BEHAVIORAL AND BRAIN SCIENCES (1989) 12, 435–502
Printed in the United States of America

Explanatory coherence

Paul Thagard
Cognitive Science Laboratory, Princeton University, 221 Nassau St., Princeton, NJ 08540
Electronic mail: *pault@confidence.princeton.edu*

Abstract: This target article presents a new computational theory of explanatory coherence that applies to the acceptance and rejection of scientific hypotheses as well as to reasoning in everyday life. The theory consists of seven principles that establish relations of local coherence between a hypothesis and other propositions. A hypothesis coheres with propositions that it explains, or that explain it, or that participate with it in explaining other propositions, or that offer analogous explanations. Propositions are incoherent with each other if they are contradictory. Propositions that describe the results of observation have a degree of acceptability on their own. An explanatory hypothesis is accepted if it coheres better overall than its competitors. The power of the seven principles is shown by their implementation in a connectionist program called ECHO, which treats hypothesis evaluation as a constraint satisfaction problem. Inputs about the explanatory relations are used to create a network of units representing propositions, while coherence and incoherence relations are encoded by excitatory and inhibitory links. ECHO provides an algorithm for smoothly integrating theory evaluation based on considerations of explanatory breadth, simplicity, and analogy. It has been applied to such important scientific cases as Lavoisier's argument for oxygen against the phlogiston theory and Darwin's argument for evolution against creationism, and also to cases of legal reasoning. The theory of explanatory coherence has implications for artificial intelligence, psychology, and philosophy.

Keywords: artificial intelligence; attribution theory; coherence, connectionism; epistemology; explanation; legal reasoning; scientific reasoning; theory evaluation

1. Introduction

Why did the oxygen theory of combustion supersede the phlogiston theory? Why is Darwin's theory of evolution by natural selection superior to creationism? How can a jury in a murder trial decide between conflicting views of what happened? This target article develops a theory of explanatory coherence that applies to the evaluation of competing hypotheses in cases such as these. The theory is implemented in a connectionist computer program with many interesting properties.

The problem of inference to explanatory hypotheses has a long history in philosophy and a much shorter one in psychology and artificial intelligence (AI). Scientists and philosophers have long considered the evaluation of theories on the basis of their explanatory power. In the late nineteenth century, Peirce discussed two forms of inference to explanatory hypotheses: *hypothesis*, which involved the acceptance of hypotheses, and *abduction*, which involved merely the initial formation of hypotheses (Peirce 1931–1958; Thagard 1988a). Researchers in artificial intelligence and some philosophers have used the term "abduction" to refer to both the formation and the evaluation of hypotheses. AI work on this kind of inference has concerned such diverse topics as medical diagnosis (Josephson et al. 1987; Pople 1977; Reggia et al. 1983) and natural language interpretation (Charniak & McDermott 1985; Hobbs et al. 1988). In philosophy, the acceptance of explanatory hypotheses is usually called *inference to the best explanation* (Harman 1973; 1986). In social psychology, attribution theory considers how people in everyday life form hypotheses to explain events (Fiske & Taylor 1984). Recently, Pennington and Hastie (1986; 1987) have proposed that much of jury decision making can be best understood in terms of explanatory coherence. For example, to gain a conviction of first-degree murder, the prosecution must convince the jury that the accused had a preformed intention to kill the victim. Pennington and Hastie argue that whether the jury will believe this depends on the explanatory coherence of the prosecution's story compared to the story presented by the defense.

Actual cases of scientific and legal reasoning suggest a variety of factors that go into determining the explanatory coherence of a hypothesis. How much does the hypothesis explain? Are its explanations economical? Is the hypothesis similar to ones that explain similar phenomena? Is there an explanation of why the hypothesis might be true? In legal reasoning, the question of explaining the hypothesis usually concerns motives: If we are trying to explain the evidence by supposing that the accused murdered the victim, we will find the supposition more plausible if we can think of reasons why the accused was motivated to kill the victim. Finally, on all these dimensions, how does the hypothesis compare against alternative hypotheses?

This paper presents a theory of explanatory coherence that is intended to account for a wide range of explanatory inferences. I shall propose seven principles of explanatory coherence that encompass the considerations just described and that suffice to make judgments of explanatory coherence. Their sufficiency is shown by the implementation of the theory in a connectionist computer program called ECHO that has been applied to more than a dozen complex cases of scientific and legal reasoning. My account of explanatory coherence thus has three parts:

© 1989 Cambridge University Press 0140-525X/89 $5.00+.00

the statement of a theory, the description of an algorithm, and applications to diverse examples that show the feasibility of the algorithm and help to demonstrate the power of the theory (cf. Marr 1982). Finally, I shall discuss the implications of the theory for artificial intelligence, psychology, and philosophy.

2. A theory of explanatory coherence

2.1. Coherence and explanation. Before presenting the theory, it will be useful to make some general points about the concepts of coherence and explanation, although it should be made clear that this paper does not purport to give a general account of either concept. The question of the nature of explanation is extremely difficult and controversial. Philosophers disagree about whether explanation is primarily deductive (Hempel 1965), statistical (Salmon 1970), causal (Salmon 1984), linguistic (Achinstein 1983), or pragmatic (van Fraassen 1980). In AI, explanation is sometimes thought of as deduction (Mitchell et al. 1986) and sometimes as pattern instantiation (Schank 1986). This paper does not pretend to offer a theory of explanation, but is compatible with any of the foregoing accounts (except van Fraassen's, which is intended to make explanation irrelevant to questions of acceptability and truth).

Nor does this paper give a general account of coherence. There are various notions of coherence in the literatures of different fields. We can distinguish at least the following:

Deductive coherence depends on relations of logical consistency and entailment among members of a set of propositions.

Probabilistic coherence depends on a set of propositions having probability assignments consistent with the axioms of probability.

Semantic coherence depends on propositions having similar meanings.

BonJour (1985) provides an interesting survey of philosophical ideas about coherence. Here, I am only offering a theory of *explanatory* coherence.

Explanatory coherence can be understood in several different ways, as

(a) a relation between two propositions,
(b) a property of a whole set of related propositions, or
(c) a property of a single proposition.

I claim that (a) is fundamental, with (b) depending on (a), and (c) depending on (b). That is, explanatory coherence is primarily a relation between two propositions, but we can speak derivatively of the explanatory coherence of a set of propositions as determined by their pairwise coherence, and we can speak derivatively of the explanatory coherence of a single proposition with respect to a set of propositions whose coherence has been established. A major requirement of an account of explanatory coherence is that it show how it is possible to move from (a) to (b) to (c); algorithms for doing so are presented as part of the computational model described below.

Because the notion of the explanatory coherence of an individual proposition is so derivative and depends on a specification of the set of propositions with which it is supposed to cohere, I shall from now on avoid treating coherence as a property of individual propositions. Instead, we can speak of the *acceptability* of a proposition, which depends on but is detachable from the explanatory coherence of the set of propositions to which it belongs. We should accept propositions that are coherent with our other beliefs, reject propositions that are incoherent with our other beliefs, and be neutral toward propositions that are neither coherent nor incoherent. Acceptability has finer gradations than just acceptance, rejection, and neutrality, however: The greater the coherence of a proposition with other propositions, the greater its acceptability.

In ordinary language, to cohere is to hold together, and explanatory coherence is a holding together because of explanatory relations. We can, accordingly, start with a vague characterization:

> Propositions P and Q cohere if there is some explanatory relation between them.

To fill this statement out, we must specify what the explanatory relation might be. I see four possibilities:

(1) P is part of the explanation of Q.
(2) Q is part of the explanation of P.
(3) P and Q are together part of the explanation of some R.
(4) P and Q are analogous in the explanations they respectively give of some R and S.

This characterization leaves open the possibility that two propositions can cohere for nonexplanatory reasons: deductive, probabilistic, or semantic. Explanation is thus sufficient but not necessary for coherence. I have taken "explanation" and "explain" as primitives, while asserting that a relation of explanatory coherence holds between P and Q if and only if one or more of (1)–(4) is true. *Incoherence* between two propositions occurs if they contradict each other or if they offer explanations that background knowledge suggests are incompatible.

The psychological relevance of explanatory coherence comes from the following general predictions concerning the acceptance of individual propositions:

If a proposition is highly coherent with the beliefs of a person, then the person will believe the proposition with a high degree of confidence.

If a proposition is incoherent with the beliefs of a person, then the person will not believe the proposition. The applicability of this to several areas of psychological experimentation is discussed in section 9.

2.2. Principles of explanatory coherence. I now propose seven principles that establish relations of explanatory coherence and make possible an assessment of the global coherence of an explanatory system S. S consists of propositions P, Q, and $P_1 \ldots P_n$. Local coherence is a relation between two propositions. I coin the term "incohere" to mean more than just that two propositions do not cohere: To incohere is to *resist* holding together. The principles are as follows:

Principle 1. Symmetry.

(a) If P and Q cohere, then Q and P cohere.
(b) If P and Q incohere, then Q and P incohere.

Principle 2. Explanation.

If $P_1 \ldots P_m$ explain Q, then:
(a) For each P_i in $P_1 \ldots P_m$, P_i and Q cohere.
(b) For each P_i and P_j in $P_1 \ldots P_m$, P_i and P_j cohere.

(c) In (a) and (b), the degree of coherence is inversely proportional to the number of propositions $P_1 \ldots P_m$.

Principle 3. Analogy.

(a) If P_1 explains Q_1, P_2 explains Q_2, P_1 is analogous to P_2, and Q_1 is analogous to Q_2, then P_1 and P_2 cohere, and Q_1 and Q_2 cohere.

(b) If P_1 explains Q_1, P_2 explains Q_2, Q_1 is analogous to Q_2, but P_1 is disanalogous to P_2, then P_1 and P_2 incohere.

Principle 4. Data Priority.

Propositions that describe the results of observation have a degree of acceptability on their own.

Principle 5. Contradiction.

If P contradicts Q, then P and Q incohere.

Principle 6. Acceptability.

(a) The acceptability of a proposition P in a system S depends on its coherence with the proposition in S.

(b) If many results of relevant experimental observations are unexplained, then the acceptability of a proposition P that explains only a few of them is reduced.

Principle 7. System Coherence.

The global explanatory coherence of a system S of propositions is a function of the pairwise local coherence of those propositions.

2.3. Discussion of the principles. Principle 1, Symmetry, asserts that pairwise coherence and incoherence are symmetric relations, in keeping with the everyday sense of coherence as holding together. The coherence of two propositions is thus very different from the nonsymmetric relations of entailment and conditional probability. Typically, P entails Q without Q entailing P, and the conditional probability of P given Q is different from the probability of Q given P. But if P and Q hold together, so do Q and P. The use of a symmetrical relation has advantages that will become clearer in the discussion of the connectionist implementation below.

Principle 2, Explanation, is by far the most important for assessing explanatory coherence, because it establishes most of the coherence relations. Part (a) is the most obvious: If a hypothesis P is part of the explanation of a piece of evidence Q, then P and Q cohere. Moreover, if a hypothesis P_2 is explained by another hypothesis P_1, then P_1 and P_2 cohere. Part (a) presupposes that explanation is a more restrictive relation than deductive implication, because otherwise we could prove that any two propositions cohere; for unless we use a relevance logic (Anderson & Belnap 1975), P_1 and the contradiction P_2 & not-P_2 imply any Q, so it would follow that P_1 coheres with Q. It follows from Principle 2(a), in conjunction with Principle 6, that the more a hypothesis explains, the more coherent and hence acceptable it is. Thus, this principle subsumes the criterion of explanatory breadth (which Whewell, 1967, called "consilience") that I have elsewhere claimed to be the most important for selecting the best explanation (Thagard 1978; 1988a).

Whereas part (a) of Principle 2 says that what explains coheres with what is explained, part (b) states that two propositions cohere if together they provide an explanation. Behind part (b) is the Duhem-Quine idea that the evaluation of a hypothesis depends partly on the other hypotheses with which it furnishes explanations (Duhem 1954; Quine 1961; see section 10.1). I call two hypotheses that are used together in an explanation "cohypotheses." Again I assume that explanation is more restrictive than implication; otherwise it would follow that any proposition that explained something was coherent with every other proposition, because if P_1 implies Q, then so does P_1 & P_2. But any scientist who maintained at a conference that the theory of general relativity and today's baseball scores together explain the motion of planets would be laughed off the podium. Principle 2 is intended to apply to explanations and hypotheses actually proposed by scientists.

Part (c) of Principle 2 embodies the claim that if numerous propositions are needed to furnish an explanation, then the coherence of the explaining propositions with each other and with what is explained is thereby diminished. Scientists tend to be skeptical of hypotheses that require myriad *ad hoc* assumptions in their explanations. There is nothing wrong in principle in having explanations that draw on many assumptions, but we should prefer theories that generate explanations using a unified core of hypotheses. I have elsewhere contended that the notion of *simplicity* most appropriate for scientific theory choice is a comparative one preferring theories that make fewer special assumptions (Thagard 1978; 1988a). Principles 2(b) and 2(c) together subsume this criterion. I shall not attempt further to characterize "degree of coherence" here, but the connectionist algorithm described below provides a natural interpretation. Many other notions of simplicity have been proposed (e.g., Foster & Martin 1966; Harman et al. 1988), but none is so directly relevant to considerations of explanatory coherence as the one embodied in Principle 2.

The third criterion for the best explanation in my earlier account was analogy, and this is subsumed in Principle 3. There is controversy about whether analogy is of more than heuristic use, but scientists such as Darwin have used analogies to defend their theories; his argument for evolution by natural selection is analyzed below. Principle 3(a) does not say simply that any two analogous propositions cohere. There must be an explanatory analogy, with two analogous propositions occurring in explanations of two other propositions that are analogous to each other. Recent computational models of analogical mapping and retrieval show how such correspondences can be noticed (Holyoak & Thagard, in press; Thagard et al. 1989). Principle 3(b) says that when similar phenomena are explained by dissimilar hypotheses, the hypotheses incohere. Although the use of such disanalogies is not as common as the use of analogies, it was important in the reasoning that led Einstein (1952) to the special theory of relativity: He was bothered by asymmetries in the way Maxwell's electrodynamics treated the case of (1) a magnet in motion and a conductor at rest quite differently from the case of (2) a magnet at rest and a conductor in motion.

Principle 4, Data Priority, stands much in need of elucidation and defense. In saying that a proposition describing the results of observation has a degree of acceptability on its own, I am not suggesting that it is indubitable, but only that it can stand on its own more successfully than can a hypothesis whose sole justification

is what it explains. A proposition Q may have some independent acceptability and still end up not accepted, if it is only coherent with propositions that are themselves not acceptable.

From the point of view of explanatory coherence alone, we should not take propositions based on observation as independently acceptable without any explanatory relations to other propositions. As BonJour (1985) argues, the coherence of such propositions is of a nonexplanatory kind, based on background knowledge that observations of certain sorts are very likely to be true. From past experience, we know that our observations are very likely to be true, so we should believe them unless there is substantial reason not to. Similarly, at a very different level, we have some confidence in the reliability of descriptions of experimental results in carefully refereed scientific journals. Section 10.4 relates the question of data priority to current philosophical disputes about justification.

Principle 5, Contradiction, is straightforward. By "contradictory" here I mean not just syntactic contradictions like P & not-P, but also semantic contradictions such as "This ball is black all over" and "This ball is white all over." In scientific cases, contradiction becomes important when incompatible hypotheses compete to explain the same evidence. Not all competing hypotheses incohere, however, because many phenomena have multiple causes. For example, explanations of why someone has certain medical symptoms may involve hypotheses that the patient has various diseases, and it is possible that more than one disease is present. Competing hypotheses incohere if they are contradictory or if they are framed as offering *the* most likely cause of a phenomenon. In the latter case, we get a kind of pragmatic contradictoriness: Two hypotheses may not be syntactically or semantically contradictory, yet scientists will view them as contradictory because of background beliefs suggesting that only one of the hypotheses is acceptable. For example, in the debate over dinosaur extinction (Thagard 1988b), scientists generally treat as contradictory the following hypotheses:

(1) Dinosaurs became extinct because of a meteorite collision.

(2) Dinosaurs became extinct because the sea level fell.

Logically, (1) and (2) could both be true, but scientists treat them as conflicting explanations, possibly because there are no explanatory relations between them and their conjunction is unlikely.

The relation "cohere" is not transitive. If P_1 and P_2 together explain Q, while P_1 and P_3 together explain not-Q, then P_1 coheres with both Q and not-Q, which incohere. Such cases do occur in science. Let P_1 be the gas law that volume is proportional to temperature, P_2 a proposition describing the drop in temperature of a particular sample of gas, P_3 a proposition describing the rise in temperature of the sample, and Q a proposition about increases in the sample's volume. Then P_1 and P_2 together explain a decrease in the volume, while P_1 and P_3 explain an increase.

Principle 6, Acceptability, proposes in part (a) that we can make sense of the overall coherence of a proposition in an explanatory system just from the pairwise coherence relations established by Principles 1–5. If we have a hypothesis P that coheres with evidence Q by virtue of explaining it, but incoheres with another contradictory hypothesis, should we accept P? To decide, we cannot merely count the number of propositions with which P coheres and incoheres, because the acceptability of P depends in part on the acceptability of those propositions themselves. We need a dynamic and parallel method of deriving general coherence from particular coherence relations; such a method is provided by the connectionist program described below.

Principle 6(b), reducing the acceptability of a hypothesis when much of the relevant evidence is unexplained by any hypothesis, is intended to handle cases where the best available hypothesis is still not very good, in that it accounts for only a fraction of the available evidence. Consider, for example, a theory in economics that could explain the stock market crashes of 1929 and 1987 but that had nothing to say about myriad other similar economic events. Even if the theory gave the best available account of the two crashes, we would not be willing to elevate it to an accepted part of general economic theory. What does "relevant" mean here? [See BBS multiple book review of Sperber & Wilson's *Relevance*, BBS 10(4) 1987.] As a first approximation, we can say that a piece of evidence is *directly* relevant to a hypothesis if the evidence is explained by it or by one of its competitors. We can then add that a piece of evidence is relevant if it is directly relevant or if it is similar to evidence that is relevant, where similarity is a matter of dealing with phenomena of the same kind. Thus, a theory of the business cycle that applies to the stock market crashes of 1929 and 1987 should also have something to say about nineteenth-century crashes and major business downturns in the twentieth century.

The final principle, System Coherence, proposes that we can have some global measure of the coherence of a whole system of propositions. Principles 1–5 imply that, other things being equal, a system S will tend to have more global coherence than another if

(1) S has more data in it;

(2) S has more internal explanatory links between propositions that cohere because of explanations and analogies; and

(3) S succeeds in separating coherent subsystems of propositions from conflicting subsystems.

The connectionist algorithm described below comes with a natural measure of global system coherence. It also indicates how different priorities can be given to the different principles.

3. Connectionist models

To introduce connectionist techniques, I shall briefly describe the popular example of how a network can be used to understand the Necker cube phenomenon (see, for example, Feldman & Ballard 1982; Rumelhart et al. 1986). Figure 1 contains a reversing cube: By changing our focus of attention, we are able to see as the front either face ABCD or face EFGH. The cube is perceived holistically, in that we are incapable of seeing corner A at the front without seeing corners B, C, and D at the front as well.

We can easily construct a simple network with the

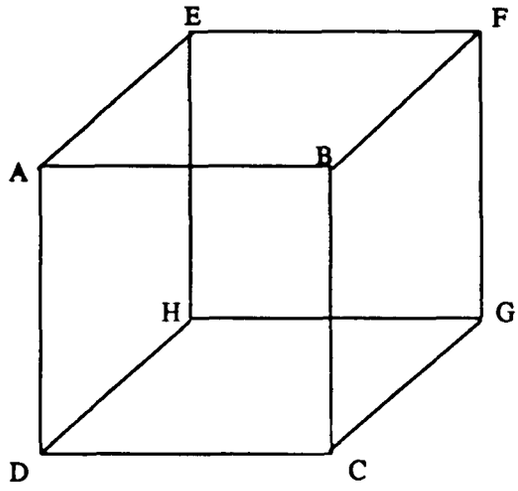

Figure 1. The Necker cube. Either ABCD or EFGH can be perceived as the front.

desired holistic property using *units*, crudely analogous to neurons, connected by links. Let Af be a unit that represents the hypothesis that corner A is at the front, while Ab represents the hypothesis that corner A is at the back. Similarly, we construct units Bf, Bb, Cf, Cb, Df, Db, Ef, Eb, Ff, Fb, Gf, Gb, Hf, and Hb. These units are not independent of each other. To signify that A cannot be both at the front and at the back, we construct an *inhibitory* link between the units Af and Ab, with similar links inhibiting Bf and Bb, and so on. Because corners A, B, C, and D go together, we construct *excitatory* links between each pair of Af, Bf, Cf, and Df, and between each pair of Ab, Bb, Cb, and Db. Analogous inhibitory and excitatory links are then set up for E, F, G, and H. In addition, we need inhibitory links between Af and Ef, Bf and Ff, and so on. Part of the resulting network is depicted in Figure 2. I have used solid lines to indicate excitatory links, and dotted lines to indicate inhibitory links.

Units can have varying degrees of *activation*. Suppose that our attention is focused on corner A, which we assume to be at the front, so that unit Af is activated. Then by virtue of the excitatory links from Af to Bf, Cf, and Df, these units will be activated. The inhibitory links from Af to Ab and Ef will cause those units to be deactivated. In turn, the excitatory links from Ab to Bb, Cb, and Db will deactivate them. Thus activation will spread through the network until all the units corresponding to the view that A, B, C, and D are at the front are activated, while all the units corresponding to the view that E, F, G, and H are at the front are deactivated.

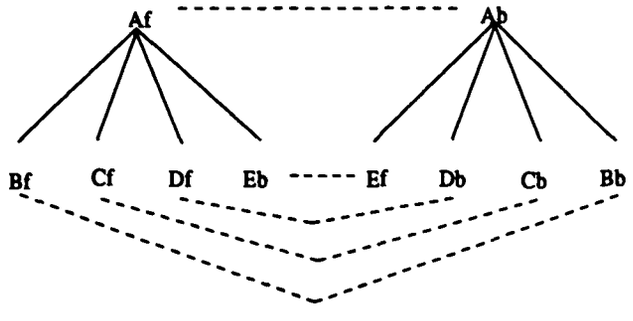

Figure 2. A connectionist network for interpreting the cube. Af is a unit representing the hypothesis that A is at the front, whereas Ab represents the hypothesis that A is at the back. Solid lines represent excitatory links; dotted lines represent inhibitory links.

Goldman has pointed out some of the attractive epistemological properties of this sort of network (Goldman 1986, Chap. 15; see also Thagard, in press a). A proposition, represented by a unit, is accepted if it is part of the best competing coalition of units and its rivals are rejected. Uncertainty consists in the absence of a clear-cut winner. Goldman argues that the connectionist view that has units representing propositions settling into either on or off states is more psychologically plausible and epistemologically appealing than the Bayesian picture that assigns probabilities to propositions.

4. ECHO

4.1. The program. Let us now look at ECHO, a computer program written in Common LISP that is a straightforward application of connectionist algorithms to the problem of explanatory coherence. In ECHO, propositions representing hypotheses and results of observation are represented by units. Whenever Principles 1–5 state that two propositions cohere, an excitatory link between them is established. If two propositions incohere, an inhibitory link between them is established. In ECHO, these links are symmetric, as Principle 1 suggests: The weight from unit 1 to unit 2 is the same as the weight from unit 2 to unit 1. Principle 2(c) says that the larger the number of propositions used in an explanation, the smaller the degree of coherence between each pair of propositions. ECHO therefore counts the propositions that do the explaining and proportionately lowers the weight of the excitatory links between units representing coherent propositions.

Principle 4, Data Priority, is implemented by links to each data unit from a special evidence unit that always has activation 1, giving each unit some acceptability on its own. When the network is run, activation spreads from the special unit to the data units, and then to the units representing explanatory hypotheses. The extent of data priority – the presumed acceptability of data propositions – depends on the weight of the link between the special unit and the data units. The higher this weight, the more immune the data units become to deactivation by other units. Units that have inhibitory links between them because they represent contradictory hypotheses have to compete with each other for the activation spreading from the data units: The activation of one of these units will tend to suppress the activation of the other. Excitatory links have positive weights; best performance occurs with weights around .05. Inhibitory links have negative weights; best performance occurs with weights around −.2. The activation of units ranges between 1 and −1; positive activation can be interpreted as acceptance of the proposition represented by the unit, negative activation as rejection, and activation close to 0 as neutrality. The relation between acceptability and probability is discussed in section 10.2.

To summarize how ECHO implements the principles of explanatory coherence, we can list key terms from the principles with the corresponding terms from ECHO:

Proposition: unit
Coherence: excitatory link, with positive weight
Incoherence: inhibitory link, with negative weight
Data priority: excitatory link from special unit
Acceptability: activation
System coherence: See the function H defined in section 4.9 below.

The following are some examples of the LISP formulas that constitute ECHO's inputs (I omit LISP quote symbols; see Tables 1–4 for actual input):

1. (EXPLAIN (H1 H2) E1)
2. (EXPLAIN (H1 H2 H3) E2)
3. (ANALOGOUS (H5 H6) (E5 E6))
4. (DATA (E1 E2 E5 E6))
5. (CONTRADICT H1 H4)

Formula 1 says that hypotheses H1 and H2 together explain evidence E1. As suggested by the second principle of explanatory coherence proposed above, formula 1 sets up three excitatory links, between units representing H1 and E1, H2 and E1, and H1 and H2.[1] Formula 2 sets up six such links, between each of the hypotheses and the evidence, and between each pair of hypotheses, but the weight on the links will be less than those established by formula 1, because there are more cohypotheses. In accord with Principle 3(a), Analogy, formula 3 produces excitatory links between H5 and H6, and between E5 and E6, if previous input has established that H5 explains E5 and H6 explains E6. Formula 4 is used to apply Principle 4, Data Priority, setting up explanation-independent excitatory links to each data unit from a special evidence unit. Finally, formula 5 sets up an inhibitory link between the contradictory hypotheses H1 and H4, as prescribed by Principle 5. A full specification of ECHO's inputs and algorithms is provided in the Appendix.

Input to ECHO can optionally reflect the fact that not all data and explanations are of equal merit. For example, a data statement can have the form

(DATA (E1 (E 2.8))).

This formula sets up the standard link from the special unit to E1, but interprets the ".8" as indicating that E2 is not as reliable a piece of evidence as E1. Hence, the weight from the special unit to E2 is only .8 as strong as the weight from the special unit to E1. Similarly, explain statements take an optimal numerical parameter, as in

(EXPLAIN (H1) E 1.9).

