Evaluating Explanations in Law, Science, and Everyday Life

Paul Thagard

Department of Philosophy, University of Waterloo, Waterloo, Ontario, Canada

ABSTRACT—*This article reviews a theory of explanatory coherence that provides a psychologically plausible account of how people evaluate competing explanations. The theory is implemented in a computational model that uses simple artificial neural networks to simulate many important cases of scientific and legal reasoning. Current research directions include extensions to emotional thinking and implementation in more biologically realistic neural networks.*

KEYWORDS—*explanation; coherence; neural networks; legal reasoning; emotion*

In *CSI* and other television crime shows, investigators collect evidence in order to determine the causes of a crime. For example, if a young woman is murdered, the police may consider as suspects the woman's boyfriend and her father. Inferences about who is the most likely culprit will be based on which hypothesis—that the boyfriend did it or that the father did it—fits best with all the available evidence. These hypotheses provide possible explanations of the evidence; for example the hypothesis that the boyfriend was the murderer may explain why his fingerprints are on the murder weapon. Conclusions about who the actual criminal was and who was innocent depend on evaluating competing explanations of the evidence.

This kind of explanatory inference is ubiquitous in human thinking, ranging from mechanical repair to medical diagnosis to scientific theorizing. When your car fails to start, you consider alternative explanations such as that it is out of gas or that the battery is dead. In medicine, a physician considers possible diseases that would explain a patient's symptoms and bases a treatment plan on what he or she thinks is the most plausible diagnosis. Psychologists publishing theoretical papers often offer sets of hypotheses that they contend provide better explanations of the results of experiments than alternative theories do.

Address correspondence to Paul Thagard, Philosophy Department, University of Waterloo, Waterloo, ON N2L 3G1, Canada; e-mail: pthagard@uwaterloo.ca.

Explanation evaluation is a mental process that is important in many areas of psychology. Cognitive psychologists have investigated causal reasoning, which often requires a person to determine the most likely cause of a surprising event. Social psychologists have studied how people explain the behavior of others. Clinical psychologists are sometimes interested in the emotion-laden reasoning by which people construct explanations of their own situations. In all these kinds of cases, people's thinking involves evaluating competing explanations of what they observe.

But explanation evaluation is not simply a matter of determining which of two or more competing hypotheses fits best with the evidence. We may also need to consider how hypotheses fit with each other, particularly when one hypothesis provides an explanation of another. This layering of hypotheses is particularly evident in legal reasoning when questions of motive are salient. Crime investigators considering whether the boyfriend or the father is the more likely murderer will naturally consider possible motives that might explain why one of them would have wanted to kill the young woman. Hence the cognitive process of explanation evaluation must consider the fit of hypotheses with each other as well as with the evidence, so that inference involves coming up with the overall most coherent picture of what happened.

This article reviews a theory of explanatory coherence that provides a psychologically plausible account of how people evaluate competing explanations. After sketching the theory, I describe how it is implemented in a computational model that uses a simple artificial neural network to evaluate competing explanations. This model has been applied to many important cases of scientific and legal reasoning. Finally, I describe current directions in the development and application of the theory of explanatory coherence, including connections with emotional thinking and implementation in more biologically realistic neural networks.

EXPLANATORY COHERENCE: THE THEORY

Table 1 lists seven principles that concisely state the theory of explanatory coherence (Thagard, 1989, 1992, 2000). These

(CDIR 424 | CDIR Journal Name | 424 Manuscript No. | B | Dispatch: 8.5.06 Journal: CDIR | CE: Blackwell | Author Received: No. of pages: 5 | PE: Sarvanan/Mini)

