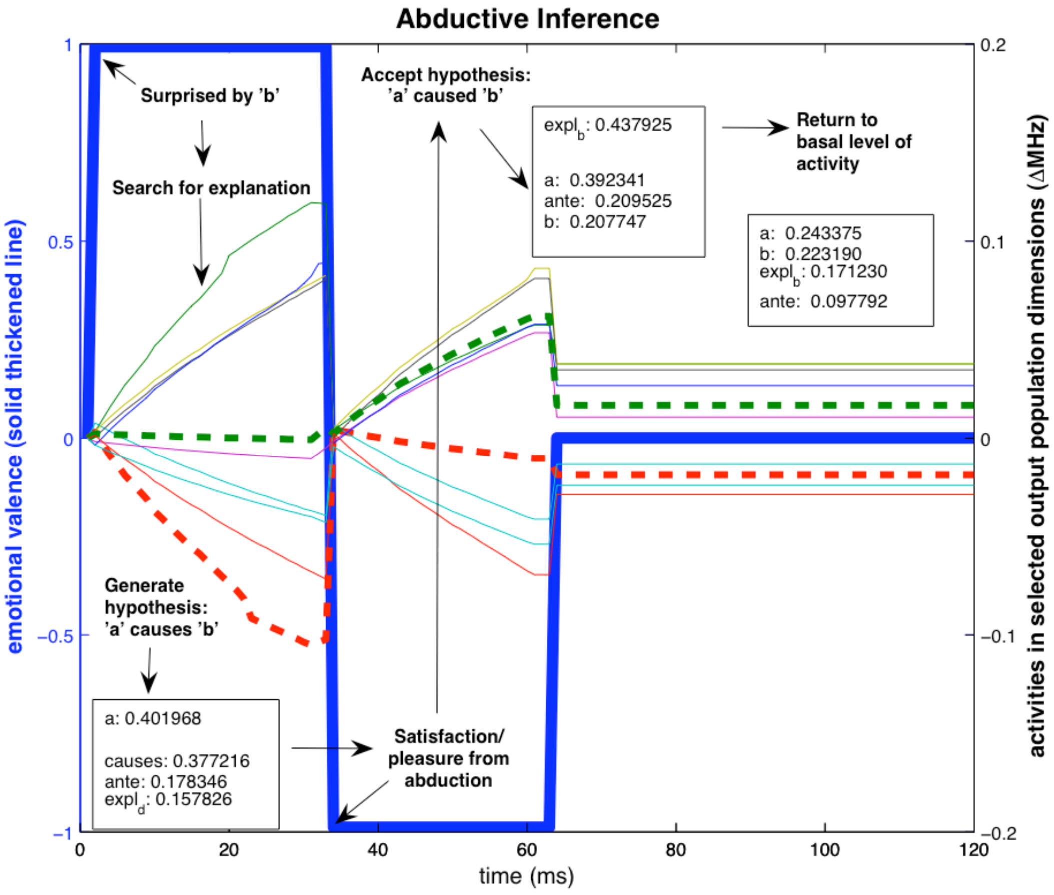

Figure 3: Neural activity in output population for abduction. For clarity, only a small (evenly spaced) selection of dimensional firing activities is displayed here (the full 2000-neuron population has 100 dimensions). Activities for two specific population dimensions are highlighted by thickened dashed or dotted lines, to demonstrate the neural activity changes in response to changing emotional valence (shown as a thickened solid line).

ECHO has been used to model numerous cases in the history of science, and has also inspired experimental research in social and educational psychology (Read and Marcus-Newhall, 1993; Schank and Ranney, 1991)   It shows how a very high-level kind of cognition, evaluating complex theories, can be performed by simple neural network performing parallel constraint satisfaction.   ECHO has a degree of psychological plausibility, but for neurological plausibility it pales in comparison to the NEF model of abduction described earlier in this section.   The largest ECHO model uses only around 200 units to encode the same number of propositions, whereas the NEF model uses thousands of spiking neurons to encode a few causal relations. Computationally, this seems inefficient, but of course the brain has many billions of neurons that provide its distributed representations.

How might one implement comparative theory evaluation as performed by ECHO within the NEF framework? Aubie and Thagard (in preparation) use the NEF to encode ECHO networks by generating a population of thousands of neurons.   Parallel constraint satisfaction is performed by transformations of neurons that carry out approximately the same calculations that occur more directly in ECHO's localist neural networks. Hence it is now possible to model evaluation of competing explanations using more biologically realistic neural networks.