The additional parameter, .9, indicates some weakness in the quality of the explanation and results in a lower than standard weight on the excitatory link between H1 and E1. In ECHO's applications to date, the additional parameters for data and explanation quality have not been used, because it is difficult to establish them objectively from the texts we have been using to generate ECHO's inputs. But it is important that ECHO has the capacity to make use of judgments of data and explanation quality when these are available.

Program runs show that the networks thus established have numerous desirable properties. Other things being equal, activation accrues to units corresponding to hypotheses that explain more, provide simpler explanations, and are analogous to other explanatory hypotheses. The considerations of explanatory breadth, simplicity, and analogy are smoothly integrated. The networks are holistic, in that the activation of every unit can potentially have an effect on every other unit linked to it by a path, however lengthy. Nevertheless, the activation of a unit is directly affected only by those units to which it is linked. Although complexes of coherent propositions are evaluated together, different hypotheses in a complex can finish with different activations, depending on their particular coherence relations. The symmetry of excitatory links means that active units tend to bring up the activation of units with which they are linked, whereas units whose activation sinks below 0 tend to bring down the activation of units to which they are linked. Data units are given priority, but can nevertheless be deactivated if they are linked to units that become deactivated. So long as excitation is not set too high (see section 12.2), the networks set up by ECHO are stable: In most of them, all units reach asymptotic activation levels after fewer than 100 cycles of updating. The most complex network implemented so far, comparing the explanatory power of Copernicus's heliocentric theory with Ptolemy's geocentric one, requires about 210 cycles before its more than 150 units have all settled. To illustrate ECHO's capabilities, I shall describe some very simple tests that illustrate its ability to handle considerations of explanatory breadth, simplicity, and analogy. Later sections on scientific and legal reasoning provide more complex and realistic examples.

4.2. Explanatory breadth. We should normally prefer a hypothesis that explains more than alternative hypotheses. If hypothesis H1 explains two pieces of evidence, whereas H2 explains only one, then H1 should be preferred to H2. Here are four formulas given together to ECHO as input:

(EXPLAIN (H1) E1)
(EXPLAIN (H1) E2)
(EXPLAIN (H2) E2)
(CONTRADICT (H1 H2))
(DATA (E1 E2))

These formulas generate the network pictured in Figure 3, with excitatory links corresponding to coherence represented by solid lines, and with inhibitory links corresponding to incoherence represented by dotted lines.

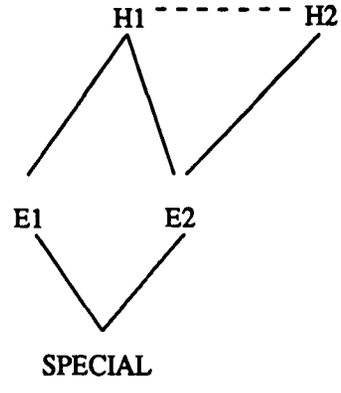

Figure 3. Explanatory breadth. As in Figure 2, solid lines represent excitatory links, whereas dotted line represents inhibitory links. Evidence units E1 and E2 are linked to the special unit. The result of running this network is that H1 defeats H2.

Activation flows from the special unit, whose activation is clamped at 1, to the evidence units, and then to the hypothesis units, which inhibit each other. Because H1 explains more than its competitor H2, H1 becomes active, settling with activation above 0, while H2 is deactivated, settling with activation below 0. (See section 4.10 for a discussion of the parameters that affect the runs, and the Appendix for sensitivity analyses.) Notice that although the links in ECHO are symmetric, in keeping with the symmetry of the coherence relation, the flow of activation is not, because evidence units get activation first and then pass it along to what explains them.

ECHO's networks have interesting dynamic properties. What happens if new data come in after the network has settled? When ECHO is given the further information that H2 explains additional data E3, E4, and E5, then the network resettles into a reversed state in which H2 is activated and H1 is deactivated. However, if the additional information is only that H2 explains E2, or only that H2 explains E3, then ECHO does not resettle into a state in which H1 and H2 get equal activation. (It does give H1 and H2 equal activation if the input says that they have equal explanatory power from the start.) Thus ECHO displays a kind of conservatism also seen in human scientists. See the discussion of conservatism in section 10.4.

4.3. Being explained. Section 4.2 showed how Principle 2(a) leads ECHO to prefer a hypothesis that explains more than its competitors. The same principle also implies greater coherence, other things being equal, for a hypothesis that is explained. Consider the following input:

(EXPLAIN (H1) E1)
(EXPLAIN (H1) E2)
(EXPLAIN (H2) E1)
(EXPLAIN (H2) E2)
(EXPLAIN (H3) H1)
(CONTRADICT H1 H2)
(DATA (E1 E2))

Figure 4 depicts the network constructed using this input. Here, and in all subsequent figures, the special evidence unit is not shown. In Figure 4, H1 and H2 have the same explanatory breadth, but ECHO activates H1 and deactivates H2 because H1 is explained by H3. ECHO thus gives more activation to a hypothesis that is explained than to a contradictory one that is not explained. If the above formulas did not include a CONTRADICT statement, then no inhibitory links would be formed, so that all units would asymptote with positive activation. Because of the decay parameter, activation is still less than 1: See the equations in the Appendix.

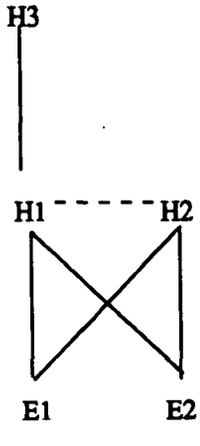

Figure 4. Being explained. H1 defeats H2 because it is explained by H3.

4.4. Refutation. According to Popper (1959), the hallmark of science is not the acceptance of explanatory theories but the rejection of falsified ones. Take the simplest case where a hypothesis H1 explains (predicts) some piece of "negative evidence" NE1, which contradicts data E1. Then E1 becomes active, deactivating NE1 and hence H1. Such straightforward refutations, however, are rare in science. Scientists do not typically give up a promising theory just because it has some empirical problems, and neither does ECHO. If in addition to explaining NE1, H1 explains some positive pieces of evidence, E2 and E3, then ECHO does not deactivate it. However, an alternative hypothesis H2 that also explains E2 and E3 is preferred to H1, which loses because of NE1. Rejection in science is usually a complex process involving competing hypothesis, not a simple matter of falsification (Lakatos 1970; Thagard 1988a, Chap. 9; section 10.1 below).

4.5. Unification. The impact of explanatory breadth, being explained, and refutation all arise from Principle 2(a), which says that hypotheses cohere with what they explain. According to Principle 2(b), cohypotheses that explain together cohere with each other. Thus, if H1 and H2 together explain evidence E, then H1 and H2 are linked. This gives ECHO a preference for unified explanations, ones that use a common set of hypotheses rather than having special hypotheses for each piece of evidence explained. Consider this input, which generates the network shown in Figure 5:

(EXPLAIN (H1 A1) E1)
(EXPLAIN (H1 A2) E2)
(EXPLAIN (H2 A3) E1)
(EXPLAIN (H2 A3) E2)
(CONTRADICT H1 H2)
(DATA (E1 E2))

Although H1 and H2 both explain E1 and E2, the explanation by H2 is more unified in that it uses A3 in both cases. Hence ECHO forms a stronger link between H2 and A3 than it does between H1 and A1 or A2, so H2 becomes activated and H1 is deactivated. The explanations by H2 are not simpler than those by H1, in the sense

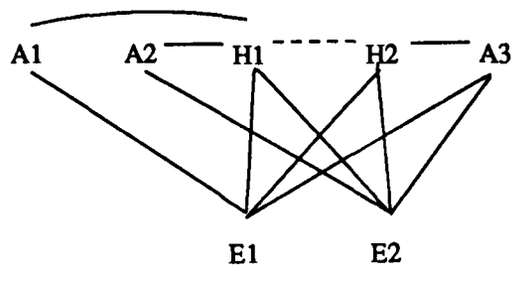

Figure 5. Unification. H2 defeats H1 because it gives a more unified explanation of the evidence.

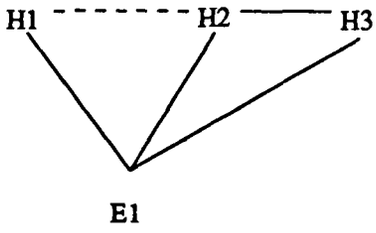

Figure 6. Simplicity. H1 defeats H2 because it gives a simpler explanation of the evidence.

of Principle 2(c), because both involve two hypotheses. ECHO's preference for H2 over H1 thus depends on the coherence of H2 with its auxiliary hypothesis and the evidence being greater than the coherence of H1 with its auxiliary hypotheses and the evidence. One might argue that the coherence between cohypotheses should be less than the coherence of a hypothesis with what it explains; ECHO contains a parameter that can allow the weights between cohypothesis units to be less than the weight between a hypothesis unit and an evidence unit.

4.6. Simplicity. According to Principle 2(c), the degree of coherence of a hypothesis with what it explains and with its cohypotheses is inversely proportional to the number of cohypotheses. An example of ECHO's preference for simple hypotheses derives from the input:

(EXPLAIN (H1) E1)
(EXPLAIN (H2 H3) E1)
(CONTRADICT H1 H2)
(DATA (E1))

Here H1 is preferred to H2 and H3 because it accomplishes the explanation with no cohypotheses. The generated network is shown in Figure 6.

Principle 2(c) is important for dealing with *ad hoc* hypotheses that are introduced only to save a hypothesis from refutation. Suppose that H1 is in danger of refutation because it explains negative evidence NE1, which contradicts evidence E1. One might try to save H1 by concocting an auxiliary hypothesis, H2, which together with H1 would explain E1. Such maneuvers are common in science: Nineteenth-century physicists did not abandon Newtonian mechanics because it gave false predictions concerning the motion of Uranus; instead, they hypothesized the existence of another planet, Neptune, to explain the discrepancies. Neptune, of course, was eventually observed, but we need to be able to discount auxiliary hypotheses that do not contribute to any additional explanations. Because the explanation of E1 by H1 and H2 is less simple than the explanation of NE1 by H1, the *ad hoc* maneuver does not succeed in saving H1 from deactivation.

4.7. Analogy. According to Principle 3(a), analogous hypotheses that explain analogous evidence are coherent with each other. Figure 7 shows relations of analogy, derived from the input:

(EXPLAIN (H1) E1)
(EXPLAIN (H2) E1)
(EXPLAIN (H3) E3)
(ANALOGOUS (H2 H3) (E1 E3))
(CONTRADICT H1 H2)
(DATA (E1 E3))

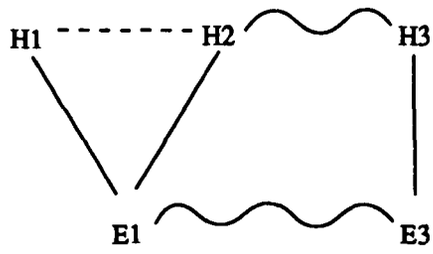

Figure 7. Analogy. The wavy lines indicate excitatory links based on analogies. H2 defeats H1 because the explanation it gives is analogous to the explanation afforded by H3.

The analogical links corresponding to the coherence relations required by Principle 3 are shown by wavy lines. Running this example leads to activation of H2 and deactivation of its rival, H1. Figures 3–7 show consilience, simplicity, and analogy operating independently of each other, but in realistic examples these criteria can all operate simultaneously through activation adjustment. Thus ECHO shows how criteria such as explanatory breadth, simplicity, and analogy can be integrated. My most recent account of inference to the best explanation (Thagard 1988a) included a computational model that integrated breadth and simplicity but left open the question of how to tie in analogy. Principle 3 and ECHO show how analogy can participate with consilience and simplicity in contributing toward explanatory power.

4.8. Evidence. Principle 4 asserts that data get priority by virtue of their independent coherence. But it should nevertheless be possible for a data unit to be deactivated. We see this both in the everyday practice of experimenters, in which it is often necessary to discard some of the data because they are deemed unreliable (Hedges 1987), and in the history of science where evidence for a discarded theory sometimes falls into neglect (Laudan 1976). Figure 8, which derives from the following input, shows how this might happen.

(EXPLAIN (H1) E1)
(EXPLAIN (H2) E2)
(EXPLAIN (H1) E3)
(EXPLAIN (H1) E4)
(EXPLAIN (H2) E2)
(EXPLAIN (H2) E5)
(EXPLAIN (H3) E3)
(EXPLAIN (H3) E5)
(EXPLAIN (H4) E4)
(EXPLAIN (H4) E5)
(CONTRADICT H1 H2)
(CONTRADICT H1 H3)
(CONTRADICT H1 H4)

These inputs lead to the deactivation of E5, dragged down by the deactivation of the inferior hypotheses H3, H4, and H5. Because E5 coheres only with propositions that are themselves unacceptable, it becomes unacceptable too. Because H1 has four excitatory links, it easily deactivates the other three hypotheses, and their negative activation brings down the initially positive activation of E5 into the negative range.

Principle 6(b) also concerns evidence, undermining the acceptability of hypotheses that explain only a small part of the relevant data. Accordingly, ECHO automati-

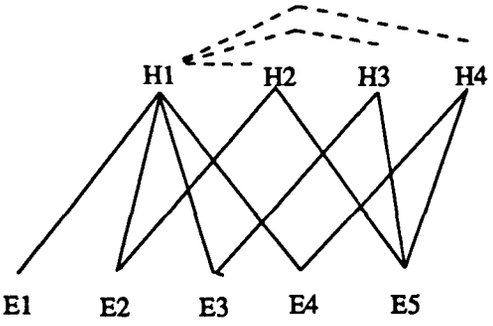

Figure 8. Downplaying of evidence. E5 is deactivated, even though it is an evidence unit, because it coheres only with inferior hypotheses.

cally increases the value of a decay parameter in proportion to the ratio of unexplained evidence to explained evidence (see Appendix). A hypothesis that explains only a fraction of the relevant evidence will thus decay toward the beginning activation level of 0 rather than become activated.

4.9. Acceptability and System Coherence. If ECHO is taken as an algorithmic implementation of the first five principles of explanatory coherence, then it validates Principle 6, Acceptability, for it shows that holistic judgments of the acceptability of a proposition can be based solely on pairwise relations of coherence. A unit achieves a stable activation level merely by considering the activation of units to which it is linked and the weights on those links. Asymptotic activation values greater than 0 signify acceptance of the proposition represented by the unit, whereas negative values signify rejection.

ECHO also validates Principle 7, System Coherence, because we can borrow from connectionist models a measure H of the global coherence of a whole system of propositions at time t:

$$H(t) = \Sigma_i \Sigma_j w_{ij} a_i(t) a_j(t) \quad (1)$$

In this equation, w_{ij} is the weight from unit i to unit j, and $a_i(t)$ is the activation of unit i at time t. This measure or its inverse has been variously called the "goodness," "energy," or "harmony" of the network (Rumelhart et al. 1986, vol. 2, p. 13). For historical reasons, I prefer a variant of the last term with the alternative spelling "harmany" (Harman 1973). Thus ECHO stands for "Explanatory Coherence by Harmany Optimization."

Equation 1 says that to calculate the harmany of the network, we consider each pair of units a_i and a_j that are linked with weight w_{ij}. Harmany increases, for example, when two units with high activation have a link between them with high weight, or when a unit with high activation and a unit with negative activation have between them a link with negative weight. In ECHO the harmany of a system of propositions increases, other things being equal, with increases in the number of data units, the number of links, and the number of cycles to update activations to bring them more in line with the weights.

4.10. Parameters. The simulations just described depend on program parameters that give ECHO numerous degrees of freedom, some of which are epistemologically interesting. In the example in section 4.2 (Figure 3), the relation between excitatory weights and inhibitory weights is crucial. If inhibition is low compared to excitation, then ECHO will activate both H1 and H2, because the excitation that H2 gets from E1 will overcome the inhibition it gets from H1. Let the *tolerance* of the system be the absolute value of the ratio of excitatory weight to inhibitory weight. With high tolerance, the system will entertain competing hypotheses. With low tolerance, winning hypotheses deactivate the losers. Typically, ECHO is run with excitatory weights set at .05 and inhibition at −.2, so tolerance is .25. If tolerance is high, ECHO can settle into a state where two contradictory hypotheses are both activated. ECHO performs well using a wide range of parameters (see the sensitivity analyses in the Appendix).

Other parameters establish the relative importance of simplicity and analogy. If H1 explains E1 by itself, then the excitatory link between H1 and E1 has the default weight .05. But if H1 and H2 together explain E1, then the weight of the links is set at the default value divided by 2, the number of cohypotheses, leaving it at .025. If we want to change the importance of simplicity as incorporated in Principle 2(c), however, then we can raise the number of cohypotheses to an exponent that represents the *simplicity impact* of the system. Equation 3 for doing this is given in the algorithm section of the Appendix. The greater the simplicity impact, the more weights will be diminished by having more cohypotheses. Similarly, the weights established by analogy can be affected by a factor representing *analogy impact*. If this is 1, then the links connecting analogous hypotheses are just as strong as those set up by simple explanations, and analogy can have a very large effect. If, on the other hand, analogy impact is set at 0, then analogy has no effect.

Another important parameter of the system is decay rate, represented by θ (see equation 4 in the Appendix). We can term this the *skepticism* of the system, because the higher it is, the more excitation from data will be needed to activate hypotheses. If skepticism is very high, then *no* hypotheses will be activated. Whereas tolerance reflects ECHO's view of contradictory hypotheses, skepticism determines its treatment of all hypotheses. Principle 6(b) can be interpreted as saying that if there is much unexplained evidence, then ECHO's skepticism level is raised.

Finally, we can vary the priority of the data by adjusting the weights to the data units from the special unit. *Data excitation* is a value from 0 to 1 that provides these weights. To reflect the scientific practice of not treating all data equally seriously, it is also possible to set the weights and initial activations for each data unit separately. If data excitation is set low, then, contrary to section 4.2, new evidence for a rejected hypothesis will not lead to its adoption. If data excitation is high, then, contrary to section 4.8, evidence that supports only a bad hypothesis will not be thrown out.

With so many degrees of freedom, which are typical of connectionist models, one might question the value of simulations, as it might seem that any desired behavior whatsoever could be obtained. However, if a fixed set of default parameters applies to a large range of cases, then the arbitrariness is much diminished. In *all* the computer runs reported in this paper, ECHO has had excitation at .05, inhibition at −.2 (so tolerance is .25), data excitation

at .1., decay (skepticism) at .05, simplicity impact at 1, and analogy impact at 1. As reported in the Appendix in the section on sensitivity analyses, there is nothing special about the default values of the parameters: ECHO works over a wide range of values. In a full simulation of a scientist's cognitive processes, we could imagine better values being *learned.* Many connectionist models do not take weights as given, but instead adjust them as the result of experience. Similarly, we can imagine that part of a scientist's training entails learning how seriously to take data, analogy, simplicity, and so on. Most scientists get their training not merely by reading and experimenting on their own but also by working closely with scientists already established in their field; hence, a scientist can pick up the relevant values from advisors. In ECHO they are set by the programmer, but it should be possible to extend the program to allow training from examples.

The examples described in this section are trivial and show merely that ECHO has some desired properties. I shall now show that ECHO can handle some much more substantial examples from the history of science and from recent legal deliberations.

5. Applications of ECHO to scientific reasoning

Theories in the philosophy of science, including computational ones, should be evaluated with respect to important cases from the history of science. To show the historical application of the theory of explanatory coherence, I shall discuss two important cases of arguments concerning the best explanation: Lavoisier's argument for his oxygen theory against the phlogiston theory, and Darwin's argument for evolution by natural selection. ECHO has also been applied to the following:

Contemporary debates about why the dinosaurs became extinct (Thagard 1988b);

Arguments by Wegener and his critics for and against continental drift (Thagard & Nowak 1988; in press);

Psychological experiments on how beginning students learn physics (Ranney & Thagard 1988); and

Copernicus's case against Ptolemaic astronomy (Nowak & Thagard, forthcoming).

Additional applications are currently under development.

5.1. Lavoisier. In the middle of the eighteenth century, the dominant theory in chemistry was the phlogiston theory of Stahl, which provided explanations of important phenomena of combustion, respiration, and calcination (what we would now call oxidation). According to the phlogiston theory, combustion takes place when phlogiston in burning bodies is given off. In the 1770s, Lavoisier developed the alternative theory that combustion takes place when burning bodies combine with oxygen from the air (for an outline of the conceptual development of his theory, see Thagard, in press b). More than ten years after he first suspected the inadequacy of the phlogiston theory, Lavoisier mounted a full-blown attack on it in a paper called "Réflexions sur le Phlogistique" (Lavoisier 1862).

Tables 1 and 2 present the input given to ECHO to represent Lavoisier's argument in his 1783 polemic against phlogiston. Table 1 shows the 8 propositions used to represent the evidence to be explained and the 12 used to represent the competing theories. The evidence concerns different properties of combustion and calcination, while there are two sets of hypotheses representing the

Table 1. *Input propositions for Lavoisier (1862) example*

Evidence	
(proposition 'E1	"In combustion, heat and light are given off.")
(proposition 'E2	"Inflammability is transmittable from one body to another.")
(proposition 'E3	"Combusion only occurs in the presence of pure air.")
(proposition 'E4	"Increase in weight of a burned body is exactly equal to weight of air absorbed.")
(proposition 'E5	"Metals undergo calcination.")
(proposition 'E6	"In calcination, bodies increase weight.")
(proposition 'E7	"In calcination, volume of air diminishes.")
(proposition 'E8	"In reduction, effervescence appears.")
Oxygen hypotheses	
(proposition 'OH1	"Pure air contains oxygen principle.")
(proposition 'OH2	"Pure air contains matter of fire and heat.")
(proposition 'OH3	"In combustion, oxygen from the air combines with the burning body.")
(proposition 'OH4	"Oxygen has weight.")
(proposition 'OH5	"In calcination, metals add oxygen to become calxes.")
(proposition 'OH6	"In reduction, oxygen is given off.")
Phlogiston hypotheses	
(proposition 'PH1	"Combustible bodies contain phlogiston.")
(proposition 'PH2	"Combustible bodies contain matter of heat.")
(proposition 'PH3	"In combustion, phlogiston is given off.")
(proposition 'PH4	"Phlogiston can pass from one body to another.")
(proposition 'PH5	"Metals contain phlogiston.")
(proposition 'PH6	"In calcination, phlogiston is given off.")

Table 2. *Input explanations and contradictions in Lavoisier (1862) example*

Oxygen explanations
(explain '(OH1 OH2 OH3) 'E1)
(explain '(OH1 OH3) 'E3)
(explain '(OH1 OH3 OH4) 'E4)
(explain '(OH1 OH5) 'E5)
(explain '(OH1 OH4 OH5) 'E6)
(explain '(OH1 OH5) 'E7)
(explain '(OH1 OH6) 'E8)

Phlogiston explanations
(explain '(PH1 PH2 PH3) 'E1)
(explain '(PH1 PH3 PH4) 'E2)
(explain '(PH5 PH6) 'E5)

Contradictions
(contradict 'PH3 'OH3)
(contradict 'PH6 'OH5)

Data
(data '(E1 E2 E3 E4 E5 E6 E7 E8))

oxygen and phlogiston theories, respectively. These propositions do not capture Lavoisier's argument completely but do recapitulate its major points. (In a slightly more complicated simulation not presented here, I have encoded the attempt by the phlogiston theory to explain the increase in weight in combustion and calcination by the supposition that phlogiston has negative weight; Lavoisier argues that this supposition renders the phlogiston theory internally contradictory, because phlogiston theorists sometimes assumed that phlogiston has positive weight.)

Table 2 shows the part of the input that sets up the network used to make a judgment of explanatory coherence. The "explain" statements are based directly on Lavoisier's own assertions about what is explained by the phlogiston theory and the oxygen theory. The "contradict" statements reflect my judgment of which of the oxygen hypotheses conflict directly with which of the phlogiston hypotheses.

These explanations and contradictions generate the network partially portrayed in Figure 9. Excitatory links, indicating that two propositions cohere, are represented by solid lines. Inhibitory links are represented by dotted lines. All the oxygen hypotheses are arranged along the top line and all the phlogiston hypotheses along the bottom, with the evidence in the middle. Omitted from the figure for the sake of legibility are the excitatory links among the hypotheses of the two theories and the links between the evidence units and the special unit. In addition to its displayed links to evidence, OH1 has excitatory links to OH2, OH3, OH4, OH5, and OH6. The link between OH1 and OH3 is particularly strong, because these two hypotheses participate in three explanations together. Figure 10, produced by a graphics program that runs with ECHO, displays the links to OH3, with excitatory links shown by thick lines and the inhibitory link with PH3 shown by a thin line. The numbers on the lines indicate the weights of the links rounded to three decimal places: In accord with Principle 2(c), weights are different from the default weight of .05 whenever multiple hypotheses are used in an explanation. If the hypotheses participate in only one explanation, then the weight between them is equal to the default excitation divided by the number of hypotheses; but weights are additive, so that the weight is increased if two hypotheses participate in more than one explanation. For example, the link between OH3 and E1 has the weight .017 (.0166666 rounded), because the explanation of E1 by OH3 required two additional hypotheses. The weight between OH3 and OH1 is .058 (.025 + .0166666 + .0166666), because the two of them alone explain E3, and together they explain E1 and E4 along with a third hypothesis in each case. OH1 and OH3 are thus highly coherent with each other by virtue of being used together in multiple explanations.

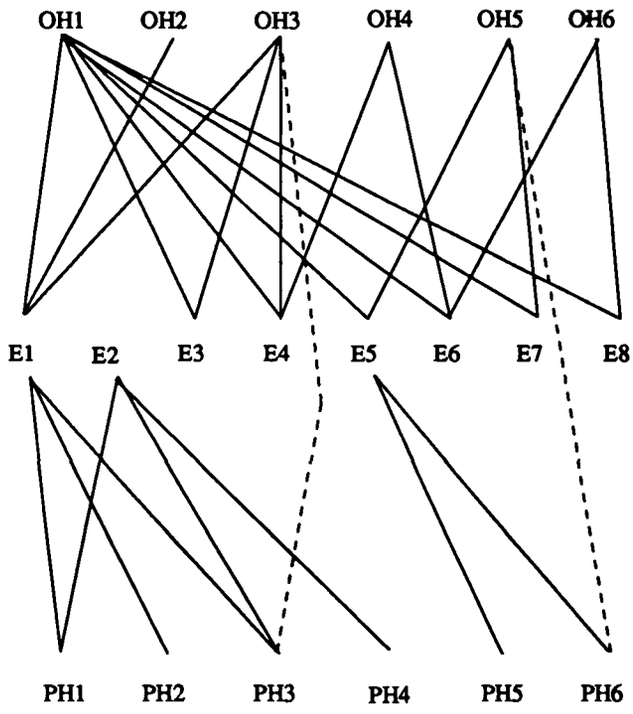

Figure 9. Network representing Lavoisier's (1862) argument. E1–E8 are evidence units. OH1–OH6 are units representing hypotheses of the oxygen theory; PH1–PH6 represent the phlogiston hypotheses. Solid lines are excitatory links; dotted lines are inhibitory.