TABLE 1
Principles of Explanatory Coherence

Principle	Statement
E1. Symmetry	Explanatory coherence is a symmetrical relation. That is, two propositions *P* and *Q* cohere with each other equally.
E2. Explanation	(a) A hypothesis coheres with what it explains, which can either be evidence or another hypothesis; (b) hypotheses that together explain some other proposition cohere with each other; and (c) the more hypotheses it takes to explain something, the lower the degree of coherence.
E3. Analogy	Similar hypotheses that explain similar pieces of evidence cohere with each other.
E4. Data priority	Propositions that describe the results of observations have a degree of acceptability on their own.
E5. Contradiction	Contradictory propositions are incoherent with each other.
E6. Competition	If *P* and *Q* both explain a proposition, and if *P* and *Q* are not explanatorily connected, then *P* and *Q* are incoherent with each other. (*P* and *Q* are explanatorily connected if one explains the other or if together they explain something.)
E7. Acceptance	The acceptability of a proposition in a system of propositions depends on its coherence with them.

principles are rather abstract, so let me explain them in terms of the legal example already introduced. The hypothesis that the boyfriend killed the woman explains the evidence that the woman is dead, so the hypothesis and the evidence cohere with each other, in accord with principle E2, Explanation. Although the relation between the hypothesis and evidence is asymmetrical, with the former explaining the latter and not vice versa, the coherence relation between them is symmetrical: They hang together equally, as indicated by principle E1, Symmetry. Principle E2 also allows the possibility of hypotheses explaining each other, as when the hypothesis that the boyfriend is the murderer is explained by the motive that he was jealous. Explanation can involve multiple hypotheses—for example that the boyfriend was both jealous and angry—that then cohere with each other. However, E2 includes a simplicity principle in clause (c), so that hypotheses that involve many hypotheses will have less coherence. For example, the theory that the woman was killed by space aliens who arrived from Alpha Centauri and singled her out for execution because of her hair color requires multiple hypotheses that lack simplicity as well as independent support. Simplicity is a matter of explaining a lot with few assumptions. The theory of explanatory coherence is neutral about what constitutes an explanation, but I have argued independently that good explanations are based on causal mechanisms (Thagard, 1999).

According to principle E3, Analogy, explanations can gain coherence by virtue of being analogous to ones already accepted. For example, if the boyfriend had a past history of being jealously angry with girlfriends and assaulting them, then these cases provide analogies that make the hypothesis that the boyfriend did it more plausible in the current case. Principle E4, Data Priority, says that observational evidence gets a degree of coherence on its own, providing a degree of priority to such observations as that the woman is dead and the boyfriend's fingerprints are on a knife found near the body. This principle does not require that observations be indubitable, but leaves open the possibility that observations could be found to be erroneous despite their initial degree of coherence.

Principles E5 and E6 deal with competing hypotheses that are incoherent with each other. E5, Contradiction, handles the most straightforward case in which two hypotheses are logically contradictory; but typically the relation between competitors is looser, as captured in E6, Competition. Normally, we treat the hypothesis that the boyfriend was the murderer as competing with the hypothesis that the father did it, even though these are not contradictory: It is logically possible that the boyfriend and the father together killed the woman. But if there is reason to suspect that the boyfriend and the father acted together in a conspiracy, then the two hypotheses—the boyfriend did it and the father did it—are explanatorily connected, so they should be treated as coherent with each other rather than incoherent. Ordinarily, however, two hypotheses that independently explain evidence will be treated as competitors that are incoherent with each other.

Finally, principle E7, Acceptance, states that we should accept and reject propositions on the basis of their overall coherence with each other. Because hypotheses and evidence can be coherent and incoherent with each other in many ways, E7 makes inference a highly complex and nonlinear process. We cannot simply accept the evidence and then accept a hypothesis and then reject its competitors, because evidence and competing hypotheses must all be evaluated together with respect to how they fit with each other. This makes explanation evaluation sound like a very mysterious holistic process, but I will now describe how a simple artificial neural network can perform the required computation.