CAUSALITY

Like most other models of explanation, these neural network models presuppose some understanding of causality. In one sense that is common in both science and everyday life, to explain something involves stating its cause.   For example, when people have influenza, the virus that infects them is the cause of their symptoms such as fever. But what is a cause? Philosophical theories of explanation correlate with competing theories of causality; for example, the deductive view of explanation fits well with the Humean understanding of causality as constant conjunction. If all *A* are *B*, then someone can understand how being *A* can cause and explain being *B*. Unfortunately, universality is not a requisite of either explanation or causality. Smoking causes lung cancer, even though many smokers never get lung cancer, and some people with lung cancer never smoked. Schematic models of explanation presuppose a primitive concept of causation without being able to say much about it. Probability theory may look like a promising approach to causality in that causes make their effects more probable than they would be otherwise, but such increased probability may be accidental or the result of some common cause. For example, the probability of someone drowning is greater on a day when much ice cream is consumed, but that is because of the common cause that more people go swimming on hot days. Sorting out causal probabilistic information from misleading correlations requires much information about probability and independence that people usually lack.

Thagard (forthcoming) conjectured that it might be possible to give a neural network account of how organisms understand causality, Suppose, in keeping with research on infants' grasp of causality, that cause is a preverbal concept based on perception and motor control (Baillargeon, Kotovsky, and Needham, 1996; Mandler, 2004). Consider an infant of a few months old, lying on its back swiping at a mobile suspended over its head. The infant has already acquired an image schema of the following form:

perception of situation + motor behavior => perception of new situation.

Perhaps this schema is innate, but alternatively it may have been acquired from very early perceptual/motor experiences in which the infant acted on the world and perceived its changes. A simple instance of the schema would be:

stationary object + hand hitting object => moving object.

The idea of a preverbal image schema for causality is consistent with the views of some philosophers that manipulability and intervention are central features of causality (Woodward, 2004). The difference between A causing B and A merely being correlated with B is that manipulating A also manipulates B in the former case but not the latter. Conceptually, the concepts of manipulation and intervention seem to presuppose the concept of causation, because making something happen is on the surface no different from causing it to happen. However, although there is circularity at the verbal level, psychologically it is possible to break out of the circle by supposing that people have from infancy a neural encoding of the causality image schema described above. This nonverbal schema is the basis for understanding the difference between one event making another event happen and one event just occurring after the other.

The causality image schema is naturally implemented within the Neural Engineering Framework used to construct the model of abductive inference. Neural populations are capable of encoding both perceptions and motor behaviors, and also capable of encoding relations between them. In the model of abductive inference described in the last section, *cause (c, e)* was represented by a neural population that encodes an HRR vector that captures the relation between a vector representing *c* and a vector representing *e*, where both of these can easily be nonverbal perceptions and actions as well as verbal representations. In the NEF model of abduction, there is no real understanding of causality, because the vector was generated automatically. In contrast, it is reasonable to conjecture that people have neural populations that encode the notion of causal connection as the result of their very early preverbal experience with manipulating objects. Because the connection is based on visual and kinesthetic experiences, it cannot be adequately formulated linguistically, but it provides the intellectual basis for the more verbal and mathematical characterizations of causality that develop later.

If this account of causality is correct, then a full cognitive model of explanation cannot be purely verbal or probabilistic. Many philosophers and cognitive scientists currently maintain that scientific explanation of phenomena consists of providing mechanisms that produce them (e.g. Bechtel and Abrahamsen, 2005; Sun, Coward, and Zenzen, 2005). A mechanism is a system of objects whose interactions regularly produce changes. All of the computational models described in this chapter are mechanistic, although they differ in what they take to be the parts and interactions that are central to explaining human thinking; for the neural network approaches, the computational mechanisms are also biological ones. But understanding of mechanism presupposes understanding of causality, in that there must be a relation between the interactions of the parts that constitutes production of the relevant phenomena. Because scientific explanation depends on the notion of causality, and because understanding of causality is in part visual and kinesthetic, future comprehensive cognitive models of explanation will need to incorporate neural network simulations of people's nonverbal understanding of causality.

CONCLUSION

This chapter has reviewed four major computational approaches to understanding scientific explanations: deductive, schematic, probabilistic, and neural network. Table 1 summarizes the different approaches to providing and generating explanations. To some extent, the approaches are complementary rather than competitive, because explanation can take different forms in different areas of science and everyday life. However, at the root of scientific and everyday explanation is an understanding of causality represented nonverbally in human brains by populations of neurons encoding how physical manipulations produce sensory changes. Another advantage of taking a neural network approach to explanation is that it becomes possible to model how abductive inference, the generation of explanatory hypotheses, is a process that is multimodal, involving not only verbal representations but also visual and emotional ones that constitute inputs and outputs to reasoning.

	target of explanation	explainers	relation between target and explainers	mode of generation
deductive	sentence	sentences	deduction	backward chaining
schema	sentence	pattern of sentences	fit	search for fit, schema generation
probabilistic	variable node	Bayesian network	conditional probability	Bayesian learning
neural network	neural group: multimodal representation	neural groups	gated activation, connectivity	search, associative learning

Table 1. Summary of approaches to computational modeling of explanation.