The numbers beneath the names in Figure 10 indicate the final activation of the named units, rounded to three decimal places. When ECHO runs this network, starting with all hypotheses at activation .01, it quickly favors the oxygen hypotheses, giving them activations greater than 0. In contrast, all the phlogiston hypotheses become deactivated. The activation history of the propositions is shown in Figure 11, which charts activation as a function of the number of cycles of updating. Figure 11 shows graphs, produced automatically during the run of the program, of the activations of all the units over the 107 cycles it takes them to reach asymptote. In each graph, the horizontal line indicates the starting activation of 0 and the y axis shows activation values ranging between 1 and -1. Notice that the oxygen hypotheses OH1–OH6 rise steadily to their asymptotic activations, while PH3 and PH6, which directly contradict oxygen hypotheses, sink to activation levels well below 0. The other phlo-

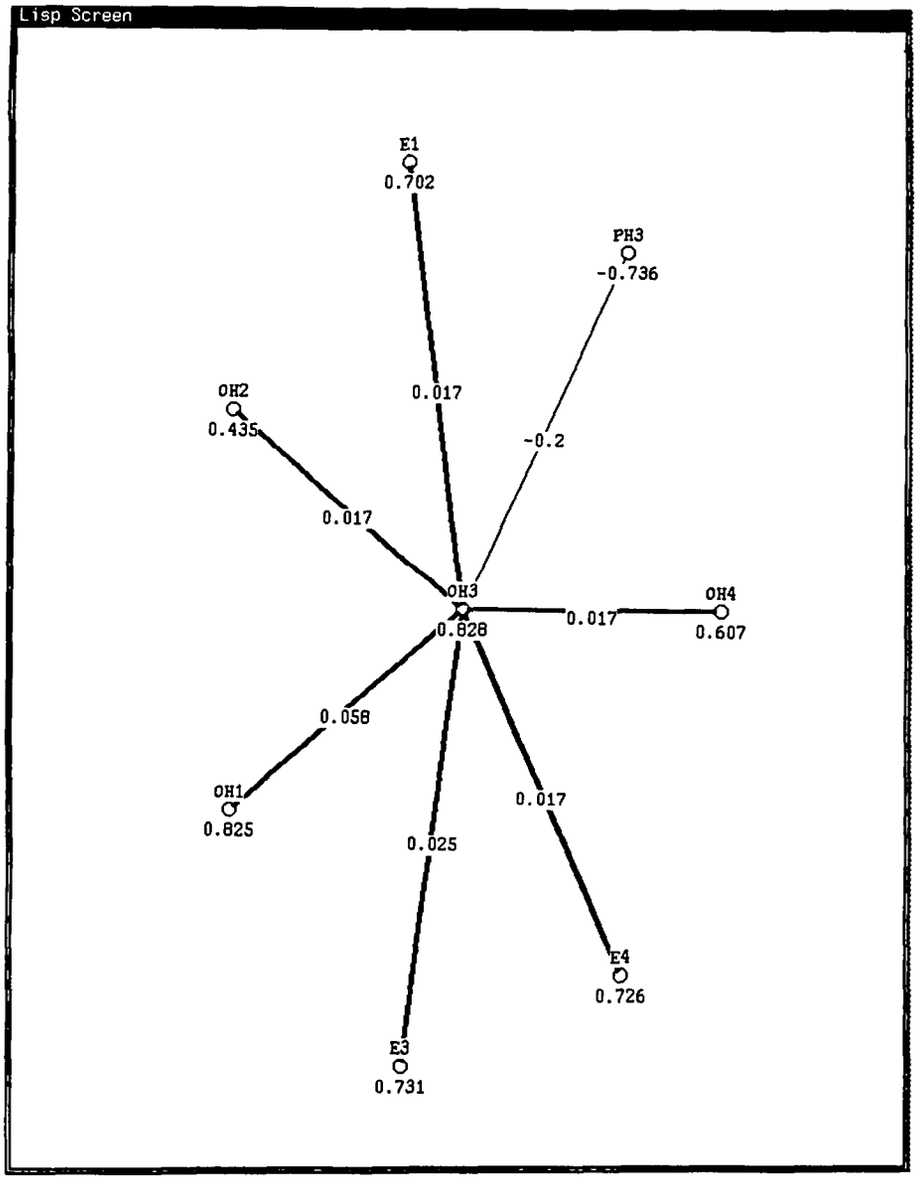

Figure 10. Connectivity of oxygen theory unit OH3. The numbers under the units are their activation values after the unit has settled. Thick lines indicate excitatory links; thin line indicates inhibitory link. Numbers on the lines indicate the weights on the links.

giston hypotheses that are not directly contradicted by oxygen hypotheses start out with positive activation but are dragged down toward 0 through their links with their deactivated cohypotheses. Thus the phlogiston theory fails as a whole.

This run of ECHO is biased towards the oxygen theory because it was based on an analysis of Lavoisier's argument. We would get a different network if ECHO were used to model critics of Lavoisier such as Kirwan (1789/1968), who defended a variant of the phlogiston theory. By the late 1790s, the vast majority of chemists and physicists, including Kirwan, had accepted Lavoisier's arguments and rejected the phlogiston theory, a turnaround contrary to the suggestion of Kuhn (1970) that scientific revolutions occur only when proponents of an old paradigm die off.

Lavoisier's argument represents a relatively simple application of ECHO, showing two sets of hypotheses competing to explain the evidence. But more complex explanatory relations can also be important. Sometimes a hypothesis that explains the evidence is itself explained by another hypothesis. Depending on the warrant for the higher-level hypothesis, this extra explanatory layer can increase acceptability: A hypothesis gains from being explained as well as by explaining the evidence. The Lavoisier example does not exhibit this kind of coherence, because neither Lavoisier nor the phlogiston theorists attempted to explain their hypotheses using higher-level hypotheses; nor does the example display the role that analogy can play in explanatory coherence.

5.2. Darwin. Both these aspects – coherence based on being explained and on analogy – were important in Darwin's argument for his theory of evolution by natural selection (Darwin 1962). His two most important hypotheses were:

DH2 – Organic beings undergo natural selection.

DH3 – Species of organic beings have evolved.

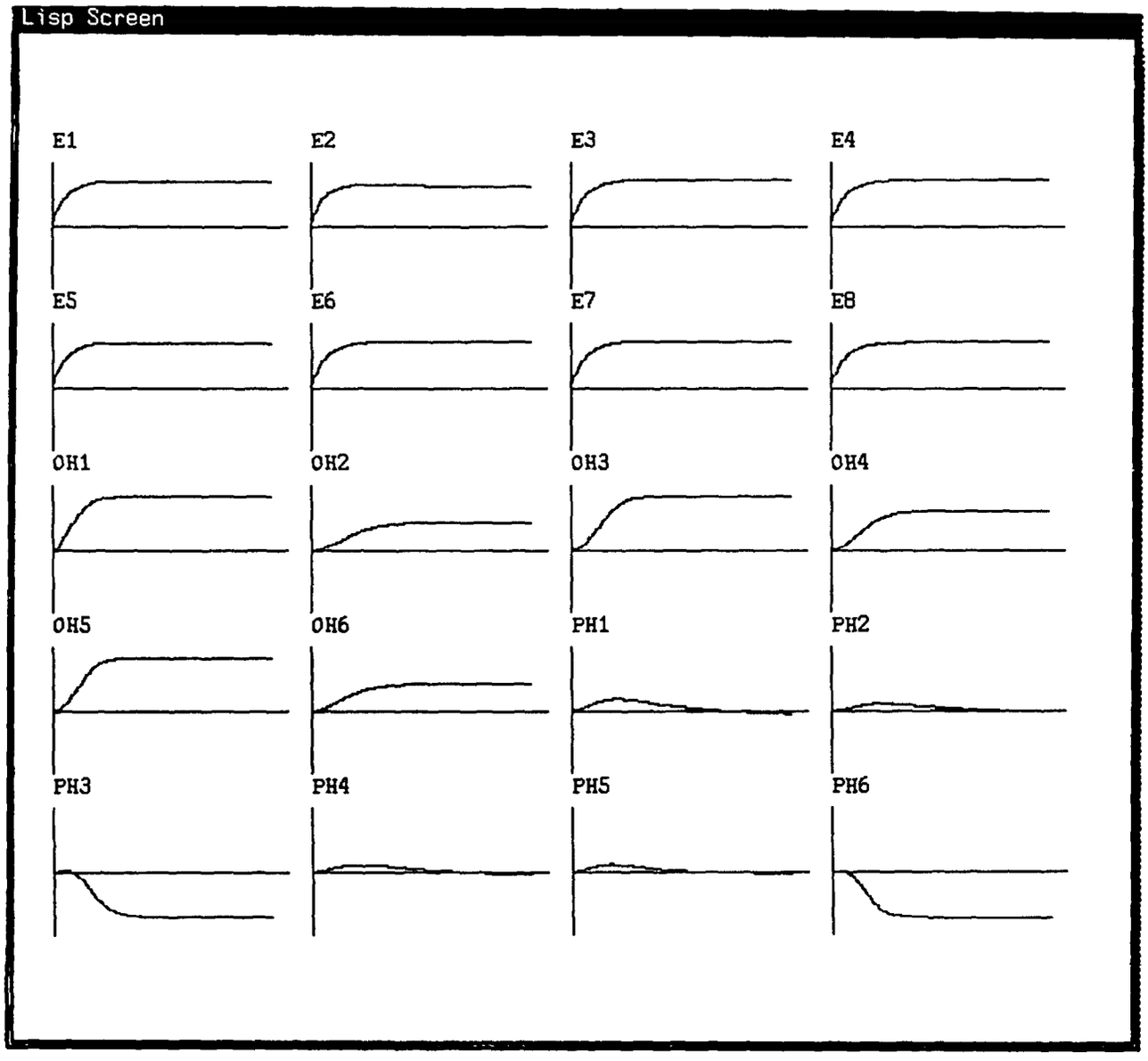

Figure 11. Activation history of Lavoisier (1862) network. Each graph shows the activation of a unit over 107 cycles of updating, on a scale of −1 to 1, with the horizontal line indicating the initial activation of 0.

These hypotheses together enabled him to explain a host of facts, from the geographical distribution of similar species to the existence of vestigial organs. Darwin's argument was explicitly comparative: There are numerous places in the *Origin* where he points to phenomena that his theory explains but that are inexplicable on the generally accepted rival hypothesis that species were separately created by God.

Darwin's two main hypotheses were not simply co-hypotheses, however, for he also used DH2 to explain DH3! That is, natural selection explains why species evolve: If populations of animals vary, and natural selection picks out those with features well adapted to particular environments, then new species will arise. Moreover, he offers a Malthusian explanation for why natural selection occurs as the result of the geometrical rate of population growth contrasted with the arithmetical rate of increase in land and food. Thus Malthusian principles explain why natural selection takes place, which explains why evolution occurs, and natural selection and evolution together explain a host of facts better than the competing creation hypothesis does.

The full picture is even more complicated than this, for Darwin frequently cites the analogy between artificial and natural selection as evidence for his theory. He contends that just as farmers are able to develop new breeds of domesticated animals, so natural selection has produced new species. He uses this analogy not simply to defend natural selection, but also to help in the explanations of the evidence: Particular explanations using natural selection incorporate the analogy with artificial selection. Finally, to complete the picture of explanatory coherence that the Darwin example offers, we must consider the alternative theological explanations that were accepted by even the best scientists before Darwin proposed his theory.

Analysis of *On the origin of species* suggests the 15 evidence statements shown in Table 3. Statements E1–E4 occur in Darwin's discussion of objections to his theory; the others are from the later chapters where he

Table 3. *Explanations and contradictions for Darwin (1962) example*

Darwin's evidence	
(proposition 'E1	"The fossil record contains few transitional forms.")
(proposition 'E2	"Animals have complex organs.")
(proposition 'E3	"Animals have instincts.")
(proposition 'E4	"Species when crossed become sterile.")
(proposition 'E5	"Species become extinct.")
(proposition 'E6	"Once extinct, species do not reappear.")
(proposition 'E7	"Forms of life change almost simultaneously around the world.")
(proposition 'E8	"Extinct species are similar to each other and to living forms.")
(proposition 'E9	"Barriers separate similar species.")
(proposition 'E10	"Related species are concentrated in the same areas.")
(proposition 'E11	"Oceanic islands have few inhabitants, often of peculiar species.")
(proposition 'E12	"Species show systematic affinities.")
(proposition 'E13	"Different species share similar morphology.")
(proposition 'E14	"The embryos of different species are similar.")
(proposition 'E15	"Animals have rudimentary and atrophied organs.")
Darwin's main hypotheses	
(proposition 'DH1	"Organic beings are in a struggle for existence.")
(proposition 'DH2	"Organic beings undergo natural selection.")
(proposition 'DH3	"Species of organic beings have evolved.")
Darwin's auxiliary hypotheses	
(Proposition 'DH4	"The geological record is very imperfect.")
(proposition 'DH5	"There are transitional forms of complex organs.")
(proposition 'DH6	"Mental qualities vary and are inherited.")
Darwin's facts	
(proposition 'DF1	"Domestic animals undergo variation.")
(proposition 'DF2	"Breeders select desired features of animals.")
(proposition 'DF3	"Domestic varieties are developed.")
(proposition 'DF4	"Organic beings in nature undergo variation.")
(proposition 'DF5	"Organic beings increase in population at a high rate.")
(proposition 'DF6	"The sustenance available to organic beings does not increase at a high rate.")
(proposition 'DF7	"Embryos of different domestic varieties are similar.")
Creationist hypothesis	
(proposition 'CH1	"Species were separately created by God.")

argues positively for his theory. Table 3 also shows Darwin's main hypotheses. DH2 and DH3 are the core of the theory of evolution by natural selection, providing explanations of its main evidence, E5–E15. DH4–DH6 are auxiliary hypotheses that Darwin uses in resisting objections based on E1–E3. He considers the objection concerning the absence of transitional forms to be particularly serious, but explains it away by saying that the geological record is so imperfect that we should not expect to find fossil evidence of the many intermediate species his theory requires. Darwin's explanations also use a variety of facts he defends with empirical arguments that would complicate the current picture too much to present here. Hence, I will treat them (DF1–DF7) simply as pieces of evidence that do not need explanatory support. The creationist opposition frequently mentioned by Darwin is represented by the single hypothesis that species were separately created by God.

Table 4 shows the explanation and contradiction statements that ECHO uses to set up its network, which is partially displayed in Figure 12. Notice the hierarchy of explanations, with the high rate of population increase explaining the struggle for existence, which explains natural selection, which explains evolution. Natural selection and evolution together explain many pieces of evidence. The final component of Darwin's argument is the analogy between natural and artificial selection. The wavy lines represent excitatory links based on analogy. Just as breeders' actions explain the development of domestic varieties, so natural selection explains the evolution of species. At another level, Darwin sees an embryological analogy. The embryos of different domestic varieties are quite similar to each other, which is explained by the fact that breeders do not select for properties of embryos. Similarly, nature does not select for most properties of embryos, which explains the many similarities between embryos of different species.

Darwin's discussion of objections suggests that he thought creationism could naturally explain the absence of transitional forms and the existence of complex organs and instincts. Darwin's argument was challenged in many ways, but based on his own view of the relevant explanatory relations, at least, the theory of evolution by natural selection is far more coherent than the creation hypoth-

Table 4. *Explanations and contradictions for Darwin example*

Darwin's explanations
(a) of natural selection and evolution
(explain '(DF5 DF6) 'DH1)
(explain '(DH1 DF4) 'DH2)
(explain '(DH2) 'DH3)

(b) of potential counterevidence
(explain '(DH2 DH3 DH4) 'E1)
(explain '(DH2 DH3 DH5) 'E2)
(explain '(DH2 DH3 DH6) 'E3)

(c) of diverse evidence
(explain '(DH2) 'E5)
(explain '(DH2 DH3) 'E6)
(explain '(DH2 DH3) 'E7)
(explain '(DH2 DH3) 'E8)
(explain '(DH2 DH3) 'E9)
(explain '(DH2 DH3) 'E10)
(explain '(DH2 DH3) 'E12)
(explain '(DH2 DH3) 'E13)
(explain '(DH2 DH3) 'E14)
(explain '(DH2 DH3) 'E15)

Darwin's analogies
(explain '(DF2) 'DF3)
(explain '(DF2) 'DF7)
(analogous '(DF2 DH2) '(DF3 DH3))
(analogous '(DF2 DH2) '(DF7 E14))

Creationist explanations
(explain '(CH1) 'E1)
(explain '(CH1) 'E2)
(explain '(CH1) 'E3)
(explain '(CH1) 'E4)

Contradiction
(contradict 'CH1 'DH3)

Data
(data '(E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14 E15))
(data '(DF1 DF2 DF3 DF4 DF5 DF6 DF7))

esis. Creationists, of course, would marshal different arguments.

For clarity, Figure 12 omits the links from DH2 to all the evidence propositions besides E5, and the links from DH2 and DH3 to DH4, DH5, and DH6. Figure 13 shows the actual connectivity of DH3. Running ECHO to adjust the network to maximize harmany produces the expected result: Darwin's hypotheses are all activated, whereas the creation hypothesis is deactivated. In particular, the hypothesis DH3 – that species evolved – reaches an asymptote at .921, while the creation hypothesis, CH1, declines to −.491. DH3 accrues activation in three ways. It gains activation from above, from being explained by natural selection, which is derived from the struggle for existence, and from below, by virtue of the many pieces of evidence it helps to explain. In addition, it receives activation by virtue of the sideways, analogy-based links with explanations using artificial selection. Figure 14 graphs the activation histories of most of the units over the 70 cycles it takes them to settle. Note that the creationist hypothesis, CH1, initially gets activation by virtue of what it explains, but is driven down by the rise of DH3, which contradicts it.

The Lavoisier and Darwin examples show that ECHO can handle very complex examples of actual scientific reasoning. One might object that in basing ECHO analyses on written texts, I have been modeling the rhetoric of the scientists, not their cognitive processes. Presumably, however, there is some correlation between what we write and what we think. ECHO could be equally well applied to explanatory relations that were asserted in the heat of verbal debate among scientists. Ranney and Thagard (1988) describe ECHO's simulation of naive subjects learning physics, where the inputs to ECHO were based on verbal protocols.

6. Applications of ECHO to legal reasoning

Explanatory coherence is also important for some kinds of legal reasoning. Most discussions of legal reasoning concern either deductive inference, in which legal principles, rules, or statutes are applied to particular cases, or analogical inference, in which past cases are used as precedents to suggest a decision in a current case (Carter 1984; Gardner 1987; Golding 1984). Recently, however, some attention has been paid to the role of explanatory inferences in legal reasoning (Hanen 1987; Pennington & Hastie 1986; 1987). These researchers are concerned primarily with inferences made by juries about factual, rather than legal, questions. In murder trials, for example, juries can be called upon to infer what happened, choosing between contradictory accounts provided by the prosecution and the defense. To get a conviction on a first-degree murder charge, the prosecution must show (1) that the accused killed the victim and (2) that the accused did so with a previously formed purpose in mind. The first proposition must account for much of the evidence; the second provides one possible explanation of the first. The defense may try to defend alternative hypotheses, such as that someone else killed the victim or that the accused acted in self-defense and therefore is innocent, or that the accused acted in the heat of the moment and is therefore guilty only of manslaughter. The defense need not provide an alternative explanation of the killing, but may undermine the explanatory coherence of the prosecution's account by providing alternative interpretations of key testimony. For example, in the Peyer murder trial discussed below, the defense tried to discredit two important witnesses for the prosecution who had come forward just before the trial (a year after the killing) by saying that they were merely seeking publicity and had not seen what they claimed.

In terms of my theory of explanatory coherence and ECHO, we can think of the prosecution and defense as advocating incompatible ways of explaining the evidence. But, as in scientific reasoning, explanatory inference in the legal domain is not simply a matter of counting which of two hypotheses explains the most pieces of evidence. More complicated organizations of hypotheses and evidence will often arise. The hypothesis that the accused intended to kill the victim will be more plausible if we can explain why the accused had it in for the victim, say, because of a previous altercation. Analogy can also play a

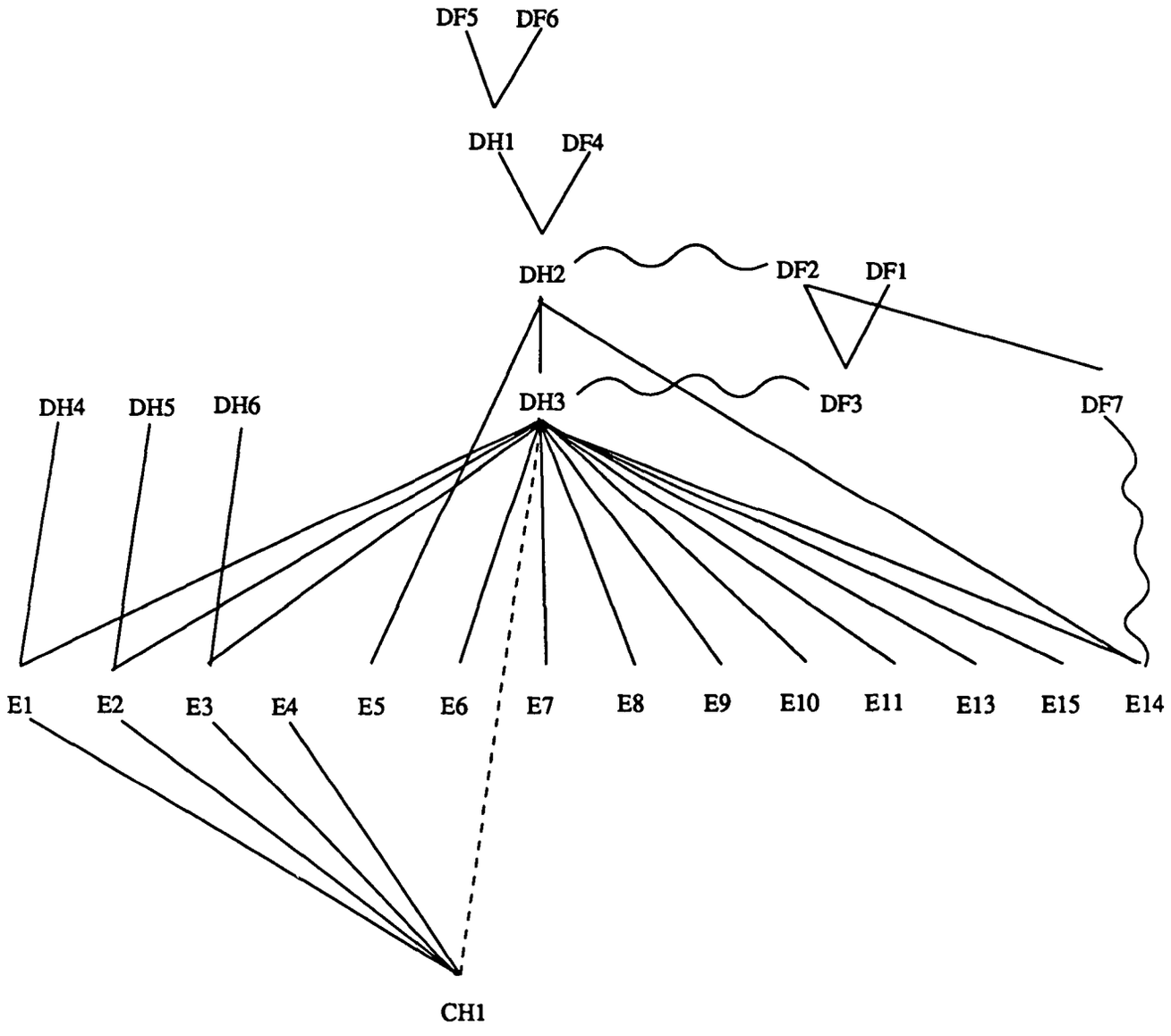

Figure 12. Network representing Darwin's (1962) argument. E1–E15 are evidence units. DH2 represents natural selection, and DH3 represents evolution of species. These defeat CH1, which represents the hypothesis that species were independently created. Solid lines are excitatory links; dotted line is inhibitory.

role: Pennington and Hastie (1986, p. 254) report that jurors sometimes evaluate the plausibility of explanations by considering how *they* would act in analogous situations. For example, a juror might reason, "If the victim had done to me what he did to the accused, then I would be angry and would want to get back at him, so maybe the accused did intend to kill the victim." Explanatory inferences can also be relevant to evaluating the testimony of a witness. If a witness who was a good friend of the accused says they were together at the time of the murder, the jury has to decide whether the best explanation of the witness's utterance is that (a) the witness really believe it or (b) the witness was lying to protect the accused.

The plausibility of a theory of explanatory coherence for legal reasoning depends on its application to real cases. ECHO has been used to model reasoning in two recent murder trials: the "preppy" murder trial in which Robert Chambers was accused of murdering Jennifer Levin in New York City and the San Diego trial in which Craig Peyer was accused of murdering Cara Knott. In both cases, there were no witnesses to the killing, so the juries had to infer on the basis of circumstantial evidence what actually happened.