EXPLANATORY COHERENCE: THE MODEL

The first step in implementing explanatory coherence computationally is to represent each proposition by a unit, a highly simplified artificial neuron that is connected to other units by excitatory and inhibitory links. As in real neurons, an excitatory link is one that enables one neuron to increase the firing of another, whereas an inhibitory link decreases firing. In the crime example, the hypothesis that the boyfriend is the murderer can

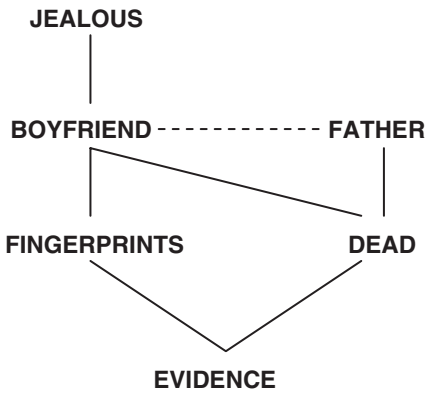

Fig. 1. Neural network modeling competing explanations for the murder of a woman. Solid lines are excitatory links between units, and the dotted line is an inhibitory link representing incoherence between competing hypotheses about who committed the crime—the dead woman's boyfriend (whose fingerprints were found on the murder weapon and who is hypothesized to have a jealous motive) or her father.

be represented by a unit called BOYFRIEND and the evidence that the woman is dead by a unit called DEAD. Then, whenever principles E2 and E3 establish relations of coherence between two propositions, the units that represent the propositions get excitatory links between them. Thus BOYFRIEND and DEAD have an excitatory link between them that is symmetrical, in accord with principle E1. Principle E4 is implemented by making an excitatory link between the special unit EVIDENCE and any unit such as DEAD that represents a proposition based on observation. Principles E5 and E6, which establish incoherence between competing hypotheses, are implemented by means of inhibitory links between units: When two hypotheses are incoherent—e.g., the boyfriend did it versus the father did it—then the units that represent the hypotheses—BOYFRIEND and FATHER—will get an inhibitory link between them.

Figure 1 depicts the simple network that evaluates competing explanations in my murder case. It includes a unit called JEALOUS that represents the hypothesis that the boyfriend was jealous, and a unit called FINGERPRINTS that represents the evidence that the boyfriend's fingerprints were found on a knife near the dead body. Notice the excitatory links between units representing coherent propositions and the inhibitory links between units representing incoherent propositions. In simulations, the links have different weights that can represent the degree of coherence or incoherence between propositions.

Representation of propositions by units that have excitatory and inhibitory links to each other makes possible the overall computation of coherence as required by principle E7. Degree of acceptance of propositions is modeled by the activation of units, which can range from 1 (acceptance) to −1 (rejection). Running the network shown in figure 1 using the algorithm given in Table 2 will lead to activation of the unit BOYFRIEND and deactivation of the unit FATHER, because the former gets more activation thanks to its additional links. This models the judgment that the hypothesis that the boyfriend did it has greater explanatory coherence.

This computational model, called ECHO, has been applied to many complex examples of legal, scientific, and everyday reasoning (Thagard 1989, 1992, 1999, 2000, 2004, 2005). Read and Marcus-Newhall (1993) found that explanations of social behavior operate in accord with principles of explanatory coherence, and Simon (2004) reviews experimental studies of legal decision making that are naturally understood in terms of coherence. Ranney and Schank (1998) discuss the relevance of explanatory coherence for understanding and improving scientific reasoning by students. Coherence models similar to ECHO have been successful in modeling other cognitive processes such as stereotype application (Kunda and Thagard, 1996).