Acknowledgements. This research was supported by the Natural Sciences and Engineering Research Council of Canada. The authors are very grateful to Chris Eliasmith for much help in applying NEF.

References

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

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

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

Baillargeon, R., Kotovsky, L., & Needham, A. (1995). The acquisition of physical knowledge in infancy. In D. Sperber, D. Premack & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 79-116). Oxford: Clarendon Press.

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

Bylander, T., Allemang, D., Tanner, M., & Josephson, J. (1991). The computational complexity of abduction. *Artificial Intelligence, 49*, 25-60.

Conklin, J., & Eliasmith, C. (2005). An attractor network model of path integration in the rat. *Journal of Computational Neuroscience, 18*, 183-203.

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

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

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

Falkenhainer, B. (1990). A unified approach to explanation and theory formation. In J. Shrager & P. Langley (Eds.), *Computational models of discovery and theory formation*. (pp. 157-196). San Mateo, CA: Morgan Kaufman.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithms and examples. *Artificial Intelligence, 41*, 1-63.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence, 24*.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Glymour, C. (2003). Learning, prediction, and causal Bayes nets. *Trends in Cognitive Sciences, 7*, 43-48.

Gopnik, A., Glymour, C., Sobel, D. M., Schultz, L. E., Kushur, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 2004*, 3-32.

Gottman, J. M., Tyson, R., Swanson, K. R., Swanson, C. C., & Murray, J. D. (2003). *The mathematics of marriage: Dynamic nonlinear models*. Cambridge, MA: MIT Press.

Harman, G. (1973). *Thought*. Princeton: Princeton University Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press/Bradford Books.

Josephson, J. R., & Josephson, S. G. (Eds.). (1994). *Abductive inference: Computation, philosophy, technology*. Cambridge: Cambridge University Press.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Kitcher, P. (1993). *The advancement of science*. Oxford: Oxford University Press.

Kitcher, P., & Salmon, W. (1989). *Scientific explanation*. Minneapolis: University of Minnesota Press.

Konolige, K. (1992). Abduction versus closure in causal theories. *Artificial Intelligence, 53*, 255-272.

Leake, D. B. (1992). *Evaluating explanations: A content theory*. Hillsdale, NJ: Erlbaum.

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

Mandler, J. M. (2004). *The foundations of mind: Origins of conceptual thought*. Oxford: Oxford University Press.

Neapolitan, R. (1990). *Probabilistic reasoning in expert systems*. New York: John Wiley.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. Cambridge, MA: MIT Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Mateo: Morgan Kaufman.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

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

Read, S., & Marcus-Newhall, A. (1993). The role of explanatory coherence in the construction of social explanations. *Journal of Personality and Social Psychology, 65*, 429-447.

Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (second ed.). Upper Saddle River, NJ: Prentice Hall.

Salmon, W. (1970). Statistical explanation. In R. Colodny (Ed.), *The nature and function of scientific theories* (pp. 173-231). Pittsburgh: University of Pittsburgh Press.

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

Schank, P., & Ranney, M. (1991). Modeling an experimental study of explanatory coherence. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 892-897). Hillsdale, NJ: Erlbaum.

Schank, R. C. (1986). *Explanation patterns: Understanding mechanically and creatively*. Hillsdale, NJ: Erlbaum.

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

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York: Springer-Verlag.

Sun, R., Coward, L. A., & Zenzen, M. J. (2005). On levels of cognitive modeling. *Philosophical Psychology, 18*, 613-637.

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

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

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

Thagard, P. (1999). *How scientists explain disease*. Princeton: Princeton University Press.

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

Thagard, P. (2003). Pathways to biomedical discovery. *Philosophy of Science, 70*, 235-254.

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

Thagard, P. (forthcoming). Abductive inference: From philosophical analysis to neural mechanisms. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Cognitive, mathematical, and neuroscientific approaches*. Cambridge: Cambridge University Press.

Woodward, J. (2003). *Scientific explanation*. Retrieved August 11, 2005, from http://plato.stanford.edu/entries/scientific-explanation/

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