6.1. Chambers. On August 26, 1986, Robert Chambers, by his own admission, killed Jennifer Levin in Central Park after the two had left a bar together. He maintained, however, that the killing was accidental, occurring when he struck her by reflex when she hurt him during rough sex. The prosecution maintained, in contrast, that he had killed her intentionally during a violent struggle. The trial took place in the first three months of 1988 and was extensively reported in the press. The following ECHO analysis is based on daily reports in the *New York Times* that described the major testimony and arguments. This information is, of course, not nearly as complete as that presented in the courtroom itself, but it suffices for

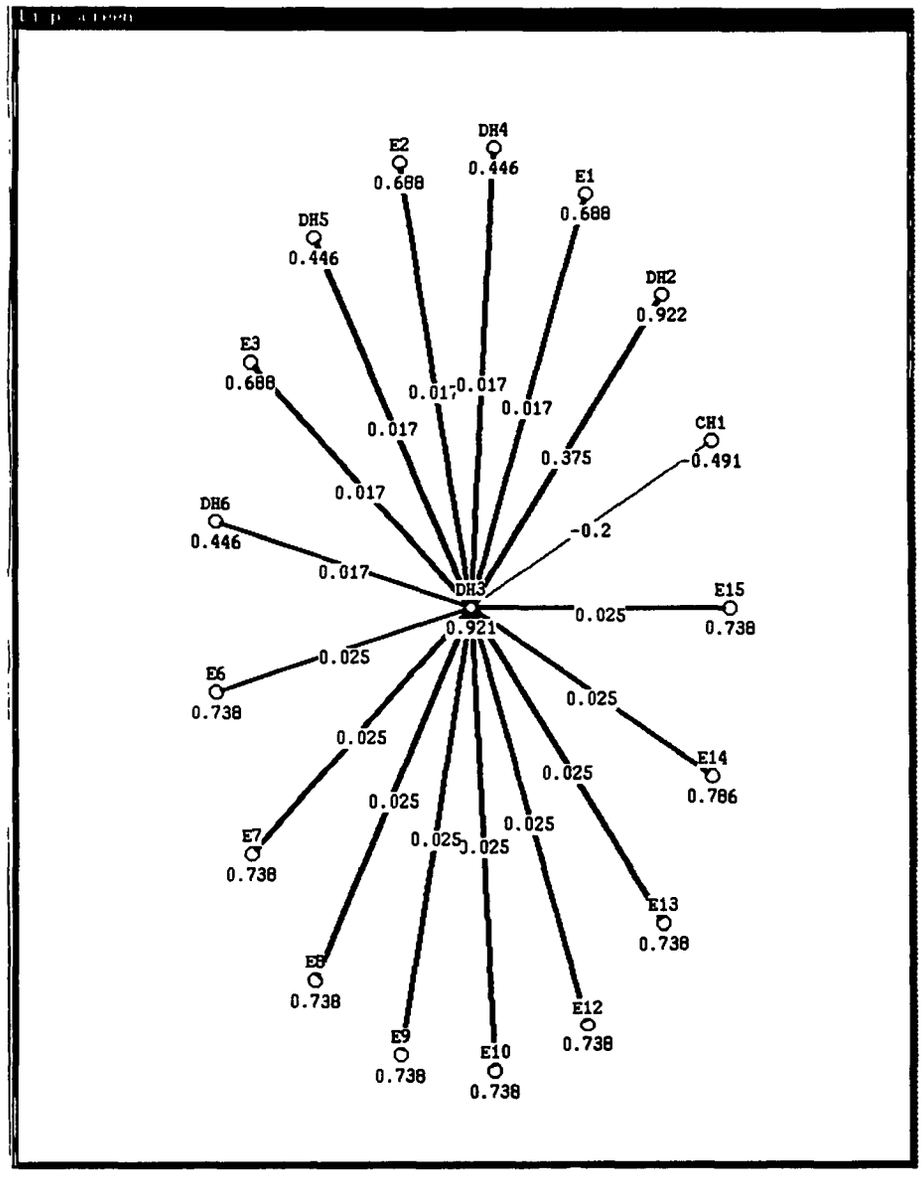

Figure 13. Connectivity of unit DH3 in the Darwin network. The numbers under the units are their activation values after the unit has settled. Thick lines indicate excitatory links; thin line indicates inhibitory link. Numbers on the lines indicate the weights on the links.

displaying the structure of a very complex explanatory inference (see also Taubman 1988).

The input to ECHO is shown in Tables 5 and 6. G1–G7 are hypotheses used by the prosecution to argue for Chambers's guilt, whereas I1–I8 present a very different explanatory account that supports his innocence. Figure 15 shows part of the network produced by this input, with excitatory links shown by solid lines and inhibitory ones shown by dotted lines (NE4 and NE9 are omitted to relieve crowding). The evidence propositions E0–E16 are indicated by number alone. Notice the layers of explanations: I3 explains I4, which explains I1, which explains E4. The prosecution's case does a better job of explaining the physical evidence using hypotheses concerning a struggle and a strangling. I have included two units, G6 and G7, to represent the question of Chambers's intent, which is crucial for deciding whether he is guilty of second-degree murder (he intended to kill her) or manslaughter (he intended merely to hurt her).

Running the network produces a clear win for G1, the main hypothesis implying Chambers's guilt. Figure 16 shows the links to G1 and the asymptotic activation of the units linked to it. Figure 17 displays the activation histories of all the units over 80 cycles. In the actual trial, the jury never got a chance to finish deciding the second-degree murder charge because a manslaughter plea-bargain was arranged during their deliberations. One important aspect that is not directly displayed in this simulation is the notion of determining guilt "beyond a reasonable doubt." Perhaps hypotheses concerning innocence should receive special activation so that hypotheses concerning guilt have to be very well supported to overcome them. Alternatively, we could require a high tolerance level so that guilt hypotheses would only be able to deactivate innocence hypotheses that were markedly inferior.

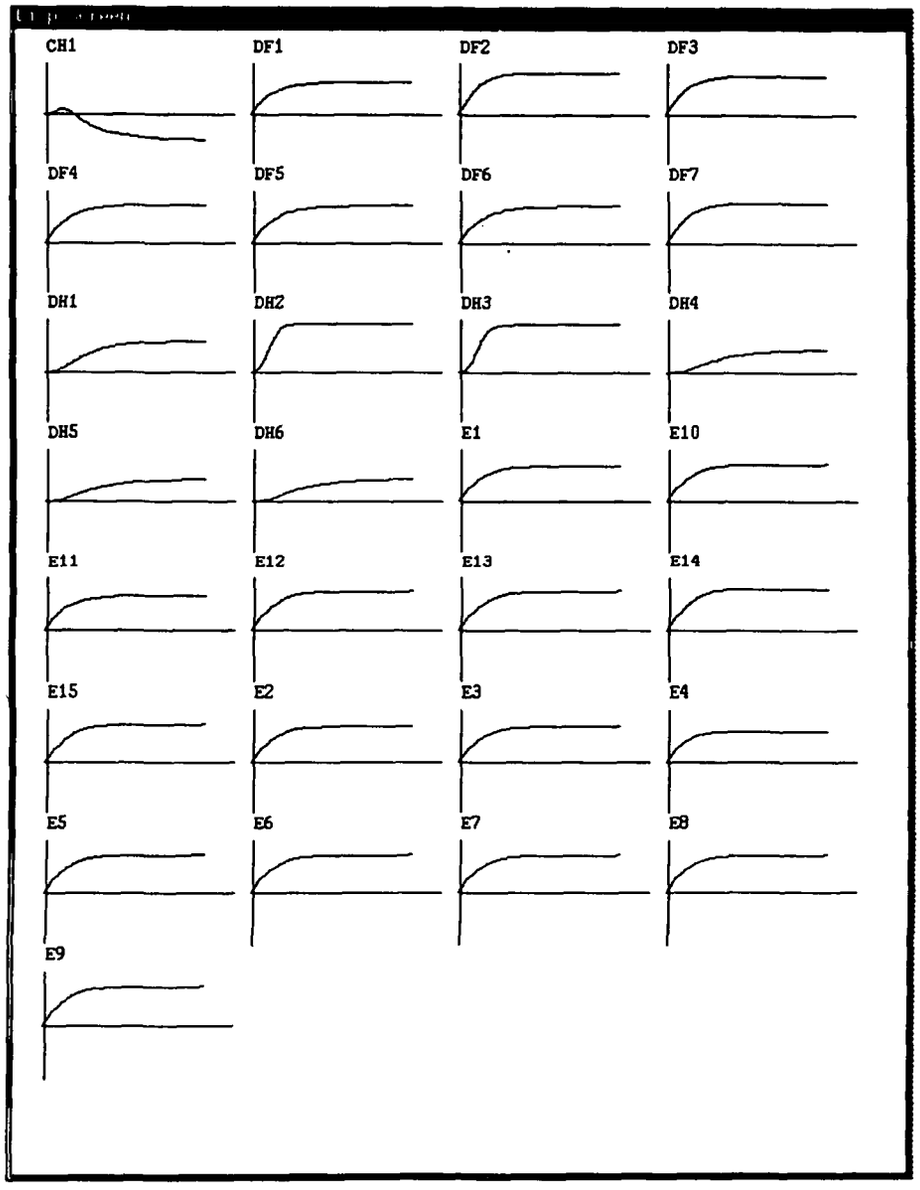

Figure 14. Activation history of the Darwin network. Each graph shows the activation of a unit over 49 cycles of updating, on a scale of −1 to 1, with the horizontal line indicating the initial activation of 0.

6.2. Peyer. Let us now consider another recent trial, where the evidence was less conclusive. Cara Knott was killed on December 27, 1986, and Craig Peyer, a veteran California Highway Patrolman, was accused. Twenty-two women, young and attractive like the victim, testified that they had been pulled over by Peyer for extended personal conversations near the stretch of road where Knott's body was found. The trial in San Diego ended February 27, 1988, and ECHO analysis is based on very extensive coverage (two full pages) that appeared the next day in the *San Diego Union* and the *San Diego Tribune.*

Tables 7 and 8 show the inputs to ECHO representing the evidence, hypotheses, and explanatory and contradictory statements in the Peyer trial. As in the Chambers representation, the G propositions are hypotheses concerning Peyer's guilt, whereas the I propositions concern his innocence. The prosecution can be understood as arguing that the hypothesis that Peyer killed Knott is the best explanation of the evidence, whereas the defense contends that the evidence does not support that claim beyond a reasonable doubt. Figure 18 shows the network ECHO sets up using the input given to it. Figure 19 shows the connectivity of the unit G1 along with the asymptotic activation of units linked to it, and Figure 20 graphs the activation histories of most of the units, omitting E1 and E2 for lack of space.

Peyer's trial ended in a hung jury, with seven jurors arguing for conviction on the second-degree murder charge and five arguing against it; the case is being retried. Figure 20 shows that ECHO finds more explanatory coherence in the guilt hypotheses than in the innocence hypotheses, although the activation of some of the I units shows that, in part, the defense had a more convincing case. Why, then, were some jurors reluctant to convict? It could, in part, be the question of establishing guilt beyond a reasonable doubt. The sensitivity analyses reported in the Appendix (see Table 11) show that ECHO rejects the hypothesis of Peyer's innocence much less strongly than it rejects the hypothesis of Chambers's innocence. With greater tolerance accruing from some-

Table 5. *Input propositions for Chambers case*

Evidence	
(proposition 'E0	"L died.")
(proposition 'E1	"L had wounds on her neck.")
(proposition 'E2	"L said she liked sex with C.")
(proposition 'E3	"L's blouse was around her neck.")
(proposition 'E4	"L's panties were not found near her.")
(proposition 'NE4	"L's panties were found near her.")
(proposition 'E5	"The police were careless about evidence.")
(proposition 'E6	"C lied to L's friend about not having seen L.")
(proposition 'E7	"C had scratches on his face and cuts on his hands.")
(proposition 'E8	"C had a broken hand.")
(proposition 'E9	"The skin on C's hand was not broken.")
(proposition 'NE9	"The skin on C's hand was broken.")
(proposition 'E10	"L's left eye was swollen and her mouth was cut.")
(proposition 'E11	"L's face was dirty.")
(proposition 'E12	"L had pinpoint hemorrhages in eye tissue.")
(proposition 'E13	"L's neck had severe hemorrhages.")
(proposition 'E14	"Bloodstains of C's type were found on L's jacket.")
(proposition 'E15	"C's fingers were bitten.")
(proposition 'E16	"C's video said he had hit her once.")
Hypotheses that Chambers is guilty	
(proposition 'G1	"C strangled L.")
(proposition 'G2	"C and L struggled.")
(proposition 'G3	"C lied about what happened.")
(proposition 'G4	"L's neck was held for at least 20 seconds.")
(proposition 'G5	"C broke his hand punching L.")
(proposition 'G6	"C intended to kill L.")
(proposition 'G7	"C intended to hurt L but not kill her.")
Hypotheses that Chambers is innocent	
(proposition 'I1	"C killed L with a single blow.")
(proposition 'I2	"The marks on L's neck were a scrape from C's watchband.")
(proposition 'I3	"L was having sadistic sex with C.")
(proposition 'I4	"L squeezed C's testicles.")
(proposition 'I5	"The police moved L's panties.")
(proposition 'I6	"C broke his hand falling on a rock.")
(proposition 'I7	"C threw L over his shoulder.")
(proposition 'I8	"C's blow triggered carotid sinus reflex.")

Note: L is Jennifer Levin; C is Robert Chambers.
Source: Data gathered from daily reports in the *New York Times* over a three-month period in 1988; see also Taubman (1988).

what higher excitation or lower inhibition, the unit representing Peyer's innocence is not deactivated.

It is also possible that matters extraneous to explanatory coherence were playing the key role in convincing some of the jurors against conviction. One juror was quoted as saying that a California Highway Patrolman with 13 years of service could never have committed a murder. This line of reasoning is represented partially by I8, which in the above simulation is swamped by G1, but a juror could give E17 (Peyer's spotless record) such a high priority that I8 could defeat G1. The simulation here is not claimed to handle all the factors that doubtless go into real jurors' decisions: "I could tell he was lying because he had shifty eyes," "The defense lawyer was such a nice man," "If he wasn't guilty of this, he was guilty of something else just as bad," and so on. But ECHO successfully handles a large part of the evidence and hypotheses in these two complex cases of legal reasoning.

7. Limitations of ECHO

It is important to appreciate what ECHO cannot do as well as what it can. The major current limitation on ECHO is that the input propositions, explanation statements, and contradiction statements are constructed by the programmer. How arbitrary are these encodings? Several different people have successfully done ECHO analyses, on more than a dozen disparate cases. In all four of the examples presented in this paper, virtually no adjustment of input was required to produce the described runs. We have not yet done the experiment of having several people analyze the same case and assessing the intercoder reliability, however. We can nevertheless maintain that the representations are not arbitrary thought experiments, because they are derived from scientific texts, newspaper reports of trials, and subject protocols.

ECHO's scope is not universal: Not every case of reason-

Table 6. *Explanations and contradictions in Chambers example*

Data
(data '(E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14 E15 E16))

Contradictions
(contradict 'G1 'I1)
(contradict 'G4 'I1)
(contradict 'G5 'I6)
(contradict 'G1 'I2)
(contradict 'G2 'I3)
(contradict 'G6 'G7)
(contradict 'E4 'NE4)
(contradict 'E9 'NE9)

Explanations supporting Chambers's innocence
(explain '(I1 I8) 'E0)
(explain '(I2) 'E1)
(explain '(I3) 'I4)
(explain '(I4) 'I1)
(explain '(I3 I5) 'E4)
(explain '(I6) 'E8)
(explain '(I6) 'NE9)
(explain '(I7) 'E12)
(explain '(I3) 'E15)
(explain '(I1) 'E16)

Explanations supporting Chambers's guilt
(explain '(G2) 'G1)
(explain '(G2) 'E3)
(explain '(G2) 'E4)
(explain '(G2) 'E7)
(explain '(G2) 'E12)
(explain '(G1) 'G4)
(explain '(G1) 'E0)
(explain '(G1) 'E1)
(explain '(G2) 'E10)
(explain '(G2) 'E11)
(explain '(G4) 'E13)
(explain '(G2) 'E14)
(explain '(G2) 'E15)
(explain '(C5) 'E8)
(explain '(C5) 'E9)
(explain '(G3) 'E16)
(explain '(G3) 'E6)
(explain '(G6) 'G1)
(explain '(G7) 'G1)
(explain '(G2) 'G7)

ing can be analyzed for ECHO's application. In doing our analyses, we try to restrict the EXPLAIN statements to cases where there is a causal relation. A research assistant attempted to use ECHO to analyze arguments in this journal for and against parapsychology (Rao & Palmer 1987; Alcock 1987), but concluded that ECHO was not appropriate. This debate largely concerns the reliability of parapsychological experiments, and ECHO is not a data analyzer. ECHO would be appropriate for this case only if there were a general parapsychological theory whose explanatory coherence could be evaluated. The general conclusion of Rao and Palmer is that parapsychological experiments are not explainable with current science, but that conclusion is not in itself an explanation of the experiments.

From a logical point of view, the analysis of explanatory relations is easily trivialized. The explanations of E1 and E2 by hypotheses H1 and H2 together can be collapsed logically by conjoining E1 and E2 into E3, and H1 and H2 into H3, so that we are left with only the boring explanation of E3 by H3. Fortunately, in real disputes in law and the history of science, such trivializations do not occur. We can easily get the appropriate level of detail by attending to the claims that scientists and lawyers make about the explanatory power of their theories. Lavoisier and the phlogiston theorists operated at roughly the same level of detail. In analyzing texts to assess explanatory coherence, I recommend the following maxim:

Detail Maxim.

In analyzing the propositions and explanatory relations relevant to evaluating competing theories, go into as much detail as is needed to distinguish the explanatory claims of the theories from each other, and be careful to analyze all theories at the *same* level of detail.

Following this maxim removes much of the apparent arbitrariness inherent in trying to adjudicate among theories.

Ideally, we would want to automate the production of the input to ECHO. This could be done either in a natural language system capable of detecting explanatory arguments (cf. Cohen, R. 1983) or, more easily, in an integrated system of scientific reasoning that formed explanatory hypotheses which could then be passed to ECHO for evaluation. PI (which is short for "processes of induction" and is pronounced "pie") is a crude version of such a system (Thagard 1988a). In PI, it is possible to represent hypotheses like those in the scientific examples discussed above using rules. One of Lavoisier's principles might be translated into the rule:

If x is combustible and x combines with oxygen, then x burns.

Like other rule-based systems, PI can use such rules to make inferences. Given a set of such rules, PI can be set the task of explaining other rules representing the evidence. While PI runs, it is possible to keep track of which rules were used in explaining which pieces of evidence. Thus explanation from this computational point of view is a process of derivation that can be inspected to determine what was actually used in deriving what. Tracing back to which hypotheses were used in deriving which evidence could generate the EXPLAIN formulas that are input for ECHO. Because PI does not have the rules of inference that permit logicians to concoct nonexplanatory deductions – for example, to infer (A or B) from A – we can identify what hypotheses played a role in explaining what pieces of evidence. Putting together all the rules to make up Lavoisier's theory and furnish explanations is a daunting task, because his writings and my summary for ECHO omit much background knowledge that would have to be dredged up and included if the derivations were to look complete. But artificial intelligence models of problem solving and learning such as PI provide at least a glimpse of how explanations can be noticed. Falkenhainer and Rajamoney (1988) describe a system that combines hypothesis formation by analogy with hypothesis evaluation by experimental design. So eventually it should be possi-

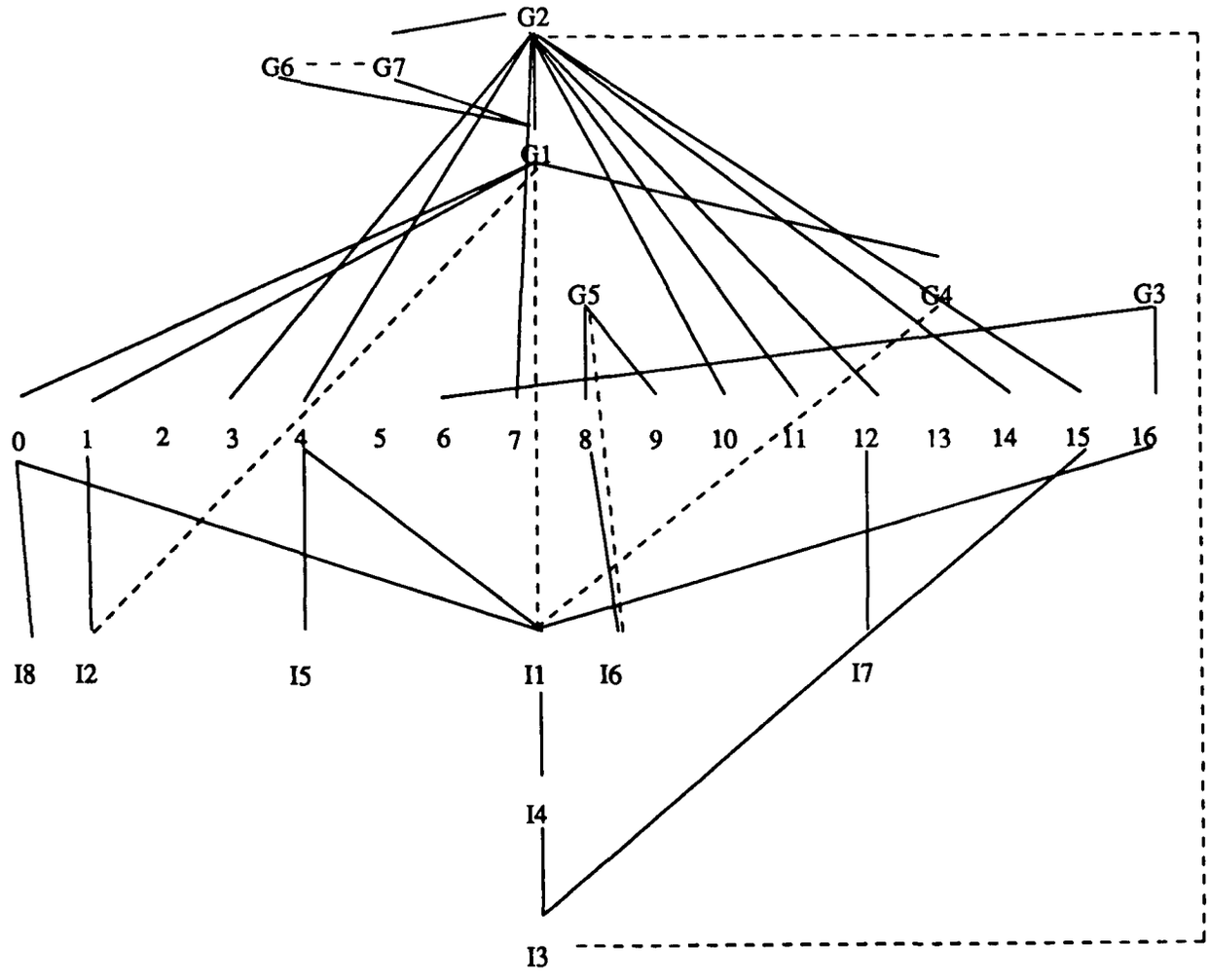

Figure 15. Network representing the Chambers trial. 1–16 are evidence units. G1–G7 represent hypotheses concerning Chambers's guilt; I1–I8 represent his innocence. Solid lines are excitatory links; dotted lines are inhibitory.

ble to integrate ECHO with a system that generates explanations and provides its input automatically.

ECHO is a very natural way of implementing the proposed theory of explanatory coherence, but one might argue for the construction of a nonconnectionist coherence model. Perhaps it could be based on simple rules such as the following:

(1) If a proposition is a piece of evidence, then accept it.

(2) If a proposition contradicts an accepted proposition, then reject it.

(3) Of two contradictory hypotheses, accept the one that coheres (by virtue of explanatory and analogical relations) with more accepted propositions and has fewer cohypotheses.

(4) If a proposition does not contradict any other propositions, accept it if it coheres with more accepted propositions than rejected ones.

Analysis suggests that an implementation of such rules could be a fair approximation to ECHO for many cases, but would lack several advantages that derive from ECHO's connectionist algorithms. First, rules such as (1) and (2) are much too categorical. ECHO is capable of rejecting a piece of evidence if it coheres only with a very inferior theory (section 4.8), just as scientists sometimes throw out data. Similarly, a hypothesis should not be rejected just because it makes a false prediction, because additional assumptions may enable it to explain the evidence and explain away the negative result. Second, the rule-based implementation would be very sensitive to the order of application of rules, requiring that the four rules stated above be applied in approximately the order given. Moreover, if a hypothesis is contradicted by two other propositions, it will be important to evaluate the other propositions first so that together they can count against the given hypothesis, otherwise it might be accepted and then knock them out one at a time. ECHO's parallelism enables it to evaluate all propositions simultaneously, so these undesirable order effects do not arise. Third, rules (3) and (4) above should not operate in isolation from one another: In our simulation of Wegener's argument for continental drift (Thagard & Nowak 1988), units representing the views that Wegener rejects become deactivated because of a *combination* of being contradicted and being coherent with rejected propositions. Fourth, the rule-based system's use of the binary categories of acceptance and rejection will prevent it from having the sensitivity of ECHO in indicating *degrees* of acceptance and rejection by degrees of activation. Fifth, the rule-based system does not come with a metric for system coherence (section 4.9). Thus, although ECHO is not the only possible means for computing coherence, its connectionist

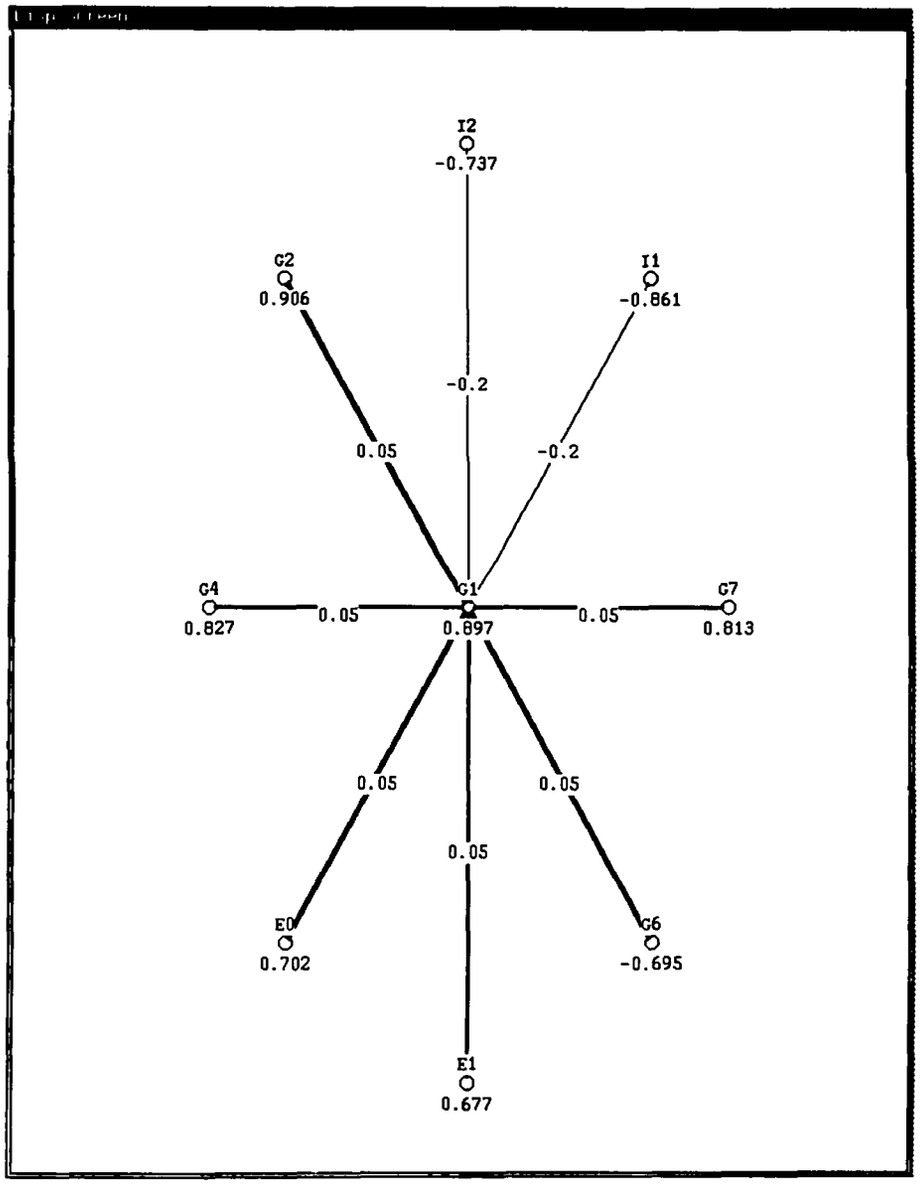

Figure 16. Connectivity of the unit G1, representing the claim that Chambers strangled Levin. The numbers under the units are their activation values after the unit has settled. Thick lines indicate excitatory links; thin lines indicate inhibitory links. Numbers on the lines indicate the weights on the links.

algorithms give it many natural advantages over alternative approaches.