CURRENT DIRECTIONS

Current research on explanatory coherence proceeds in several directions, especially extension to emotional reasoning and modeling using more biologically realistic neural networks. Explanation evaluation is often a highly emotional enterprise. A scientist with a favorite theory will react to a challenging alternative not merely with disbelief but possibly also with annoyance or even more negative emotions. In legal cases, the prosecution and the defense will have very different emotional attitudes toward the prospect of the accused being convicted, and obviously the accused and his or her supporters will react with intensely negative emotions toward the prospect of conviction. Ideally, the judge and jury are supposed to be neutral, but they are as prone as anyone else to affective biases. In the simple case in Figure 1, the mother of the murdered woman

TABLE 2
Algorithm for Running a Neural Network That Computes Coherence

Step	Procedure
1	Set the activation of all units to 0, except EVIDENCE, which gets activation 1.
2	Repeatedly and in parallel spread activation among the units, with the new activation of a unit being determined by (a) its previous activation, (b) its excitatory and inhibitory links with other units, (c) the activations of the units to which it is linked, and (d) a decay factor.
3	Spread activation until the network has settled—that is, until the activation of all units has stabilized, with no unit changing activation. Typically, it only takes about 100 cycles of repeatedly updating activations for the network to settle. Then interpret propositions whose units have positive activation as being accepted, and interpret propositions whose units have negative activation as being rejected.

TABLE 3
Principles of Emotional Coherence

Principle	Statement
1. Valence	Elements have positive or negative emotional valences as well as degrees of acceptability. Elements can be concepts, propositions, or other representations.
2. Connection	Elements can have positive or negative emotional connections to other elements.
3. Determination	The valence of an element is determined by the valences and acceptability of all the elements to which it is connected.

would probably be more emotionally inclined to view as guilty whichever man—the boyfriend or the father—she liked least.

Accordingly, I extended explanatory coherence and its kindred coherence theories into a theory of emotional coherence (Thagard, 2000, 2003), which applies to a range of elements besides the propositions used in explanatory coherence. The theory can be summed up in the principles in Table 3. People attach negative valences to concepts such as vomit, actions such as doing boring chores, and propositions such as that it will be a cold winter. Our overall emotional reaction to a situation requires balancing the positive and negative valences of all the elements of the situation.

How this balancing can work is shown by a computational model called HOTCO, for "hot coherence." HOTCO works like ECHO, spreading activation between units in a network of artificial neurons, but also attaches numerical valences between units and spreads them based on their positive and negative emotional connections. For example, if a unit for vomit is connected positively to a unit for whiskey, then negative valence will spread from the former to the latter. Figure 2 shows an emotional expansion of the network in Figure 1 from the perspective of someone who strongly dislikes the boyfriend but likes the father. Disliking the boyfriend creates a positive valence for inferring that he is the murderer. If valences are allowed to influence activations, then the network in Figure 2 will be emotionally biased toward believing that the boyfriend is the murderer.

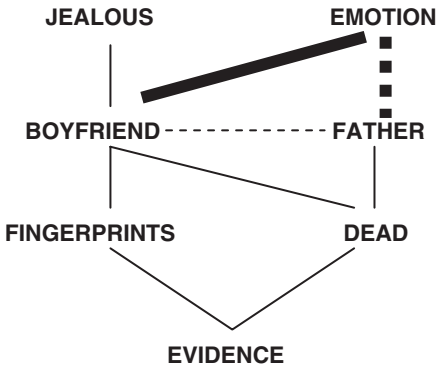

Fig. 2. Neural network modeling competing explanations for a woman's murder plus emotional bias affecting a person's judgment of the case. Solid lines are excitatory links between units, and dotted lines are inhibitory links. Thick lines indicate positive and negative emotional connections. The unit BOYFRIEND gets a positive emotional link because it represents the hypothesis that the disliked boyfriend is the murderer.

HOTCO has been used to model emotional bias in the notorious trial of O.J Simpson, who was acquitted of murdering his former wife (Thagard, 2003). Many observers thought that the evidence showed Simpson to be guilty, and an explanatory-coherence simulation using ECHO, based on the details of the trial, judged that the best explanation of the case is that Simpson was the murderer. However, adding in emotional bias in favor of Simpson and against the Los Angeles Police Department produces a HOTCO simulation that duplicates the actual decision of the jury to acquit Simpson. HOTCO has also been used to simulate psychological experiments involving other sorts of emotional biases including political attitudes and racial prejudice.