Finally, as an implementation of a theory of explanatory coherence, ECHO is only as good as the principles in that theory. The seven principles of explanatory coherence seem now to be complete enough to characterize a wide range of cases of hypothesis evaluation, but they are themselves hypotheses and therefore subject to revision.

8. Implications for artificial intelligence

The theory of explanatory coherence and its implementation in ECHO have implications for research in the areas of artificial intelligence, cognitive psychology, and philosophy. Like the evaluation of scientific theories, the evaluation of philosophical and computational theories is a comparative matter. While discussing the computational, psychological, and philosophical significance of the approach proposed here, I shall compare it with similar research in these fields.

8.1. Connectionism. Very recently, other researchers have also suggested connectionist models for the evaluation of explanatory hypotheses. Peng and Reggia (in press) describe a connectionist model for diagnostic problem solving. Theoretically, it differs from my proposal most in that it does not use constraints involving simplicity (in the sense indicated by Principle 2[c]), analogy, and the desirability of a hypothesis being explained as well as explaining. Their implementation differs from ECHO most strikingly in that it does not use inhibitory links between units representing incompatible hypotheses, but instead has nodes competing for activation from the output of a source node. Goel et al. (1988) propose an architecture that chooses the best explanation by considering explanatory coverage of data, number of hypotheses, and prior plausibility of hypotheses. ECHO uses the

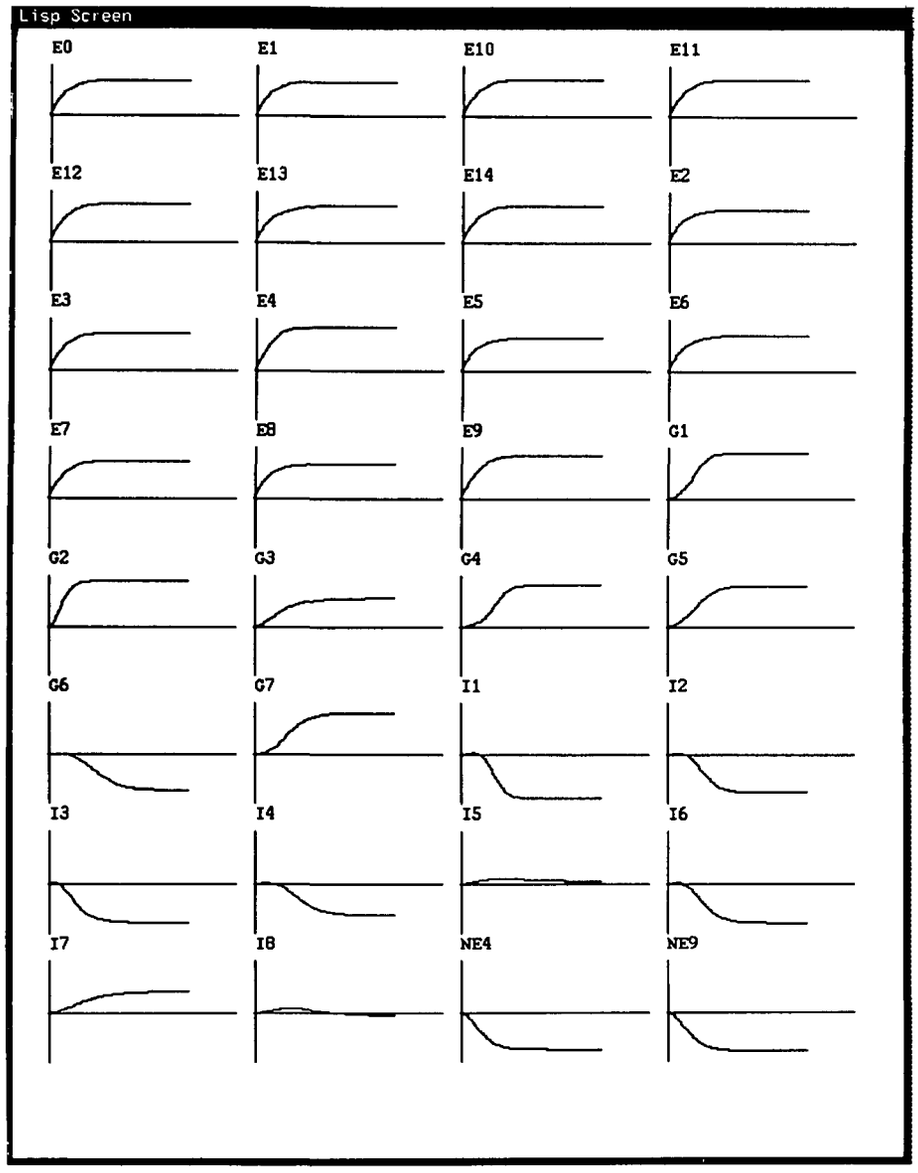

Figure 17. Activation history of the Chambers network. Each graph shows the activation of a unit over 59 cycles of updating, on a scale of −1 to 1, with the horizontal line indicating the initial activation of 0.

first two of these criteria, but not the third, because in the domains to which it has been applied, plausibility appears to be determined by explanatory coherence alone.

Parallel constraint satisfaction models somewhat similar to ECHO have been proposed for other phenomena: analogical mapping (Holyoak & Thagard, in press), analog retrieval (Thagard et al. 1989), discourse processing (Kintsch 1988), and word pronunciation retrieval (Lehnert 1987). (See also surveys by Feldman & Ballard 1982 and Rumelhart et al. 1986.) These systems differ from Boltzmann machines and back-propagation networks (Rumelhart et al. 1986) in that they do not adjust weights while the network is running, only activations.

ECHO's connectionist character may prompt immediate boos or cheers from different partisan quarters. Currently, debate rages in cognitive science concerning competing methodologies. We can distinguish at least the following approaches to understanding the nature of mind and intelligence:

(1) Straight neuroscience, studying neurons or sections of the brain

(2) Computational models of actual neurons in the brain

(3) Connectionist models using distributed representations, so that a concept or hypothesis is a pattern of activation over multiple units

(4) Connectionist models using localist representations, in which a single unit represents a concept or proposition

(5) Traditional artificial intelligence models using data structures such as frames and production rules

(6) Psychological experiments

(7) Mathematical analysis

(8) Theoretical speculation

ECHO falls into (4), but I reject as *methodological imperialism* the opinion that other approaches are not worth pursuing as well. In the current neonatal state of cognitive science, restrictions on ways to study the mind are

Table 7. *Input propositions for the Peyer example*

Evidence	
(proposition 'E1	"Knott's body and car were found on a frontage road near I-15.")
(proposition 'E2	"22 young women reported being talked to at length by Peyer after being stopped near where Knott's body was found.")
(proposition 'E3	"Calderwood said that he saw a patrol car pull over a Volkswagon like Knott's near I-15.")
(proposition 'E4	"Calderwood came forward only at the trial.")
(proposition 'E5	"Calderwood changed his story several times.")
(proposition 'E6	"6 fibers found on Knott's body matched Peyer's uniform.")
(proposition 'E7	"Ogilvie said Peyer quizzed her about the case and acted strangely.")
(proposition 'E8	"Dotson said Olgivie is a liar.")
(proposition 'E9	"Anderson and Schwartz saw scratches on Peyer's face the night of the killing.")
(proposition 'E10	"Martin said she saw Peyer pull Knott's Volkswagon over.")
(proposition 'E11	"Martin came forward only just before the trial.")
(proposition 'E12	"Anderson says she saw Peyer wipe off his nightstick in his trunk.")
(proposition 'E13	"Anderson did not say anything about the nightstick when she was first interrogated.")
(proposition 'E14	"Bloodstains found on Knott's clothes matched Peyer's blood.")
(proposition 'E15	"12,800 other San Diegans had blood matching that on Knott's clothes.")
(proposition 'E16	"A shabby hitchhiker was lunging at cars near the I-15 entrance.")
(proposition 'E17	"Peyer had a spotless record with the California Highway Patrol.")
Hypotheses that Peyer is guilty	
(proposition 'G1	"Peyer killed Knott.")
(proposition 'G2	"Knott scratched Peyer's face.")
(proposition 'G3	"Fibers from Peyer's uniform were transferred to Knott.")
(proposition 'G4	"Peyer pulled Knott over.")
(proposition 'G5	"Calderwood was reluctant to come forward because he wanted to protect his family from publicity.")
(proposition 'G6	"Peyer like to pull over young women.")
(proposition 'G7	"Peyer had a bloody nightstick.")
(proposition 'G8	"Anderson was having personal problems when first interrogated.")
Hypotheses that Peyer is innocent	
(proposition 'I1	"Someone other than Peyer killed Knott.")
(proposition 'I2	"Calderwood made his story up.")
(proposition 'I3	"The 6 fibers floated around in the police evidence room.")
(proposition 'I4	"Ogilvie lied.")
(proposition 'I4A	"Ogilvie is a liar.")
(proposition 'I5	"Peyer' scratches came from a fence.")
(proposition 'I6	"Martin lied.")
(proposition 'I7	"Anderson was mistaken about the nightstick.")
(proposition 'I8	"Peyer is a good man.")

Source: Analysis based on coverage by the *San Diego Union* and *San Diego Tribune* on February 28, 1988.

clearly premature. By pursuing all eight strategies, we can hope to learn more about how to investigate the nature of mind. As suggested by my juxtaposition of PI and ECHO in section 7, I see no great incompatibility between connectionist systems and traditional symbolic AI. Much is to be gained from developing hybrid systems that exploit the strengths of both research programs (Hendler 1987; Lehnert 1987).

Despite ECHO's parallelism, and use of a vague neural metaphor of connections, I have not listed neural plausibility as one of its advantages, because current knowledge does not allow any sensible mapping from nodes of ECHO representing propositions to anything in the brain. For the same reason, I have not used the term "neural net." Parallelism has its advantages independent of the brain analogy (Thagard 1986).

8.2. Probabilistic networks. The account of explanatory coherence I have given bears some similarity to Pearl's (1986; 1987) work on belief networks. Pearl also represents propositions as nodes linked by inferential dependencies and uses a parallel algorithm to update numerical values assigned to the nodes. The major difference between ECHO and Pearl's networks, however, is that he construes numerical values as the *probabilities* of the propositions, and weights between nodes as conditional probabilities. Thus, in contrast to links established by coherence relations, Pearl's links are asymmetric, because in general the probability of P given Q is not equal to the probability of Q given P.

Although Pearl's probabilistic approach appears promising for domains such as medical diagnosis where we can empirically obtain frequencies of cooccurrence of diseases and symptoms and thus generate reasonable conditional probabilities, it does not seem applicable to the cases of explanatory coherence I have been considering. What, for example, is the conditional probability of burned objects gaining in weight given the hypothesis

Table 8. *Explanations and contradictions in the Peyer example*

The case for Peyer's guilt
(explain '(G1) 'G2)
(explain '(G1) 'G3)
(explain '(G1) 'G7)
(explain '(G1) 'E1)
(explain '(G6) 'E2)
(explain '(G4) 'E3)
(explain '(G5) 'E4)
(explain '(G3) 'E6)
(explain '(G1) 'E7)
(explain '(G2) 'E9)
(explain '(G1) 'E10)
(explain '(G7) 'E12)
(explain '(G8) 'E13)
(explain '(G1) 'E14)

The case for Peyer's innocence
(explain '(I1) 'E1)
(explain '(I2) 'E4)
(explain '(I2) 'E5)
(explain '(I3) 'E6)
(explain '(I4) 'E7)
(explain '(I4A) 'E8)
(explain '(I4A) 'I4)
(explain '(I5) 'E9)
(explain '(I6) 'E10)
(explain '(I6) 'E11)
(explain '(I7) 'E12)
(explain '(I7) 'E13)
(explain '(I1 E15) 'E14)
(explain '(I1) 'E16)
(explain '(I8) 'E17)

Contradictions
(contradict 'G1 'I1)
(contradict 'G5 'I2)
(contradict 'G7 'I7)
(contradict 'G1 'I8)
(contradict 'G2 'I5)
(contradict 'G3 'I3)

Data
(data '(E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14 E15 E16 E17))

that oxygen is combined with them? It would be 1 if the hypothesis entailed the evidence, but it does so only with the aid of the additional hypothesis that oxygen has weight, and some unstated background assumption about conservation of weight. To calculate the conditional probability, then, we need to be able to calculate the conjunctive probability that oxygen has weight and that oxygen combines with burning objects, but these propositions are dependent to an unknown degree. Moreover, what is the probability that the evidence is correct? In contrast to the difficulty of assigning probabilities to these propositions, the coherence relations established by my principles are easily seen directly in arguments used by scientists in their published writings. When frequencies are available because of empirical studies, probabilistic belief networks can be much more finely tuned than my coherence networks, but they are ill-suited for the kinds of nonstatistical theory evaluation that abounds in much of science and everyday life.

One clear advantage to the probabilistic approach is that the properties of probabilities are naturally understood using the axioms of probability and their natural interpretation in terms of games of chance. Acceptability, as indicated in ECHO by activation levels, has no such precise interpretation. (See section 10.2 for further discussion of probability versus acceptability.)

8.3. Explanation-based learning. In machine learning, a rapidly growing part of AI, the term "explanation-based" is used to distinguish cases of knowledge-intensive learning from cases of simple learning from examples (see, for example, DeJong & Mooney 1986). Rajamoney and DeJong (1988) discuss the problem of "multiple explanations" and describe a program that does simulated experiments to select an explanatory account. This program is Popperian in spirit, in that the experiments concerning electricity and heat flow serve to refute all but one of the competing hypotheses. (Science is rarely so neat; see sections 4.4 and 10.4.) Systems that deal with more complex theories than those occurring in Rajamoney and DeJong's system will need a more comparative method of choosing among multiple explanations such as that found in ECHO.

Recently, there has been growing attention in AI to "abduction," construed as the construction and selection of competing explanatory hypotheses. (Peirce applied "abduction" only to hypothesis formation, but the term is used in many quarters to apply to hypothesis evaluation as well.) Abduction has been investigated in the domains of medical diagnosis (Josephson et al. 1987; Pople 1977; Reggia et al. 1983), natural language understanding (Hobbs et al. 1988), and folk psychology (O'Rorke et al. 1988). My account shares with these models the aim of finding the most comprehensive explanation, but it differs in both theory and implementation. The biggest theoretical difference is that my principles of explanatory coherence also favor hypotheses that are explained and fare well on considerations of simplicity and analogy. Leake (1988) describes a program for evaluating individual explanations, a problem different from selecting a hypothesis on the basis of how well it explains a wide range of evidence.

9. Implications for psychology

The theory of explanatory coherence described here is intended to describe approximately the way people reason concerning explanatory hypotheses (see section 10.5 for further discussion of the descriptive *and* normative character of the theory). The psychological relevance of explanatory coherence is evident in at least three important areas of psychological research: attribution theory, discourse processing, and conceptual change. After sketching how explanatory coherence is germane to these topics, I shall illustrate the testability of the ECHO model.

9.1. Attribution. Because the inferences that people make about themselves and others generally depend on causal theories, social psychology is a very rich domain for a

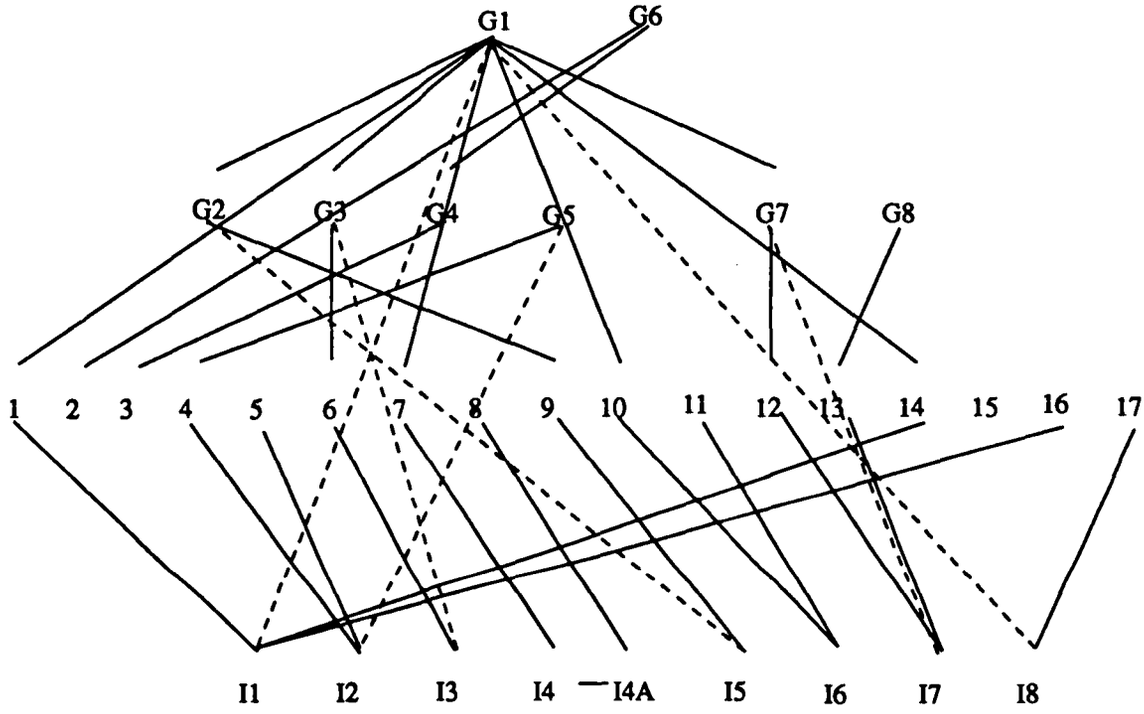

Figure 18. Network representing the Peyer trial. 1–17 are evidence units. G1–G8 represent hypotheses concerning Peyer's guilt; I1–I8 concern his innocence. Solid lines are excitatory links; dotted lines are inhibitory.

theory of explanatory coherence, and *attribution theory* has been a major focus of research for several decades. Research on attribution "deals with how the social perceiver uses information in the social environment to yield causal explanations for events" (Fiske & Taylor 1984, p. 21). Much of the theorizing about attribution can be understood in terms of explanatory coherence. For example, we can interpret the *correspondent inference theory* of Jones and Davis (1965) as saying that we accept hypotheses about the dispositional attributes of other people on the basis of the hypotheses providing coherent explanations of their behavior. Jones and Davis's discussion of the *analysis of noncommon effects* can be understood as saying that we infer that someone has one of a set of intentions because that intention explains some aspects of their behavior that the other intentions do not. Our inferences about other people's dispositions will also depend on the available alternative explanations of their behavior, such as coercion, social desirability, social role, and prior expectations. Explanatory coherence theory does not address the question of how people form these kinds of hypotheses, but it does show how people can select from among the hypotheses they have formed. I conjecture that if cases of attributional inferences were analyzed in sufficient detail to bring out the relevant data and hypotheses, preferences for situational or dispositional explanations would follow from the nature of the explanatory networks.

As we saw in the preppy murder trial, jurors often have to infer the intentions of witnesses and of the accused. Pennington and Hastie (1986; 1987) have interpreted the results of their experiments on juror decision making by hypothesizing that jurors make judgments based on considerations of explanatory coherence; their cases look ripe for ECHO analysis. Of course, ECHO does not model all the kinds of reasoning involved in these experiments. In particular, it does not model how the jurors process the statements about evidence and combine them into explanatory stories. But it does give an account of how jurors choose between stories on the basis of their explanatory coherence.

9.2. Discourse processing. The problem of recognizing intention in utterances can be understood in terms of explanatory coherence.[2] Clark and Lucy (1975) advocated a stage model of comprehension, according to which a literal meaning for an utterance is calculated before any nonliteral meanings are considered. In contrast, Gibbs (1984) and others have argued that hearers are able to understand that "Can you pass the salt?" is a request, without first interpreting it as a question. In explanatory coherence terms, we can think of competing hypotheses – that the utterance is a request and that it is a question – as simultaneously being evaluated with respect to what they explain and how they themselves are explained. To take an extreme example, the utterance might even be construed as an insult if it was expressed in a nasty tone of voice and if we had reason to believe that the utterer wanted to be insulting. Parallel evaluation of the different explanations of the utterance results in an appropriate interpretation of it.

Trabasso et al. (1984) have argued that causal cohesiveness is very important for story comprehension. They analyze stories in terms of networks of causally related propositions that are similar to ECHO's explanatory networks except that there are no links indicating contradictions. Comprehension differs from theory evaluation in lacking easily identified alternatives competing for acceptance. Still, it is possible that some mechanism similar to ECHO's way of activating a subset of mutually coherent propositions may be involved in reaching a satisfactory understanding of a story. Text comprehen-

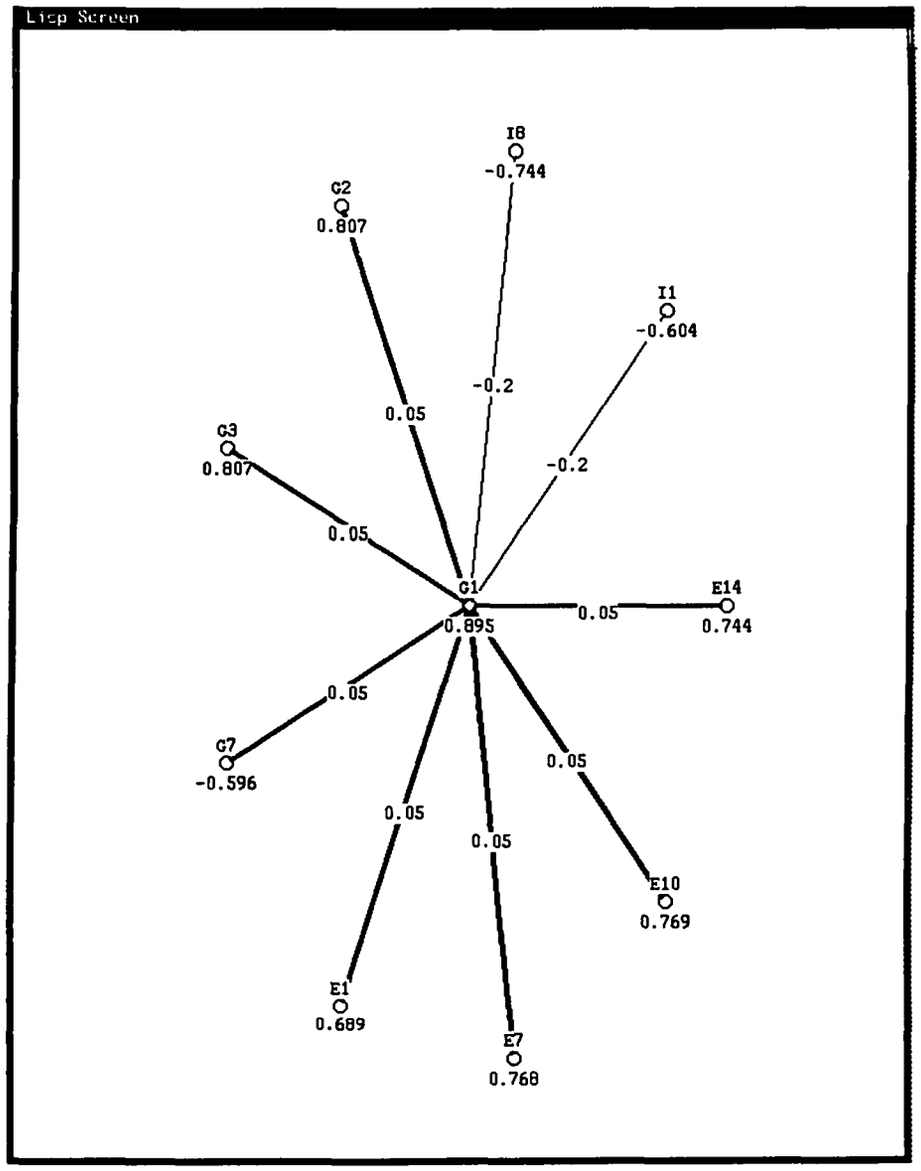

Figure 19. Connectivity of the unit G1, representing Peyer's guilt. The numbers under the units are their activation values after the unit has settled. Thick lines indicate excitatory links; thin lines indicate inhibitory links. Numbers on the lines indicate the weights on the links.

sion obviously involves many processes besides inferences about causal or explanatory coherence, but ECHO-like operation may nevertheless contribute to the necessary task of appreciating the causal cohesiveness of a story.

9.3. Belief revision and conceptual change. Ranney and Thagard (1988) describe the use of ECHO to model the inferences made by naive subjects learning elementary physics by using feedback provided on a computer display (Ranney 1987). Subjects were asked to predict the motion of several projectiles and then to explain these predictions. Analyses of verbal protocol data indicate that subjects sometimes underwent dramatic belief revisions while offering predictions or receiving empirical feedback. ECHO was applied to two particularly interesting cases of belief revision with propositions and explanatory relations based on the verbal protocols. The simulations captured well the dynamics of belief change as new evidence was added to shift the explanatory coherence of the set of propositions.

The theory of explanatory coherence sketched here has the capacity to explain major conceptual changes such as those that have been hypothesized to occur in scientific revolutions (Kuhn 1970; Thagard, in press b) and in children (Carey 1985). Because ECHO evaluates a whole network of hypotheses simultaneously, it is capable, when new data are added, of shifting from a state in which one set of hypotheses is accepted to a state in which an opposing set is accepted. This shift is analogous to the Gestalt switch described in section 3, except that scientists rarely shift back to a rejected view. Developmental psychologists have speculated about the existence of some kind of "transition mechanism" that could shift a child forward from a primitive conceptual scheme to an advanced one. We currently have insufficient experimental data and theoretical understanding to know whether knowledge development in children has the somewhat

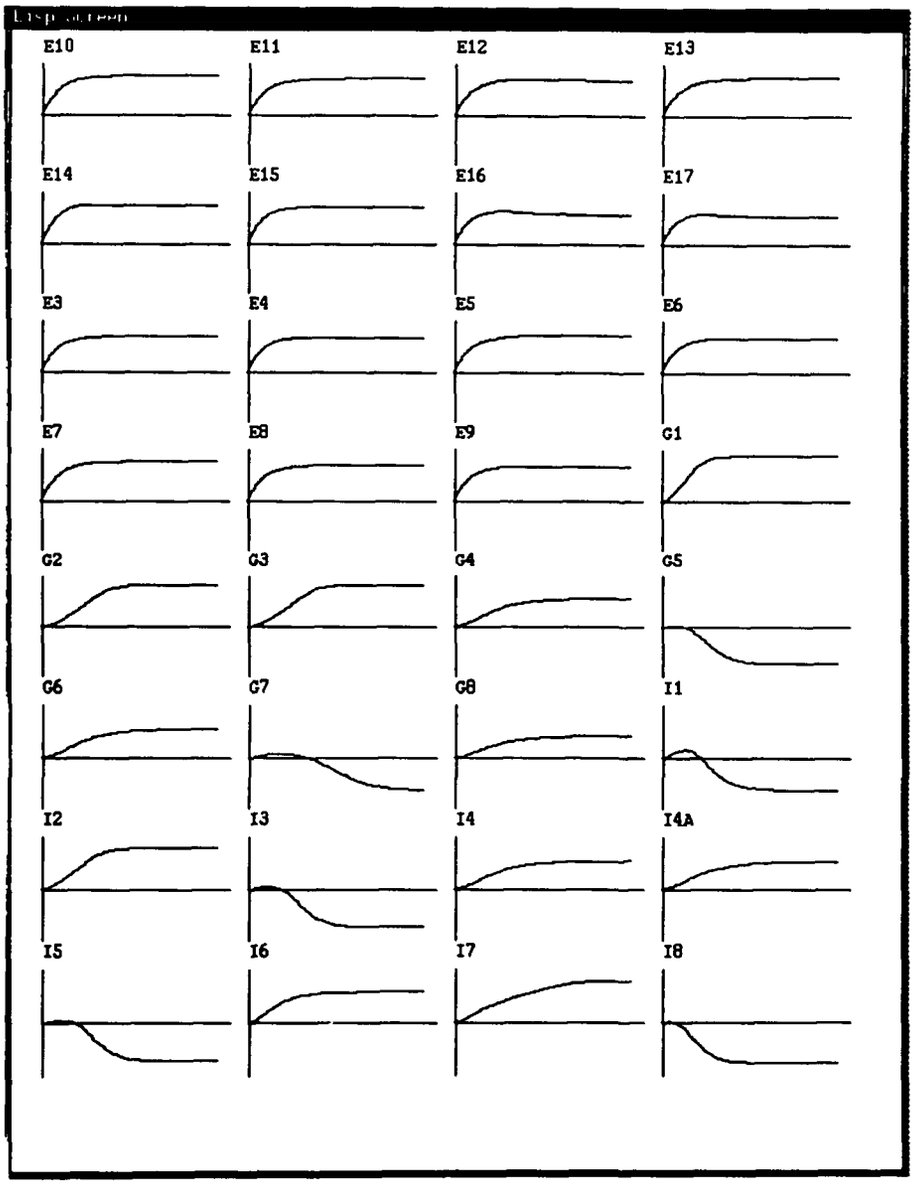

Figure 20. Activation history of the Peyer network. Each graph shows the activation of a unit over 54 cycles of updating, on a scale of −1 to 1, with the horizontal line indicating the initial activation of 0.

precipitous nature attributed to scientific revolutions. But if children do undergo dramatic changes in conceptual systems because they have acquired a more coherent way of understanding their worlds, then ECHO may be very useful for modeling the transition.

9.4. Testability. So far, my discussion of the psychological relevance of ECHO has been merely suggestive, showing that explanatory coherence judgments may be plausibly considered to contribute to important kinds of inferential behavior. A defense of ECHO as a psychological model, however, will require controlled experiments that provide a much finer-grained evaluation of the theory of explanatory coherence. Fortunately, there appears to be great potential for testing explanatory coherence theory and the ECHO model by comparing the performance of human subjects with ECHO-based predictions about qualitative and quantitative features of the acceptance and rejection of hypotheses. Michael Ranney and I are planning several studies in which subjects will be given textual descriptions of scientific and legal debates. We want to determine whether, when ECHO is run with inputs derived from subjects' own analyses of debates, the analyses predict their conclusions. We also want to determine whether manipulating textual descriptions of evidence, explanations, and contradictory hypotheses will affect the confidence that subjects have in different hypotheses in a way that resembles how manipulations affect ECHO's activation levels. It will also be interesting to find out whether important transitional points in the amount of evidence and explanation that tend to tip ECHO's activations over to new sets of accepted beliefs correspond to major shifts in subjects' beliefs at the same points. Such transitions, in both subjects and ECHO, are described in Ranney and Thagard (1988).

The methodology here is to use ECHO to test the psychological validity of the theory of explanatory coherence embodied in the seven principles in section 2.2. These principles in themselves are too general to have direct experimental consequences, but their implemen-

tation in ECHO makes possible very detailed predictions about the conclusions people will reach and the relative degree of confidence they will have in those conclusions. Merely proposing experiments does not, of course, show the psychological validity of the theory or the model, but it does show their joint testability. The theory of explanatory coherence presented in this paper has been well explored computationally, but I hope the above section shows that it is also suggestive psychologically.

10. Implications for philosophy

In philosophy, a theory of explanatory coherence is potentially relevant to metaphysics, epistemology, and the philosophy of science. In metaphysics, a coherence theory of truth, according to which a proposition is said to be true if it is part of a fully coherent set, has been advocated by idealist philosophers such as Bradley and Rescher (Bradley 1914; Rescher 1973; see also Cohen, L. J. 1978). In epistemology, the view that justified reasoning involves the best total explanatory account has been urged by Harman (1973; 1986) and contested by Goldman (1986) and Lehrer (1974). The theory of explanatory coherence in this paper is not aimed primarily at questions of truth or justification, but rather at the philosophy of science and of law, illuminating the kinds of reasoning used to justify the acceptance and rejection of scientific and legal hypotheses. The account of explanatory coherence offered here is as compatible with a correspondence theory of truth – according to which the truth of a proposition depends on its relation to an independent reality – as it is with a theory that attempts to define truth in terms of coherence.

10.1. Holism. A major concern in epistemology and philosophy of science concerns whether inference is holistic. According to Quine (1961, p. 41), "our statements about the external world face the tribunal of sense experience not individually but only as a corporate body." In a similar vein, Harman (1973, p. 159) writes that "inductive inference must be assessed with respect to everything one believes." Behind these holistic views is the antifoundationalist assumption that it is impossible to provide isolated justifications for isolated parts of our system of beliefs. Quine's position is based on his rejection of the analytic-synthetic distinction and on the view of Duhem (1954) that deducing an observation statement from a hypothesis always involves a complex of other hypotheses, so that no hypothesis can be evaluated in isolation. Because predictions are usually obtained from sets of hypotheses, observations that contradict the predictions do not provide grounds for rejecting any particular hypothesis, only for concluding that there is at least one false hypothesis. Harman argues that reasoning is inference to the best explanation, which includes both inference to hypotheses that explain the evidence and inference to what is explained. Inferential holism is therefore suggested by the following considerations:

(1) Hypotheses cannot be refuted and confirmed in isolation.

(2) Hypothesis evaluation must take into account the total sum of relevant evidence.

(3) The acceptability of a proposition is a function not only of what it explains but also of its being explained.

Unfortunately, holism brings many problems with it. Hegel (1967, p. 81) said that "the true is the whole," but he insisted that it should not be taken to be a crude, undifferentiated whole. If sets of hypotheses must be evaluated together, and everything is potentially relevant to everything else, how can we make a reasonable judgment about which hypotheses to maintain and which to reject? Kuhn's (1970) influential account of theory change as shifts in whole paradigms has been taken by some to imply that there is no rationality in science. Fodor (1983) has concluded from Quinean holism not only that philosophers have failed to provide a reasonable account of scientific confirmation, but even that cognitive science is unlikely ever to provide an account of such central psychological processes as hypothesis selection and problem solving (see Holland et al. 1986, Chap. 11, for a rebuttal).

My theory of explanatory coherence and its implementation in ECHO are holistic in that the acceptability of a hypothesis potentially depends on its relation to a whole complex of hypotheses and data. But there is nothing mystical about how ECHO uses pairwise relations of local coherence to come up with global coherence judgments. Although evidence units can be deactivated, just as data are sometimes ignored in scientific practice, the evidence principle gives some priority to the results of observation.

Although ECHO does not exhibit simplistic Popperian falsification, it need not succumb to the various strategies that can be used to save a hypothesis from refutation. The strongest direct evidence *against* a hypothesis is pointing out that it has implications that contradict what has been observed. One way of saving the hypothesis from an objection of this sort is to use an auxiliary hypothesis to explain away the negative evidence. Section 4.6 showed how simplicity considerations can prevent this strategem from working. Another way of saving a hypothesis in the face of negative evidence is to modify its cohypotheses. As Duhem and Quine pointed out, if H1 and H2 together imply some NE1 that contradicts a datum E1, then logic alone does not tell whether to reject H1, H2, or both. In ECHO, which hypotheses are deactivated depends on other relations of explanatory coherence. If H1 contributes to fewer explanations than H2, or if H1 contradicts another highly explanatory hypothesis, H3, then H1 will be more likely to be deactivated than H2.

Although ECHO makes it possible for a set of hypotheses to be accepted or rejected as a whole, it also admits the possibility of more piecemeal revision. Perrin (1988, p. 115) reports that the conversion of phlogiston theorists to the oxygen theory sometimes took several years, with the converts gradually accepting more and more of Lavoisier's views. ECHO's networks such as the oxygen-phlogiston one shown in Figure 9 do not connect everything to everything else. Explanatory relations may produce relatively isolated packets of coherent hypotheses and evidence; these may sometimes be accepted or rejected independent of the larger theory.

10.2. Probability. My account of theory evaluation contrasts sharply with probabilistic accounts of confirmation that have been influential in philosophy since Carnap (1950). Salmon (1966), for example, advocates the use of Bayes's theorem for theory evaluation, which, if P(H,E) stands for the probability of H given E, can be written as:

$$P(H,E) = \frac{P(H)P(E,H)}{P(E)}. \quad (2)$$

Consider what would be involved in trying to apply this to Lavoisier's argument against the phlogiston theory. We would have to take each hypothesis separately and calculate its probability given the evidence, but it is totally obscure how this could be done. Subjective probabilities understood as degrees of belief make sense in contexts where we can imagine people betting on expected outcomes, but scientific theory evaluation is not such a context. How could we take into account that alternative explanations are also being offered by the phlogiston theory? The issue is simplified somewhat if we consider only likelihood ratios for the oxygen and phlogiston theories – that is, the ratio of P(E, oxygen) to P(E, phlogiston) – but we still have the problem of dealing with the probability of the conjunction of a number of oxygen hypotheses whose degree of dependence is indeterminate. As I argued in section 8.2, probabilities have marginal relevance to qualitative explanatory inferences in science and law.

My account of coherence based on explanation contrasts markedly with probabilistic accounts. A set of propositions, S, is probabilistically coherent if there is a real-valued function that gives an assignment of values to the propositions consistent with the axioms of probability (Levi 1980). This constraint is very different from the ones governing explanatory coherence that ECHO shows to be sufficient for accepting and rejecting hypotheses. The real numbers that are degrees of activation of propositions in ECHO are clearly not probabilities, because they range from 1 to −1, like the certainty factors in the AI expert system MYCIN (Buchanan & Shortliffe 1984). Note that two contradictory propositions can both have activation greater than 0, if neither is substantially more coherent with the evidence than the other.

Probabilities, ranging from 0 to 1, are often interpreted as degrees of belief, but this interpretation obscures the natural distinction between acceptance and rejection, belief and disbelief. I do not just have low confidence in the proposition that the configuration of the stars and planets at birth affects human personality; I reject it as false. One advantage of probability theory, however, is that it provides rules for calculating the probabilities of conjunctions and disjunctions. In contrast, the acceptability (in my sense) of "P and Q" and "P or Q" is not defined, because such composite propositions do not, in general, figure in explanations. One can concoct cases of disjunctive explanations ("He said he was flying in from either New York or Philadelphia, and the weather is very bad in both places, so that explains why he's delayed"), but I have never encountered one in a scientific or legal context. Explanations depending on conjunctions of cohypotheses are common, but ECHO has no need to calculate the acceptability of "P and Q," because relations of explanatory coherence tell you all you need to know about P and Q individually. The apparent advantage of probability theory is much weakened in practice by the fact that the calculation of conjunctive and disjunctive probabilities requires knowledge of the extent to which the two propositions are independent of each other. Such information is easily gained when one is dealing with games of chance and in other cases where frequencies are available, but it is hard to come by in cases of scientific and legal reasoning: For Lavoisier, what was the conditional probability of OH1 (pure air contains oxygen principle) given OH2 (pure air contains matter of fire and heat)?

10.3. Confirmation theory. My view of theory evaluation based on explanatory coherence can also be contrasted with confirmation theory, according to which a hypothesis is confirmed by observed instances (Glymour 1980; Hempel 1965). The cases I discussed in detail in this paper are typical, I would argue, of the general practice in scientific argumentation that theories are not justified on the basis of particular observations that can be derived from them. Rather, observations are collected together into generalizations. These generalizations are sometimes rough, describing mere tendencies. In this process, particular observations can be tossed out as faulty or irrelevant. Theory evaluation starts with the explanation of the generalizations, not with particular observations. Lavoisier, for example, did not defend his theory by pointing to particular confirming observations such as his measurements indicating that a sample of burned phosphorus gained weight on a particular day in 1772. Rather, his central claim in defense of his theory is that it explains why objects in general gain weight when burned. Qualitative confirmation theory also does not in itself suggest how simplicity, analogy, competition, and being explained can play a role in theory evaluation.

10.4. Explanationism and conservatism. Now let me turn to a brief discussion of philosophical views that are much closer to my account of explanatory coherence. My discussion of hypotheses is compatible with the "explanationism" of Harman (1986) and Lycan (1988). Unlike them, however, I am not trying to give a general account of epistemic justification: I hold that there are other legitimate forms of inference besides inference to the best explanation. The principle of data priority assumes that results of observation start with a degree of acceptability that derives from their having been achieved by methods that lead reliably to true beliefs. This justification is closer to the "reliabilism" of Goldman (1986), a position which has problems, however, in justifying the acceptance of hypotheses (Thagard, in press a). My own view of justification is that explanation and truth are *both* epistemic goals that need to be taken into account as part of a larger process of justifying inferential strategies (Thagard 1988a, Chap. 7).

The other major difference I have with Harman and Lycan is that they both advocate *conservatism* as a supplement to considerations of explanatory coherence. Harman says we should try to maximize explanatory coherence while minimizing change. I view conservatism as a *consequence* of explanatory coherence, not as a separate factor in brief revision. In ECHO, we get a kind of conservatism about new evidence, as I showed in section 4.2. For ECHO, new evidence that does not cohere with what has been accepted is not treated equally with old evidence. In addition to conservatism about new evidence, there is a kind of conceptual conservatism inherent in any cognitive system: If an alternative theory requires a network of concepts which differs from my own, then I cannot evaluate the new system until I have effortfully acquired that system of concepts (Thagard, in press b).

Hence, an existing set of views will be conservatively favored until the alternative is fully developed.

The conservatism favored by Harman and Lycan seems most plausible, not for actual scientific cases, but for imagined ones in which a trivial variant of an accepted theory appears as an alternative. Suppose H1 gets high activation as the best explanation of E1 and E2, and then H2 is proposed to explain them both. If H1 and H2 are contradictory, then ECHO readjusts activation so that H1 and H2 are virtually at the same level. What if H2 is just a trivial variant of H1? Then H2 does not really contradict H1, so they can both be highly active without any problem. One might worry that the system will quickly be cluttered with trivial variants, but in a full computational system, that would be taken care of by having pragmatic constraints on what hypotheses are generated (Holland et al. 1986).

As a final comparison, consider the complementary views on explanatory unification of Kitcher (1981). He describes how powerful theories such as Darwin's and Newton's provide unification by applying similar *patterns* of explanation to various phenomena. That this should contribute to explanatory coherence is a consequence of my theory, for if H1, H2, and H3 are all used to explain the evidence, we get the result not just that each coheres with the evidence, but also that they cohere with each other. Moreover, degrees of coherence are cumulative, so that the more two hypotheses participate in explaining different pieces of evidence, the more they cohere with each other. (See the simple example in section 4.5.)

The major philosophical weakness of my account of explanatory coherence concerns the nature of explanation. This paper has bypassed the crucial question of what explanation is. Fortunately, to apply the principles of explanatory coherence and to generate input for ECHO, it is not necessary to have an exact analysis of the nature of explanation. We can take for granted the explanatory relations described by scientists such as Lavoisier and Darwin, or we can get an approximation using a computational system such as PI, as described in section 7. For an outline of what a computational account of explanation might look like, see Thagard (1988a, Chap. 3).

10.5. The descriptive and the normative. Philosophy differs from psychology primarily in its concern with normative matters – how people *ought* to reason rather than how they do reason. For some philosophers, any analysis that smacks of psychology has disqualified itself as epistemology. From this perspective, one faces the dichotomy: Is my theory of explanatory coherence normative or is it merely descriptive? In accord with Goldman (1986) and Harman (1986), I reject this rigid dichotomy, maintaining that descriptive matters are highly relevant to normative issues (see Thagard 1988a, Chap. 7). The seven principles of explanatory coherence are intended to capture both what people generally do and what they ought to do. By no means do they constitute a full theory of rationality. There are undoubtedly cases where people deviate from explanatory coherence – for example, preferring a hypothesis because it makes them happy rather than because of the evidence for it (Kunda 1987). Racial or other types of prejudice may prevent jurors from taking a piece of evidence seriously. Various other biases (Nisbett & Ross 1980) may intrude to throw off considerations of explanatory coherence. Much psychological experimentation and modeling is needed to show when people's reasoning can be accounted for in terms of explanatory coherence and when it is affected by other factors. This work can go hand in hand with refinements in the normative aspects of the theory.

11. Conclusion

I conclude with a brief survey of the chief accomplishments of the theory of explanatory coherence offered here.

First, it fits directly with the actual arguments of scientists such as Lavoisier and Darwin who explicitly discuss what competing theories explain. There is no need to postulate probabilities or contrive deductive relations. The theory and ECHO have engendered a far more detailed analysis of these arguments than is typically given by proponents of other accounts. Using the same principles, it applies to important cases of legal reasoning as well.

Second, unlike most accounts of theory evaluation, this view based on explanatory coherence is inherently comparative. If two hypotheses contradict each other, they incohere, so the subsystems of propositions to which they belong will compete with each other. As ECHO shows, successful subsystems of hypotheses and evidence can emerge gracefully from local judgments of explanatory coherence.

Third, the theory of explanatory coherence permits a smooth integration of diverse criteria such as explanatory breadth, simplicity, and analogy. ECHO's connectionist algorithm shows the computability of coherence relations. The success of the program is best attributed to the usefulness of connectionist architectures for achieving parallel constraint satisfaction, and to the fact that the problem inherent in inference to the best explanation is the need to satisfy multiple constraints simultaneously. Not all computational problems are best approached this way, but parallel constraint satisfaction has proven to be very powerful for other problems as well – for example, analogical mapping (Holyoak & Thagard, in press).

Finally, my theory surmounts the problem of holism. The principles of explanatory coherence establish pairwise relations of coherence between propositions in an explanatory system. Thanks to ECHO, we know that there is an efficient algorithm for adjusting a system of propositions to turn coherence relations into judgments of acceptability. The algorithm allows every proposition to influence every other one, because there is typically a path of links between any two units, but the influences are set up systematically to reflect explanatory relations. Theory assessment is done as a whole, but a theory does not have to be rejected or accepted as a whole. Those hypotheses that participate in many explanations will be much more coherent with the evidence, and with each other, and will therefore be harder to reject. More peripheral hypotheses may be deactivated even if the rest of the theory they are linked to wins. We thus get a holistic account of inference that can nevertheless differentiate between strong and weak hypotheses. Although our hypotheses face evidence only as a corporate body, evidence and relations of explanatory coherence suffice to separate good hypotheses from bad.

Table 9. *Algorithms for processing input to* ECHO

1. Input: (PROPOSITION NAME SENTENCE)
   Create a unit called NAME and an index for it.
   Store SENTENCE with NAME.
2. Input: (EXPLAIN LIST-OF-PROPOSITIONS PROPOSITION)
   Make excitatory links[a] between each member of LIST-OF-PROPOSITIONS and PROPOSITIONS.
   Make excitatory links[a] between each pair of LIST-OF-PROPOSITIONS.
   Record what explains what.
3. Input: (CONTRADICT PROPOSITON-1 PROPOSITION-2)
   Make an inhibitory link between PROPOSITION-1 and PROPOSITION-2.
4. Input: (DATA LIST-OF-PROPOSITIONS)
   For each member of LIST-OF-PROPOSITIONS, create an excitatory link from the special evidence unit with the weight equal to the data excitation parameter, unless the member is itself a list of the form (PROPOSITION WEIGHT). In this case, the weight of the excitatory link between the special unit and PROPOSITION is WEIGHT.
   If there are unexplained data propositions, increase the decay rate parameter by multiplying it by the ratio of the total number of evidence proppositions to the number of explained evidence propositions.

[a]The weights on these links are determined by equation 3 given in the text. Weights are additive: If more than one EXPLAIN statement creates a link between two proposition units, then the weight on the link is the sum of the weights suggested by both statements.

12. APPENDIX

Technical details of ECHO

For those interested in a more technical description of how ECHO works, this appendix outlines its principle algorithms and describes sensitivity analyses that have been done to determine the effects of the various parameters on ECHO's performance.

12.1. Algorithms. As I described in section 4.1, ECHO takes as input PROPOSITION, EXPLAIN, CONTRADICT, and DATA statements. The basic data structures in ECHO are LISP atoms that implement units with property lists that contain information about connections and the weights of the links between units. Table 9 describes the effects of the four main kinds of input statements. All are very straightforward, although the

Table 10. *Algorithms for network operation*

1. Running the network:
   Set all unit activations to an initial starting value (typically .01), except that the special evidence unit is clamped at 1.
   Update activations in accordance with (2) below.
   If no unit has changed activation more than a specified amount (usually .001), or if a specified number of cycles of updating have occurred, then stop.
   Print out the activation values of all units.
2. Synchronous activation updating at each cycle:
   For each unit u, calculate the new activation u in accord with equations 3 and 4 in the text, considering the old activation of each unit u' linked to u.
   Set the activation of u to the new activation.

EXPLAIN statements require a calculation of the weights on the excitatory links. The equation for this is:

$$weight(P,Q) = default\ weight\ /\ (number\ of\ cohypotheses\ of\ P)^{(simplicity\ impact)} \quad (3)$$

Here simplicity impact is an exponent, so that increasing it lowers the weight even more, putting a still greater penalty on the use of multiple assumptions in an explanation. In practice, however, I have not found any examples where it was interesting to set simplicity impact at a value other than 1.

After input has been used to set up the network, the network is run in cycles that synchronously update all the units. The basic algorithm for this is shown in Table 10. For each unit j, the activation a_j, ranging from -1 to 1, is a continuous function of the activation of all the units linked to it, with each unit's contribution depending on the weight w_{ij} of the link from unit i to unit j. The activation of a unit j is updated using the following equation:

$$a_j(t+1) = a_j(t)(1-\theta) + \begin{cases} net_j(max - a_j(t)) & \text{if } net_j > 0 \\ net_j(a_j(t) - min) & otherwise \end{cases} \quad (4)$$

Here θ is a decay parameter that decrements each unit at every cycle, min is minimum activation (-1), max is maximum activation (1), and net_j is the net input to a unit. This is defined by

$$net_j = \Sigma_i w_{ij} a_i(t) \quad (5)$$

Repeated updating cycles result in some units becoming activated (getting activation > 0) while others become deactivated (activation < 0).

12.2. Sensitivity analyses. Multiple connected localist networks sometimes exhibit instability, failing to settle into stable activation patterns because complexes of mutually excitatory units produce activation oscillations. As Figures 11, 14, 17, and 20 suggest, ECHO's networks are generally stable, usually requiring

Table 11. *Network information for four major examples*

	Units	Links	Cycles to settle	Excitation ceiling	Inhibition floor
Lavoisier	20	49	107	.13	−.18
Darwin	29	70	49	.06	−.16
Chambers	34	59	63	.17	−.07
Peyer	34	54	78	.08	−.13

fewer than 100 units of updating for all units to reach asymptotic levels. Pearl (1987) devotes considerable effort to rearranging probabilistic networks so that they will be singly connected and hence stable. Fortunately, the networks set up by ECHO in accord with the theory of explanatory coherence do not require any alteration to settle into stable activations. Whereas a probabilistic network may need links specifying the conditional probabilities of p given q, q given r, r given s, and s given p, such cyclic paths rarely arise in ECHO because the "explain" relation sets up hierarchies of units rather than cycles. ECHO undergoes activation oscillations only when the excitation parameter is high relative to inhibition, for example, in the Chambers case, if excitation has a value of .17 instead of .05. ECHO is efficient: In each of the four major examples, a complete run, including network creation and settling, takes less than a minute of cpu time on a Sun 3/75 workstation. Because networks with hundreds more units and thousands more links than ECHO's networks have run successfully in ACME, a similar program that does analogical mapping (Holyoak & Thagard, in press), I see no problem in scaling ECHO up to run on much larger examples.