One obvious problem with both the ECHO and HOTCO artificial-neural-network models is that they are extremely unlike the biological neural networks used by brains. They use single neuron-like units to represent entire propositions and concepts that real brains distribute over vast numbers of neurons that collectively correspond to many different representations. Moreover, the units in ECHO and HOTCO operate with simple rates of activation and valence, whereas real neurons have spiking patterns that give them more representational power than mere rates of spiking can achieve. (A rate of activation is how often a neuron fires in a given interval, whereas a spiking pattern is a specific sequence of firing and resting.) Another respect in which ECHO and HOTCO are biologically unrealistic is that they do not organize neurons into the kinds of functional areas found in the brain. Hence they do not interface with models of inference, such as the GAGE model of Wagar and Thagard (2004), that use distributed representations with spiking neurons organized into brain regions. More neurologically realistic neural networks that integrate cognition and emotion have the potential to illuminate many aspects of human inference and decision making. For example, the GAGE model sheds light on why moral reasoning is so easily distorted by conflicts of interest (Thagard, in press).

My research group is now working on more biologically realistic computational models of high-level reasoning within the neural-engineering framework of Eliasmith and Anderson (2003). We already have a model that generates explanatory hypotheses using representations made of populations of spiking neurons (Thagard and Litt, in press). This model also captures the emotional inputs to explanation, such as puzzlement, and the emotional outputs, such as satisfaction, and can model hypotheses and evidence represented in sensory modalities as well as

verbally. Another new model under development attributes people's tendency in decision making to differently evaluate gains and losses to separate brain processes involving different neurotransmitters. Thus work is underway to determine how explanatory and emotional coherence can be computed in biologically realistic neural networks.

Recommended Reading

Thagard, P. (1992). (See References)

Thagard, P. (2000). (See References)

Thagard, P. (in press). *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge, MA: MIT Press.

Wagar, B.M., & Thagard, P. (2004). (See References)

Acknowledgments—This research has been funded by the Natural Sciences and Engineering Research Council of Canada. I am grateful to the editor and anonymous referees for helpful comments.

REFERENCES

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

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*, 284–308.

Ranney, M., & Schank, P. (1998). Toward an integration of the social and the scientific: Observing, modeling, and promoting the explanatory coherence of reasoning. In S.J. Read & L.C. Miller (Eds.), *Connectionist models of social reasoning and social behavior* (pp. 245–274). Mahwah, NJ: Erlbaum.

Read, S.J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, *65*, 429–447.

Simon, D. (2004). A third view of the black box: Cognitive coherence in legal decision making. *University of Chicago Law Review*, *71*, 511–586.

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

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

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

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

Thagard, P. (2003). Why wasn't O.J. convicted? Emotional coherence in legal inference. *Cognition and Emotion*, *17*, 361–383.

Thagard, P. (2004). Causal inference in legal decision making: Explanatory coherence vs. Bayesian networks. *Applied Artificial Intelligence*, *18*, 231–249.

Thagard, P. (2005). Testimony, credibility, and explanatory coherence. *Erkenntnis*, *63*, 295–317.

Thagard, P. (in press). The moral psychology of conflicts of interest: Insights from affective neuroscience. *Journal of Applied Philosophy*.

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

Wagar, B.M., & Thagard, P. (2004). Spiking Phineas Gage: A neurocomputational theory of cognitive-affective integration in decision making. *Psychological Review*, *111*, 67–79.

Author Query Form

Journal	**CDIR**
Article	**424**

Dear Author,

During the copy-editing of your paper, the following queries arose. Please respond to these by marking up your proofs with the necessary changes/additions. Please write your answers clearly on the query sheet if there is insufficient space on the page proofs. If returning the proof by fax do not write too close to the paper's edge. Please remember that illegible mark-ups may delay publication.

Query No.	Description	Author Response
.	No Queries	