Table 11 shows, for each major example, the size of the networks created and the number of cycles of activation updating it takes for them to settle using the default parameter values of .05 for excitation, −.2 for inhibition, .1 for data excitation, and .05 for decay. Experiments have shown that ECHO exhibits the behavior described in the text over a wide range of values for these parameters. For example, in the Lavoisier example, no important differences in the results occur if the decay, excitation, inhibition, and data excitation parameters are all halved or doubled. In general, lowering positive parameters and making inhibition closer to 0 tends to prolong settling time. Increasing decay tends to flatten the activation curves, both positive and negative, keeping them closer to 0. Increasing data excitation leads evidence units to have higher asymptotic activation. Varying excitation and inhibition systematically reveals that there is a critical value for each. If excitation is high relative to inhibition, then the system shows much "tolerance" and does not deactivate inferior hypotheses. Table 11 lists excitation ceilings and inhibition floors for the four major examples. The excitation ceilings are the maximum values that excitation can have without activating units representing inferior hypotheses; inhibition here is constant at the default value of −.2. The excitation values at which networks become unstable are well above these ceilings. The inhibition floors are the minimum values that inhibition must have without failing to deactivate units representing inferior hypotheses; excitation here is constant at the default value of .05. The excitation ceiling and the inhibition floor indicate the most important respects in which quantitative parameter changes in ECHO have qualitative effects. Keep in mind that the excitation ceilings and inhibition floors listed in Table 11 are based on a fixed value for, respectively, inhibition and excitation. Varying these values will produce different floors and ceilings, so that the range of possible parameter values is much larger than Table 11 portrays.

ACKNOWLEDGMENTS

The development of the ideas in this paper has benefited from discussions with Gilbert Harman, Michael Ranney, Gregory Nowak, Frank Doring, and other members of a discussion group on explanatory coherence. For ideas about connectionist models, I am indebted to Keith Holyoak and Stephen Hanson. I am grateful to Ziva Kunda, Dan Hausman, Phil Johnson-Laird, Robert McCauley, Yorick Wilks, William Bechtel, James Hendler, Robin Dawes, Paul O'Rorke, and several anonymous referees for helpful comments on previous drafts. Thanks to Greg Nelson for the graphics program used for some of the figures, and to Robert McLean for writing a C version of ECHO. (LISP and C versions of ECHO are available on request.) This research was supported by a grant from the James S. McDonnell Foundation to Princeton University and by a contract from the Basic Research Office of the Army Research Institute for the Behavioral and Social Sciences.

NOTES

1. From here on, I shall be less careful about distinguishing between units and the propositions they represent.
2. I owe this suggestion to Daniel Kimberg.

Open Peer Commentary

Commentaries submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Explanation and acceptability

Peter Achinstein

Department of Philosophy, Johns Hopkins University, Baltimore, MD 21218

Thagard proposes a theory of explanatory coherence that is based on what he admits to be a primitive, undefined concept of explanation. It is an essential part of this theory that explanatory coherence is intimately tied to acceptability. ("We should accept propositions that are [explanatorily] coherent with our other beliefs.") My challenge is this: Can such a theory work and be illuminating if explanation remains undefined?

Let us consider two ways one might construe Thagard's explanation sentences of the form "P explains Q." First, they might be construed simply as proposed explanations. For example, we say that the *Book of Genesis* explains the origin of the universe. In saying this, we do not necessarily imply that the explanation is good, or correct, or that it even meets minimal standards that we have for explanations. We may simply mean that it has been proposed by those who believe these things. (For a definition of this nonevaluative sense of "explain," see Achinstein 1983, Chapters 2 and 3.) I doubt that Thagard has this sense of explanation in mind, because it bears no obvious connection to acceptability.

Second, let us shift to an evaluative sense of explanation. Which one? This question needs to be raised because of a standard objection to explanatory accounts of acceptability. Suppose that we have a set of observed data and a hypothesis h that explains all of them. The objection to concluding that h is acceptable on such grounds is that some incompatible hypothesis h′, which also explains the observed data, will usually be constructible. If so, then, unless h has some independent support, h is not acceptable, or at least no more so than h′.

One standard form of explanation found particularly in the quantitative sciences involves deductive derivation. Suppose that data O_1, O_2, . . . are deductively derivable from hypothesis h together with background information b. The following probability theorem is provable: If h has at least one incompatible competitor h′ that together with b also entails O_1, O_2, . . . , and is such that $p(h'/b) \geq p(h/b)$, then for any n, no matter how large, $p(h/O_1, \ldots O_n \,\&\, b) \leq .5$. That is, if h has a competitor that entails the same data as h, and whose probability on the background information alone is at least as high as that of h, then the probability of h will not rise above ½, no matter how many data h entails. This holds even if scientists are unaware of the competitor h′. If the acceptability of a hypothesis requires a probability greater than ½, then, in such circumstances, h is not acceptable despite its success in entailing all the data and

despite the fact that proponents of h are unable to think of any competitor to h that entails all the data. Nor is ½ sacred here. We can show, for example, that if there is such a competitor to h whose probability on b is at least .8, then h's probability cannot rise above .2.

Indeed, if h entails each of the Os, then it doesn't necessarily follow even that h's probability increases. If each of the Os is "old" evidence, known with certainty to be true, then h's probability remains constant. It increases if and only if the Os are new phenomena whose probability is less than 1. The rub is that although h's probability rises as it entails more and more new phenomena, it may forever remain extremely low, much lower than that of a competitor, and fail to approach anything in the acceptable range.

Thagard is dubious about assigning probabilities to scientific hypotheses, and in any case, he claims, rejection is different from low probability. I don't find these objections decisive. The competitive hypothesis theorem does not require that we be able to assign a precise probability to h or its competitor h'. It assumes only that h's probability on the background information, whatever it is, is no greater than that of h'. In that case, whatever h's probability on the data is, it cannot get very high. Furthermore, although low probability is not necessarily the same as rejection, it is possible to give rules of acceptance and rejection that are based on probabilities (see Levi 1967). Thagard needs to demonstrate that there is some connection between his notion of explanation and acceptability. But without some account of explanation (and of acceptability), I don't see that any connection is guaranteed. Thagard cannot simply assume–as he does in his Principle 6–that there is some reasonable concept of explanation that insures acceptability.

As far as explanation by deduction is concerned, Thagard makes it clear that, for other reasons, he "assume(s) that explanation is more restrictive than implication." But (assuming he does not deny that many explanations do involve deductive derivations), what additional conditions is he willing to impose? This question is crucial because not all plausible conditions will thwart the previous probability theorem. An important nineteenth-century proponent of an explanatory view of acceptability was William Whewell. He regarded the wave theory of light as acceptable not simply because it afforded derivational explanations of various observed optical phenomena, but because (a) the phenomena derived were not all of the same type (what he called "consilience"), and (b) the hypotheses of the theory that provided the explanation "run together" (as he put it) in a way that goes beyond simply explaining the data. Yet it is possible to show that such additional conditions–though considerably stronger than those imposed by the standard deductive model of explanation–are not sufficient to dodge the effects of the "competitor" probability theorem. (For details, see Achinstein, forthcoming.)

Finally, there are, to be sure, evaluative conditions on explanations that will insure the acceptability of the explanatory hypothesis–for example, require that the explanatory hypothesis be true, or that it be highly probable given all the observed data (as does Hempel's, 1965, standard deductive model of explanation). The problem is that if these are the requirements added to derivability (or whatever else explanations are supposed to exhibit), then explanation becomes redundant. If what we care about is h's acceptability, then h's truth or high probability will insure that. Why bring in explanation?

When weak explanations prevail

Carl Bereiter and Marlene Scardamalia

Centre for Applied Cognitive Science, Ontario Institute for Studies in Education, Toronto, Canada M5S 1V6

Electronic mail: *c_bereiter@utoroise.bitnet and scardamalia@utoroise.bitnet*

Thagard presents a psychologically interesting theory implemented in a computer model so straightforward that other investigators can readily test its applications and limitations and even fiddle with its procedures. Our comments are stimulated by initial attempts to use ECHO to analyze arguments and as an educational device for application with children.

Explanatory coherence shows us how strong theories win out over weak ones, even though they are more vulnerable to troublesome facts. If we examine instances where weak theories survive against stronger ones, however, we see a kind of argument that is not represented in Thagard's theory, although it seems to play a predominant role in everyday thinking. Thagard presents a model of argument to the best explanation. The issue from this standpoint is, Does the hypothesis explain the facts? Bartlett (1958) concluded from his studies of everyday thinking that people are not much concerned with accounting for facts. Instead, they settle quickly on a belief and retain it as long as facts more or less support it. For them the issue is, Do the facts support the hypothesis?

This latter kind of thinking has been prominent in the last two decades of debate about the heritability of intelligence. In this debate, a relatively weak theory, environmentalism, has fared very well both in scientific circles and in popular opinion against a stronger heredity-plus-environment theory that purports to explain the environmentalist facts plus a number of others–such as the magnitudes of various kinship correlations, regressions to the mean, and within-family variability of IQ. (See Urbach, 1974,) for an analysis of this controversy from a Lakatosian perspective.) The main thrust of environmentalist arguments has not been defending the explanatory power of environmentalist hypotheses but attacking the quality of evidence brought forth by hereditarians (Kamin 1974). It may be noted that present-day creationists are gathering a following by using the same kind of argument. Within cognitive science, a somewhat similar argument develops when people criticize the validity of thinking-aloud protocol data. And, of course, attacking the evidence is a major form of argument in court trials.

These examples have in common a view of hypotheses as being *upheld* by data. Environmentalism and creationism win because their factual claims are unassailable, being mostly common knowledge, whereas those of the opposing theories are contestable. This is a radically different view from that taken by Thagard. If social scientists were to approach the heredity/environment issue in a manner consistent with Thagard's theory, they would begin by agreeing that there are facts needing explanation–that so and so reported such and such correlation between the IQs of double-first cousins, and so forth. An argument like Kamin's, which produces a different explanation for every fact, would fare poorly against an argument that explains all the facts from a few coherent hypotheses.

Are we describing a kind of bad thinking that Thagard's theory ought to help overcome, or is there some merit in this alternative approach to explanation? Both, we think, are true. In a "mature" scientific controversy, irrelevancies have been shaken out, and there remains a set of mutually recognized facts that need explaining. In such a situation, the stronger theory–the one best able to account for the recognized facts–ought to prevail. That is a situation that Thagard's theory, implemented in ECHO, seems to handle nicely. In the murkier situations of ordinary life and the soft sciences, however, it is often uncertain whether the facts in need of explanation have been properly identified or are to be trusted. In such cases, there may be reasons why a weaker theory should prevail.

We have been investigating explanatory hypotheses in a case of special interest to Canadians, that of Ben Johnson, who was stripped of an Olympic gold medal for alleged use of anabolic steroids. The facts that need explaining consist mainly of laboratory test results indicating steroid use. There are a number of other facts, however, that figure in some explanations–for instance, that Johnson customarily drank sarsaparilla tea after a race, that the bag containing his flask of tea was unguarded during the race, and that strangers were seen in its vicinity. We may call these *contextual* facts. These facts themselves do not need explaining; there is nothing "suspicious" about them. Nevertheless, an explanation gains plausibility if it weaves these facts into its story. A simple "spiked sarsaparilla" theory fails because of its inability to account for laboratory results indicating long-term steroid use. But a more complex theory–which has Johnson taking steroids plus a masking drug, and enemies spiking his sarsaparilla in order to defeat the masking drug–starts to sound like a contender because it not only accounts for the critical facts but also incorporates a variety of contextual facts.

In criminal cases, contextual facts are typically used in arguing motive, opportunity, and disposition to commit the crime. None of these are vital issues if there is only one explanation that satisfactorily accounts for the "suspicious" facts–that is, the facts recognized as needing explanation. But when alternative explanations are tenable, the one that makes better use of the surrounding contextual facts is rightly to be preferred. Similarly, the environmentalist explanation of IQ differences gains strength because it weaves in many contextual facts about cultural differences, social conditions, and historical antecedents, whereas hereditarian hypotheses deal with little besides kinship data and test scores. Some environmentalists have woven in historical facts to support a conspiracy theory about hereditarians, thus casting a general cloud of suspicion over the hereditarians' factual claims–again, a common courtroom strategy, but one that in its own way contributes greatly to the coherence of an argument.

Although Thagard's theory does not deal with contextual facts, it seems that his ECHO program is quite happy to accommodate them. Perhaps contextual facts should receive less sustaining activation than facts needing explanation, and perhaps the connection weight between hypotheses and facts that are merely used should be less than the weight between hypotheses and facts that they explain. However, in our limited experiments with Thagard's examples and in our Ben Johnson case, there does not seem to be any need to modify activation levels or weights. Sensible results are obtained by treating all facts as equal and all positive connections as equal. For reasons to be given later, however, we do not regard this as a good sign.

In another respect ECHO seems to be more limited than the theory it implements. To settle on a winner, ECHO needs contradictions, which enter the network as negative connection weights. Again, this is not a problem in a "mature" scientific controversy, where basic disagreements have been identified. But in many ambiguous or undeveloped areas of inquiry, there may be competing explanations that do not clearly conflict. They may occupy different levels of description, for instance. In such cases, ECHO can lead to unfortunate results. This difficulty, incidentally, was discovered by a group of 11-year-olds who were using ECHO to test their explanations in the Ben Johnson case. It is more easily illustrated, however, with Thagard's Darwin example.

Suppose we enter with another hypothesis into the set of propositions constituting the Darwin case. Call it the Satanic hypothesis:

The Devil is responsible for differences.

This hypothesis contradicts the creationist hypothesis, but it does not contradict any Darwinian hypotheses. It doesn't explain any of the evidence, but on the other hand, it isn't incompatible with any of the evidence either. When tested in competition with the Darwinian and creationist hypotheses, ECHO gives it a final activation level of .67, which puts it ahead of three of the five Darwinian hypotheses. It gains this status solely by virtue of contradicting a hypothesis that is defeated by other hypotheses.

There seems to be two ways, then, in which Thagard's theory and its ECHO implementation need to be augmented to deal with a larger range of explanatory problems. There needs to be a way to compare competing but noncontradictory hypotheses, especially hypotheses at different levels of description. And there needs to be a way to attribute coherence both to the explanation of facts by hypotheses and to the use of contextual facts in explanations. However, the theory ought to be able to distinguish between the two. Otherwise, elaborate stories that fail to account for crucial facts will tend to defeat incisive theories that explain crucial facts without reference to contextual information. As educational researchers, we are particularly keen on the use of ECHO as a way of guiding students toward argument to the best explanation and away from the weaker kinds of explanation that seem more prevalent. We were amazed in our initial trials to find elementary school children taking naturally to questions such as, "What hypotheses explain this fact?" and, "What facts does this hypothesis explain?" They quickly caught on to what ECHO was doing, saw it as reasonable, and became interested in experimenting with the effects of suppressing certain facts or introducing new hypotheses. Perhaps, as Bartlett concluded, human beings do not usually think this way. But we see reason to hope that they could learn to do so.

Explanatory coherence as a psychological theory

P. C.-H. Cheng and M. Keane

Human Cognition Research Laboratory, The Open University, Milton Keynes MK7 6AA, England

Electronic mail: *pch_cheng@vax.acs.ou.ac.uk and mt_keane@vax.acs.ou.ac.uk*

If Thagard's theory is to be viewed as a psychological theory, its principles need to be amended considerably. Furthermore, the need for such amendments suggests that a purely parallel model may not be optimal. Two main problems are evident from the psychological perspective.

First, any psychological theory must acknowledge human processing limitations. People are unlikely to have ECHO's unlimited processing power to consider all of the interdependencies between a theory's propositions and the evidence. For instance, the jurors empaneled at a fraud trial will likely find the "propositions" involved difficult to evaluate because of the introduction of new concepts, the large quantity of evidence, and their interrelationships. A more realistic psychological account of theory evaluation would hence be one in which new propositions and evidence are gradually assimilated in a piecemeal fashion. Such a view of theory evaluation is supported by the research of the *new experimentalists* (Ackermann 1985; Franklin 1986; Galison 1987; Hacking 1983) in the area of the philosophy of science, which contrasts with the *holism* espoused by Thagard. However, to achieve this sort of piecemeal evaluation, ECHO would have to break the network up into smaller subsets of the complete set of propositions and evidence and to operate on those subsets in a more serial fashion. This introduces several problems: 1. It seems unlikely that the combination of smaller subsets would aggregate to produce the same result as the complete set of propositions processed in parallel (because of the nature of such connectionist models). 2. There is also the attendant question of how the different subsets might be combined. 3. A further set of processes would be required to

select subsets for consideration, and these would have to be specified by the theory.

Second, in the model, explanation is instantiated as links between the various units in the network. However, because the psychological processes that underlie explanation are not specified, one could just as well have substituted "connected to" or "associated with" for the term "explains" throughout the article and the model would not need to be changed. An adequate psychological theory must specify the processes involved in "explaining." In certain sciences, researchers spend a significant amount of time determining whether a proposition really *explains* something, weighing up strong and weak senses of a proposition, and attempting to separate ancillary, *ad hoc* proposals from the basic tenets of a theory. Thagard hints at this problem with the concept of explanation when he says that a causal sense of "explain" was used in the analyses, but this is inconsistent with his initial statement that the theory should be a theory of any form of explanation. From a modelling perspective, the determination of what it means to explain and what constitutes a theoretical proposition have all the hallmarks of heuristic, evaluative processes. Although such processes could be modelled in a connectionist fashion, a traditional symbolic treatment seems more directly applicable.

In conclusion, a psychological theory would require considerable additions not provided for in Thagard's theory. The nature of these changes also recommends a conventional symbolic model rather than a connectionist one.

Assimilating evidence: The key to revision?

Michelene T. H. Chi

Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA 15260

Electronic mail: *micki%olrdca@vms.cis.pittsburgh.edu*

Thagard's theory of explanatory coherence has several exciting and profound applications in psychology. Two crucial but unresolved issues in psychology are: (1) How does conceptual change occur? and (2) What kind of "transition mechanism" accounts for these changes? These two questions most typically arise in the domains of learning and development. In learning, one manifestation of this issue concerns the transition from holding a naive theory of the physical world to holding a scientific or Newtonian view. In development, the issue concerns the transition from one stage (such as preoperational) to another, more advanced stage of thought (such as concrete operational). It is commonly thought that the shift from one kind of thought to another, either from preoperational to operational, or from pre-Newtonian to Newtonian, depends on the adoption of a set of interrelated beliefs. (This wholesale adoption is sometimes called radical restructuring.) The dilemma has always been trying to identify the "mechanisms" that enabled this transition to take place.

The most promising aspect of Thagard's theory is that it could potentially uncover precisely what factors can contribute to restructuring (or to conceptual change) without postulating an explicit mechanism that is responsible for the transition. That is, by implementing ECHO in a connectionist framework with parallel constraint satisfaction, the model has the capability of settling into a state "naturally," thereby achieving restructuring without identifying specific mechanisms for it. Thus, in some sense, ECHO has bypassed the problem of identifying the "transition mechanism" that has puzzled psychologists for decades. The implication of ECHO is that manipulating a few coherence relations in a piecemeal way might in fact produce dramatic shifts in one's theoretical orientation or frame of thought.

Although ECHO has this potential, what has ECHO accomplished so far? To understand what could have caused the transition (i.e., to understand what caused one theory to be more coherent than another), Thagard needs to model conceptual transitions directly. This is almost an impossible task in the historical context, somewhat less difficult in the developmental context, but perhaps feasible in a learning context. Thagard has attempted to model such a transition in the learning case by modeling the belief revisions that a student underwent in explaining the trajectories of projectiles while offering predictions and receiving feedback. In the data cited in Ranney and Thagard (1988), the shift exhibited by subject S.P.I. from a non-Newtonian to a Newtonian framework occurred primarily from encoding new evidence that either confirmed existing Newtonian hypotheses or contradicted existing non-Newtonian hypotheses. The advantage of this demonstration is that Ranney and Thagard could model the shift in conceptual change without postulating the formulation of new hypotheses, as was necessary in the historical cases (for example, Lavoisier had many new hypotheses that were not entertained by Stahl). This is fortunate because the mechanism by which new hypotheses are formulated is as yet little understood, as Thagard knows. What appears to have caused a shift from pre-Newtonian to Newtonian conceptions is modeled as the occurrence of new evidence either provided by the experimenter, or entertained by the student, evidence that either confirmed or contradicted the student's existing hypotheses (the student initially had both Newtonian and non-Newtonian hypotheses). Thus, in general, Thagard's applications of his theory to the learning domain, as well as to historical cases, point to two critical mechanisms needed for restructuring: the acquisition or formulation of new hypotheses and the encoding or entertaining of new evidence.

Unfortunately for ECHO's plausibility as a model of human performance, the majority of psychological evidence regarding conceptual change contradicts an implicit assumption underlying ECHO's analyses of Ranney's data. It is often found (in Piagetian research, for example) that confronting a child with evidence that contradicts the child's hypothesis usually does not lead to the child's rejection of that hypothesis. This suggests that the crux of the matter may not lie in the straightforward provision of new evidence, as modeled in Ranney and Thagard. Rather, the crucial insight for the subject is to realize that a particular hypothesis explains a particular piece of evidence. (This is the simplest case; I will avoid adding qualifications to all other cases, such as realizing that two hypotheses are contradictory, and so on.) Hence, what underlies theory revision may be precisely the willingness to adopt or reject a belief that a particular piece of evidence is explainable by a specific hypothesis. (This is currently built into ECHO as a given.) This willingness may in turn depend on the representation of a subject's current conception.

There is another issue that is relevant to psychology: What does ECHO's network of explanations represent? By modeling ECHO in a connectionist framework, Thagard is implying that the connectivity per se ought to inform the person whose mind the network embodies that a particular theory is more or less coherent than another. Presumably, if the network of Lavoisier's explanations is an accurate reflection of his memory, then it is not surprising that Lavoisier is convinced that his theory is the current one. This raises an interesting dilemma, however: Two contemporaneous theorists who hold opposing views would presumably know about each other's hypotheses as well as the evidence that each theory's hypotheses would explain. And yet, two contemporaneous theorists would not come to the same evaluation of their respective theories, as predicted from ECHO. This means that their representations must be different somehow. How they might differ can easily be seen in the arguments entered into by the theorists. Many of these arguments question the assumption that a particular hypothesis

explains a particular piece of evidence. This goes back to the previous point that the critical "insight" is the willingness to assimilate into one's representation that a particular piece of evidence is explained by a particular hypothesis. As psychological evidence shows, one is unwilling to encode a piece of evidence and its interpretation if it conflicts with one's existing hypotheses. Thus, we have cycled back to the original question of how exactly individuals revise their initial sets of beliefs in a significant way.

One other issue relating to ECHO's feasibility as a human model concerns ECHO's exhibition of apparently superhuman capabilities. In the behavioral decision-making literature, it has consistently been found that simple linear combinations of evidence are better at predicting outcomes such as success in graduate school (Dawes 1971) or the severity of Hodgkin's disease (Einhorn 1972) than human experts (e.g., physicians in the case of diagnosing Hodgkin's disease). The usual interpretation of these data is that humans excel at evaluating individual pieces of evidence with respect to a hypothesis but are extremely poor at integrating multiple pieces of evidence. The same superhuman reasoning ability may be exhibited by ECHO in that it can resolve two discrepant views given all their explanatory links, whereas humans only evaluate each individual explanation. The psychological community anxiously awaits further empirical tests to clarify these important issues.

Two problems for the explanatory coherence theory of acceptability

L. Jonathan Cohen

The Queen's College, Oxford University, Oxford OX1 4AW, England

Thagard's analysis of reasoning about acceptability is an interesting new contribution to the field. However, it fails to meet at least two requirements that any such analysis should aim to satisfy.

1. Consider a situation in which a known fact, E, needs to be explained and two rival hypotheses, H_1 and H_2, are proposed for the task. Suppose that H_1 explains E and also another known fact, F. Suppose that H_2 explains E and also predicts a hitherto unknown fact, P; and suppose too that this prediction is observationally or experimentally confirmed and that H_2 also explains P. In that case, Thagard's system would allow H_1 and H_2 to have equal acceptability. But in the history of science, most researchers have been inclined to attach greater value, other things being equal, to a hypothesis that generates new knowledge than to one that merely explains what we already know (Bacon 1859; Lakatos 1970; Leibniz 1865). Good scientific ideas have heuristic as well as explanatory power. They look to the future as well as to the past. This feature is reflected in any Bayesian analysis of reasoning about the evaluation of hypotheses, because p(H/E) increases, other things being equal, as p(E) decreases. It is also reflected in the Baconian method of relevant variables (Cohen 1989, p. 152). However Thagard's analysis, however, makes no allowance for the merit of predictive novelty, and, if widely adopted, would distort the evaluation of scientific hypotheses in a way that might be seriously detrimental to the progress of human enquiry. Of course, Thagard could tack on an eighth principle that would attach appropriate value to predictive novelty alongside explanatory coherence, but this would be an *ad hoc* modification of his theory, whereas the merit of predictive novelty is an integral consequence of both Bayesian and Baconian analyses. By Thagard's own standard of simplicity, therefore, a Bayesian or Baconian analysis is preferable in this respect.

2. Another feature of both Bayesian and Baconian accounts is that they offer a systematic, logical syntax for evaluating hypotheses, because the former is tied to the mathematical calculus of chance and the latter to a generalised modal logic (Cohen 1989). It is thus possible in both systems to infer, for instance, the degree of acceptability of a conjunction of two independent hypotheses from the respective degrees of acceptability that each has on its own, or to infer that where H_2 is the disjunction of H_1 with some other proposition, H_2's degree of acceptability must be at least as great as that of H_1.

In Thagard's theory, however, as he himself points out, the acceptability of "P and Q" or of "P or Q" is not defined. So, in general, no such inferences are possible, and Thagard argues that this does not matter. ECHO, he says, has no need to calculate the acceptability of "P and Q," because relations of explanatory coherence tell you all you need to know about P and Q individually. Clearly this assumes that acceptability is of interest only in relation to single hypotheses. It is as if the ultimate purpose of research were to provide a list of individual hypotheses with high acceptability values. Though such a program might conceivably satisfy those who have a certain kind of purely intellectual interest in science, however, it falls far short of what the practical interests of technology require. When you are building a plane, for example, you rely on many more than just one hypothesis, and the acceptability value of the conjunction of these hypotheses is very much at issue.

Thagard also claims that his measure of acceptability is applicable to forensic reasoning about matters of fact. Yet on that topic there has been extensive discussion in recent years–in the literatures of jurisprudence, philosophy, and statistics–about how a conjunction's degree of acceptability relates to the degrees of acceptability of its several conjuncts (e.g., Allen 1986; Cohen 1977; Dawid 1987; Eggleston 1983; Kaye 1986; Schum 1986; Williams 1979). There is a serious problem about whether a Bayesian measure can be applied in such cases, or whether a Baconian one is needed; the problem arises in regard to both the criminal standard of proof (proof beyond reasonable doubt) and the civil standard (proof on the preponderance of evidence). Thagard's theory of explanatory coherence, however, is not even a candidate for consideration as a measure of acceptability here, because it allows no application to the problem.

An example will make the point clearer. Imagine a civil case against an insurance company in which the plaintiff has to prove two independent points–that he has a paid-up automobile insurance policy of a certain kind and that his accident was due to such and such circumstances. Suppose he proves each point with a probability of .6. Apparently he has proved his case as a whole with a mathematical probability of only .36, and yet the legal standard of proof may seem to require a probability of more than .5 for him to win. Well, perhaps there are ways for a Bayesian analysis to get around this difficulty, or perhaps the Bayesian analysis should be replaced by one in terms of Baconian probability. But at least we have to take seriously the problem of how to evaluate the acceptability of a conjunction in relation to the acceptability of its conjuncts. It will not do to imply, as Thagard implies, that the problem does not exist.

I am not claiming that Thagard's theory of acceptability has no valid areas of application. My point is only that it clearly fails to do justice to two kinds of context in which evaluations of acceptability are important in our culture–namely, the heuristic dimension of evaluation in science and the evaluation of conjunctions in technology and the courts. There are, of course, trade-offs to be calculated in relation to any measure of acceptability. But it looks as though Thagard's system is inferior in important respects both to Bayesian and to Baconian measures.

Thagard's Principle 7 and Simpson's paradox

Robyn M. Dawes
Department of Social Sciences, Carnegie-Mellon University, Pittsburgh, PA 15213
Electronic mail: *rd1b@andrew.cmu.edu*

Although I share Thagard's admiration for the work of Pennington and Hastie (1987; 1988), I interpret it somewhat differently. Thagard argues that they show that a jury's verdict depends on the "explanatory coherence" of the prosecution's story compared to that of the defense. I believe they do more. Their "story model" of jury verdict is not just a descriptive one, in which they can argue *post hoc* that coherence has been achieved. Rather, it has the strong implication that the *order* in which evidence is presented will influence the verdict.

Just as Thagard does not present a theory of explanation, Pennington and Hastie do not present a theory of what constitutes a "good story." They do note, however, that the order in which evidence is presented can affect the "goodness" of a story, and in their work they show that the order of the evidence does indeed affect the verdict. In contrast, Bayesian analyses of jury verdicts (and other conclusions) are independent of the order in which evidence is obtained and presented. (The final posterior odds comparing two hypotheses consist of the ratio of the probability of the intersection of all the evidence and one hypothesis divided by the probability of the intersection of all the evidence and the other hypothesis; intersection is commutative.) In addition to being inconsistent with a Bayesian analysis, Hastie and Pennington's conclusion that order has an effect is compatible with our experience in forming judgments.

Thagard does have an analogous "strong implication" in his model. Specifically, the coherence of a set of propositions is dependent on binary coherence (and in addition, the coherence of a single proposition depends in turn on the set in which it is embedded; section 2.1). That implication also conflicts with other analyses, such as *all* probabilistic ones. The reason is that such analyses allow the possibility of a Simpson's paradox reversal in the relation between evidence and hypotheses, whereas pairwise analysis does not.

This paradox is illustrated in Table 1. Principle 7 is in section 2.1 of Thagard's target article. The entries are compound probabilities, involving two equally likely hypotheses, H_1 and H_2, and two bits of evidence, e_1 and e_2. H_1 is more probable given e_1 than given its negation, and it is likewise more probable given e_2 than given its negation. The posterior odds comparing H_1 and H_2 are respectively 3/2 and 4/1. But H_1 is *less* probable given the combination of e_2 with e_1 than it is given e_2 alone. The odds are 3/1, not 4/1.

Is such a combination purely hypothetical? No. Let H_1 refer to Jill's hypothesized preference for Mortimer over Jack; let e_1 refer to the evidence that she has accepted a date with Mortimer to a particular dance at the time Jack calls her to invite her to that dance; and let e_2 refer to the evidence that she turns down Jack's invitation. The structure of Table 1 indicates that it is perfectly rational to "discount" the impact of the rejection of Jack given knowledge of the otherwise damning (to Jack) information that she has accepted an invitation from Mortimer. Other examples of such reversals can be found in Tribe (1971) and Falk and Bar-Hillel (1983) (e.g., a suspect's being seen in a bar slightly drunk 15 minutes prior to a crime committed 12 minutes away and quite drunk 15 minutes afterwards constitutes an "alibi," whereas either sighting alone can be interpreted as evidence of guilt). For a discussion of the role of such Simpson's paradox reversals in thought and science, see Messick and van de Geer (1981).

Of course, it is very difficult to establish a general principle. Hastie and Pennington do not establish that in all cases at all times the order of the evidence makes a difference, but then they do not have to, for a Bayesian analysis indicates that order shouldn't ever make a difference. They have an easier job than Thagard.

Table 1. *Simpson's paradox illustrated*

	e_1	$\sim e_1$	
e_2	.3	.1	H_1
$\sim e_2$	.0	.1	

	e_1	$\sim e_1$	
e_2	.1	.0	H_2
$\sim e_2$	.1	.3	

I am concerned, however, that Thagard does not stick to his basic binary hypothesis. For example, we read at the end of section 2 that P_1 and P_2 may "together explain" Q, whereas "P_1 and P_3 together explain not-Q." I do not understand how we get to "together" from the premise that we only speak "derivatively" of the explanatory coherence of a set "as determined by their pairwise coherence." To be reasonable, that is, to deal with reversals, the system must incorporate such possibilities. The problem, however, is to determine what the system means when it is extended in this way (this problem seems analogous to determining what "a connectionist" interpretation of the Necker Cube means other than that its interpretation in three dimensions is not self-contradictory); see section 3 of Thagard's target article. Unless Thagard's analysis and its realization in ECHO strictly follow the binary hypothesis, they become–to me anyway–indistinguishable from a verbal description of what the people meant to be modeled *could* have thought, had they been reasonable people. (They were.)

Perhaps the problem about the relation between the completeness of Thagard's analysis and its own internal coherence arises because he is trying to do too much: He wants to reduce both scientific and lay thinking to an associative basis, as if there were only one way of thinking logically about a problem. This approach follows that of Cohen (1981), who argues that because people have no way of thinking rationally that transcends their own thinking processes, these processes must be regarded on an *epistemological* basis as defining rationality. The problem with that conclusion is that when people think about a problem that involves logical coherence, they can think about it in many ways, and in fact they see the logical characterization problem as existing apart from their own thought processes. As in perception (Neisser 1976), their orientation is to discover what is "out there"; and although admitting the analytic rather than empirical nature of reasoning to determine it, they nevertheless

use reasoning as part of the process of "discovery." They become–to use the phrase of Davis and Hersh (1981) in describing how mathematicians actually think about their problems–"closet Platonists," even though they cannot justify Platonism. Consider, for example, the Paul Halmos tournament problem. The person in charge of reserving a squash court for a 53-person tournament may compulsively figure out multiple systems of byes to determine how often it must be reserved, only to "discover" suddenly that the answer is 52–because 52 entrants must be eliminated and one is eliminated as the result of each match.

Thus, arguments are rejected or accepted in a much less coherent manner than Thagard's analysis suggests; some occasionally even lead to bad conclusions. Perhaps the attempt to provide a single principle that will lead both to a conclusion based on certain evidence *and* to its subsequent rejection is just too ambitious. After all, some higher courts overturn the decisions of lower ones because their reasoning was improper (e.g., the famous Collins case 1968[1]), and some scientific arguments overturn others. Such reversals of conclusions are not just a matter of discovering a new evidence node, with everyone's associations between the existing ones changed only through the relationship of the new node to them. If it were, new experiments or discoveries would be "crucial" in the sense that Thagard implies they aren't.

I'm not saying that Thagard's goal is impossible to obtain, I just have my doubts. If Thagard can deal with Simpson's paradox, that doubt will be lessened but not eliminated.

NOTE

1. The lower court allowed 1 minus the "exclusion probability" (that a randomly constructed couple would have the characteristics of the accused couple) to be interpreted as the probability that the accused couple was guilty. The higher court ruled that the appropriate probability was that the accused couple was the guilty one given that the couple committing the crime had the specified characteristics. (68 Cal. 2nd *438 P* 2nd 33 66 *California Reporter,* 1968.)

Is Thagard's theory of explanatory coherence the new logical positivism?

Eric Dietrich

Department of Philosophy, Program in Philosophy and Computer & Systems Sciences, State University of New York, Binghamton, NY 13901
Electronic mail: *dietrich@bingvaxu.cc.binghamton.edu*

I view Thagard's theory of explanatory coherence as philosophy of science, and, if I ignore the program ECHO, I find his ideas refreshing and important, in part because his theory is outside the formalist legacy left to us by logical positivism. Thagard, it seems to me, has helped make respectable the idea that scientific explanation is a multifaceted enterprise and that the tools formalists dearly love–logic and probability–constitute merely one of the facets, and a small one at that. Of course, Thagard is not the only one trying to make this idea respectable; we may at long last be ridding ourselves of the shackles of logic and logical positivism. I will return to this in the conclusion.

I do have some reservations about Thagard's theory, however. Simply put, I'm not sure what his theory of explanatory coherence is a theory of. Viewing his theory as philosophy of science required a conscious choice on my part, because there is at least one other way of viewing his theory: as psychology. Both views seem to me to have problems. I will begin with the philosophy-of-science view.

1. Explanatory coherence as philosophy of science. Thagard begins and ends his target article with discussions of explanation and methods for distinguishing good hypotheses from bad ones. Here he is clearly attempting to locate his theory in the space of competing theories in the philosophy of science. But, if his theory of explanatory coherence is philosophy, then what is the program ECHO for? The claim that scientists accept or reject hypotheses based on how their various parts cohere and how the hypotheses cohere with one another can be made without using a computer program. Harman has done it (1973), as have Kuhn (1977) and Kitcher (1981). In fact, scientists themselves find this claim quite plausible, especially those like astronomers and paleontologists who cannot run experimental tests of their theories. Moreover, the program actually interferes with Thagard's argument. For example, ECHO invites such questions as, Why does Thagard select the activation levels and excitatory and inhibitory weights he does? and, Why is the number of cycles ECHO takes to settle important? The answer to the first of these questions is that it makes ECHO settle more quickly (see sect. 4.2 and 12.2), and, as near as I can tell, the answer to the second is that if ECHO settles quickly, we spend less money on compute-time. From a philosophy-of-science perspective, these answers are irrelevant.

I have another problem with ECHO, namely, that it must have *facts, hypotheses,* and *evidence* distinguished for it ahead of time. But what are the criteria for distinguishing among these? I tried an example myself, not using ECHO, but using PLECHUPP (Playing with ECHO Using Pencil and Paper) on Poincare's explanation (1952, pp. 46–63) of the Eureka Phenomenon (having a sudden insight into a problem) and on the competing explanation that insight comes from following rules. I found that distinguishing among hypotheses, evidence, and facts in this case was rather arbitrary. And because I wanted Poincare's model to win, I wasn't sure that the way I distinguished between hypotheses and evidence didn't beg the question against the competing, rule-based explanation. Thagard is aware of this problem, I think (see Thagard's discussion of Lavoisier's and Darwin's arguments, and section 7), but being aware of a problem is not solving it. Moreover, it is more in the spirit of Thagard's theory, viewed as philosophy of science, that propositions should change their status as hypotheses or evidence based on pressures for the propositions to cohere in certain ways.

One last problem with Thagard's theory construed as philosophy of science is his taking the notion of explanation as a primitive. This move makes his whole project seem question-begging. Philosophers of science want to know what explanation *is*. Because Thagard's goal (in my view) is to develop a theory of scientific explanatory coherence, it is perhaps all right to assume some notion of explanation as a primitive for the short term, but then he is not free to criticize other philosophers of science who are attempting to explain explanation on the grounds that their accounts do not do what his does. Thus, Thagard's criticism of Salmon (1966) and Glymour (1980) seems irrelevant and unfair because they are trying to do what he is not: explain scientific explanation.

2. Explanatory coherence as psychology. The existence of ECHO makes more sense (but not much more) if Thagard's theory is viewed as a psychological one about how humans come to believe a certain hypothesis. Thagard is quite right when he says that his seven principles "are too general to have direct experimental consequences." If Thagard's theory is psychological, then to test it he will need detailed predictions regarding what ECHO networks look like when they settle, how long (relatively) it takes them to settle, and what the weights and activation levels are. And it would be interesting if ECHO's shifts in beliefs correspond to those of human subjects. (Of course, it is not the program ECHO that is relevant to this project, but the equations implemented in the program.) It seems extraordinarily unlikely, however, that the few equations Thagard has for testing his psychological theory actually capture the dynamics of human belief change and fixation. I for one believe that one day we will have such equations; perhaps Thagard's theory will make that day arrive sooner rather than later.

3. Conclusion. According to Steve Downes, a colleague of mine (personal communication), Thagard's confusion over

whether he is doing philosophy of science or cognitive psychology has a dark interpretation. He might think that understanding how science works is equivalent to understanding how human individuals work. One hopes that Thagard does not think this, because such a project leaves out the social aspects of scientific explanation and is therefore doomed.

I have a dark interpretation of my own for Thagard's confusion. Clearly, the star of Thagard's target article is the program ECHO. Thagard is also clearly proposing a theory in the philosophy of science. If we recall that one of the hallmarks of logical positivism was its reliance on technical, formal devices derived from logic for solving problems in the philosophy of science, we can perhaps see a *new* positivism, a "computational positivism," moving in to take up where logical positivism left off. Computer programs and a reliance on logic have already virtually ruined artificial intelligence and cognitive science (Dietrich, in press a; in press b). Philosophy may be the next to go–again.

On the testability of ECHO

D. C. Earle

Department of Psychology, Washington Singer Laboratories, University of Exeter, Exeter EX4 4QG, England
Electronic mail: *earle.dc@exeter.ac.uk*

Thagard's theory of explanatory coherence and its connectionist implementation in ECHO is a significant achievement with some interesting possibilities for future development. As the implementation of a theory of hypothesis evaluation in a philosophy of science, ECHO has the particularly pleasing property of being able to disregard contradictory evidence under certain circumstances. This capability is necessary for any sophisticated philosophy of science if it is to accord with the history of science, but frequently the provision of such a capability has a disturbingly *ad hoc* nature: In ECHO the ability to disregard evidence is an intrinsic property of the program.

Thagard rejects a rigid dichotomy between normative and descriptive matters and proposes that the theory of explanatory coherence can be applied to hypothesis evaluation in the philosophy of science, in legal reasoning, and in psychology. As such, ECHO is presented as a model of the behaviour of scientists, the behaviour of jurors, and the behaviour of subjects in experiments in psychology. It is intended that ECHO should be testable, as must be required of any scientific theory. In this respect, a major concern is whether the initial conditions for the application of ECHO to a particular situation are sufficiently constrained to provide the testability and to enable the unequivocal interpretation of results required of a scientific theory.

Consider the application of ECHO to a case of hypothesis evaluation in science. Suppose that ECHO prefers one hypothesis over another; and suppose that the scientific community is divided, with different groups of scientists supporting one or the other hypothesis. Are we to conclude that one group of scientists is behaving rationally and the other irrationally–for example, using arguments that are extraneous to explanatory coherence, such as a prior and otherwise unsupported belief that incoheres with the hypothesis preferred by ECHO? Such a possibility is suggested in a case of legal reasoning in the discussion of the Peyer case, and a similar argument might be presented to account for the creationists' refusal to accept Darwinian theory. However, an alternative interpretation is that in this hypothetical example, the specific application of ECHO models the reasoning of one group of scientists but not that of the other–that the failure lies not with the lack of rationality of one group of scientists but with ECHO. It may be argued that hypotheses are evaluated in relation to the wider set of beliefs of the individual, and that some of these beliefs may be relatively immune to disconfirmations or may be supported by different sets of evidence. If the initial conditions for the application of ECHO were altered to include these other beliefs, then it is possible that a different end state would be reached.

There are a number of related difficulties here. One concerns the question of what is to be counted as rational and what is to be counted as irrational; another concerns the decision as to what is to be included in the initial conditions and what omitted; and a third concerns the weights and activation levels to be given initially to particular items. It is shown that different end states may be reached, depending on the initial settings of the parameters and on the priority given to a certain piece of data. If one group of scientists values a piece of evidence more than another group, then this may well account for the difference in the decisions of the groups. Part of the problem is the large number of free parameters in ECHO, which necessarily make it flexible in its predictions.

Thagard suggests that the input to ECHO could be automated, but in view of the argument presented here, this suggestion is unconvincing, except perhaps for certain well-defined cases. If, however, it is accepted that a failure by ECHO to match the behaviour of an individual, or a group of individuals, in a certain application cannot be interpreted reliably as a failure of rationality on the part of the individuals rather than a failure by ECHO to model the reasoning of the individuals, then on the basis of a similar argument, the success of ECHO to predict hypothesis evaluation is just as equivocal. Human behaviour and ECHO's predictions may be consistent, but for different reasons. Subjects in reasoning tasks are notoriously bad at describing their reasoning processes (Nisbett & Wilson 1977); the initial conditions for an application of ECHO are not easily established from subjects' protocols or from other descriptions of the reasoning process. Without an independently verifiable way of establishing the initial conditions, including the values for the large number of free parameters, the testability of ECHO as a model of hypothesis evaluation must remain limited. In this matter, ECHO is no different from other work in artificial intelligence where algorithms may successfully match human performance but where it is difficult to establish that the underlying processes are the same.

A major challenge for the future development of ECHO will be to find ways of independently establishing the intial conditions for particular applications of ECHO. Enough has been accomplished already to make this endeavour worthwhile, and section 9 suggests that progress is already being made.

What's in a link?

Jerome A. Feldman

International Computer Science Institute, 1947 Center St., Berkeley, CA 94704-1105
Electronic mail: *jfeldman@icsi.berkeley.edu*

One of the hopes for connectionist modeling techniques has been that they will provide a useful scientific language for efforts in various behavioral and brain sciences. Thagard's target article is a beautiful example of how this is beginning to work out in practice. The details of Thagard's theory doubtless need refinement, but the case for expressing competing approaches as networks of positive and negative influences seems convincing. The formulation is no mere recasting of logical or probabilistic arguments–here the *interactions* of the elements determine the outcome. This is potentially of great importance, and my one disappointment with the article is that more attention was not paid to foundational questions.

One of the principal attractions of both logic and probability theory is that each has a relatively clean and well-understood formal semantics. Even if networks of weights and activity levels are better for describing many phenomena, they will not be fully

acceptable without some interpretation of the formalism. This is much more pressing in a philosophically based application such as this one than in a neurophysiological model, for example. The fact that the same basic rules for network generation apply across examples encourages one to believe that there might be principled relations in explanatory coherence. The seven principles and their mapping onto networks provide an informal semantics, but no foundation. No one should expect a complete solution in such a preliminary exploration, but it was surprising to find no acknowledgment that there was an issue. If the links don't represent probabilities (and they don't), what do they represent? There is a practical side to this problem. If we wanted to apply ECHO to an unsolved decision problem, how would we know what choices of weights were admissible?

Coherence: Beyond constraint satisfaction

Gareth Gabrys and Alan Lesgold

Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA 15260

We offer two observations about Thagard's very important work in this commentary. First, we note that the theory represents a specialization of constraint satisfaction systems, making the specifics of the specialization of particular importance. Second, we muse about what it means to do dialectical thinking, to use one's built-in coherence processor as a tool.

Constraint satisfaction. Thagard's principles provide a mapping between a set of explanations and a connectionist computational structure within which coherences can be calculated. In essence, Thagard asserts that explanatory coherence is a constraint satisfaction problem. For sets of explanations, observations, and assertions that can be structured according to his seven principles, a connectionist constraint satisfaction algorithm can find a good fit to the constraints formed by these explanations, observations, and assertions and can assess the contribution each one makes to that fit. Pointing out that this can be done is important. However, once the basic approach has been proposed, further discussion must focus on the question of which details of the approach matter. Principles 2(a), 2(b), and 5 equate explanations with constraints. Principles 1, 6(a), and 7 are general properties of parallel distributed constraint satisfaction models, described in explanatory coherence terms. Principles 2(c), 3, 4, and 6(b) offer unique flavoring to the general proposal. Specifically, 2(c) reduces the weight of explanations requiring cohypotheses, 3 describes how analogy can set up explanatory links, 4 sets up a bias for data propositions, and 6(b) increases the decay of hypotheses when there is unexplained evidence. These four principles respectively account for the influence of simplicity, analogy, evidence, and comprehensiveness on explanatory coherence. We expect that future debate will focus on them.

Given a formalized set of explanations, Thagard's principles can organize them into a constraint structure. But how are hypotheses and explanations formalized? Thagard recognizes this problem and suggests several principles that reduce the arbitrariness of the formalization process. However, the extent of reduction necessarily results in a loss of information. Presumably, humans have additional reasoning mechanisms layered on top of the basic capability Thagard presents. Thagard recognizes that not all arguments can be represented in ECHO, and points out that he is dealing only with causal explanations. It will be intriguing to see what other aspects of reasoning and argumentation can be built on top of ECHO.

Toward dialectical process models. In his examples of jury reasoning, Thagard demonstrates how his approach can provide a useful framework for understanding and measuring explanatory coherence. However, jury reasoning is a fact-finding process that is very different from the jurisprudential reasoning used in deciding cases at law. Indeed, juries are instructed to consider the facts only, and to take instruction in the law from the judge. When cases are argued at law, different principles apply, such as *stare decisis*, the principle that prior decisions should not, in general, be overturned. Ashley (1988) has suggested that case-based reasoning is a more dialectical process than principle-driven reasoning because it involves a search for the particular features that differentiate precedent cases that were decided for the plaintiff from those decided for the defendant. He suggests that one needs to pay attention to the particular factors that distinguish cases decided each way, introducing cut scores on these factors.

For example, consider the importance in a trade secrets case of the number of people to whom the secret has been divulged. Ashley suggests that instead of weighting this value in a constraint satisfaction system, legal reasoning is better modeled by setting cut points on the dimension. For example, a case is weakened if more than a few people have been told the secret. However, on rare occasions a successful case has been made even when thousands were told the secret. The chore then becomes local rather than global–to analyze the links to particular features in order to determine what distinguishes this otherwise aberrant case. One could build constraint satisfaction systems to do that, too, but they would be driven by precedent rather than by causal reasoning. By forcing (clamping) large weights on links from nodes representing cases to nodes representing the two possibilities of the plaintiff's or the defendant's having prevailed in the decisions for those cases, a model would focus on a subset of the weightings known to be generally relevant. The task would be one of discovering, rather than building from, an explanation.

One way this might be done is by alternately operating in global *stare decisis* mode as just described, or in a more local mode in which subsets of the known relationships concerning a set of cases are separately examined. More generally, we hope that it will be possible to build on the level of reasoning described by Thagard's theory to model more reflective thinking. Perhaps such reflection involves the temporary construction of candidate systems of assertions, observations, and explanations that are then subject to the "built-in" coherence analyzing mechanism. Such candidate subsystems might be handled by the kind of attentional gating mechanisms recently introduced by Schneider and Detweiler (1987).

We are intrigued by the possibility of systems that combine both forms of reasoning explicitly, that is, by dialectical systems that use several different weighting schemes and then analyze how they differ. It seems worthwhile to try to model a higher plane of dialectical reasoning that is served by, but goes beyond, constraint satisfaction. Thagard has shown us how humans may quickly evaluate explanations–we still need more work on how explanations are generated in the first place.

ACKNOWLEDGMENT

We thank Arlene Weiner for her insightful comments on our remarks and for helping us sharpen them.

What does explanatory coherence explain?

Ronald N. Giere

Center for Philosophy of Science, University of Minnesota, Minneapolis, MN 55455

Thagard begins his target article by asking, "Why did the oxygen theory of combustion supersede the phlogiston theory?" Answering such questions has been a major goal of the philosophy of science for as long as it has existed as a recognizable discipline. Rejecting answers deriving from the philosophy of

logical empiricism, Thagard would replace or supplement the resources of logic with those of the cognitive sciences, particularly artificial intelligence. This approach is gaining adherents within the philosophy of science community (Darden 1983; Giere 1988; Glymour et al. 1987; Nersessian 1984). In spite of my sympathies with the general approach, it seems to me that Thagard is still a long way from answering the questions he poses.

Thagard claims that as a matter of psychological fact, individual scientists reason and evaluate theories according to something like his model of explanatory coherence. ECHO, he says, "can handle very complex examples of actual scientific reasoning." The revolutions in question took place, therefore, because the scientists involved individually reasoned to similar conclusions. This claim is not adequately supported by the evidence Thagard presents.

The explanatory relationships modeled by ECHO are obtained by an intuitive analysis of texts written by major architects of scientific revolutions, such as Lavoisier and Darwin. In both of these cases, the texts in question were written long after the principals had themselves become convinced of the correctness of their views. Moreover, in each case the scientist's purpose in producing his text was not to record the thought processes by which he became convinced of the correctness of his theory, but to establish his claim on the theory and to persuade others of its value. These purposes are so different that there is considerable reason to doubt that a text produced in the latter context would provide much insight into the former processes.

In his target article, Thagard refers to "the input given to ECHO to represent Lavoisier's argument in his 1783 polemic against phlogiston." Later he repeatedly refers to "Darwin's argument." These phrases suggest that what Thagard is really modeling is not scientists' reasoning but the structure of their arguments, presented in what might be called "the context of persuasion." The most the model of explanatory coherence explains, then, is why a scientist's presentation favors his view over those of his rivals. It presents his hypotheses as more explanatorily coherent with the data and each other than those of the opponents.

Thagard recognizes the objection that he might only be "modeling the rhetoric of the scientists, not their cognitive processes." However, his reply that "there is some correlation between what we write and what we think" fails to meet the objection. Of course there is "some correlation." The question is whether there is enough. Thagard provides little independent reason to suppose that there is sufficient correlation to use *what* one *writes* as an indicator not only of what, but *how*, one *thinks*. Why should written arguments constructed after the fact to persuade others be a good indicator of one's original cognitive processes?

Now suppose we grant, for example, that Lavoisier's presentation exhibits his theory as possessing greater explanatory coherence than phlogiston theories. We still have no explanation of why there was a revolution, which is to say, why *others* adopted Lavoisier's presentation as their own. Because Thagard treats "explains" as a primitive, he must agree with competitors' claims about what explains what. Moreover, as Thagard allows, different scientists may assign greater weight to some explanatory relationships than to others. Thus, applying Thagard's own model to the writings of Lavoisier's opponents would probably result in their presentations of the phlogiston theory exhibiting greater explanatory coherence than Lavoisier's theory. Even if some of these opponents later adopted Lavoisier's arguments, we are left without any explanation of *why* they changed their minds.

The traditional philosophical objection to coherence theories is that there is not enough of a connection between internal coherence and representational fidelity to the external world. It requires only theoretical ingenuity to construct a highly coherent explanatory network. Why should that provide much basis for thinking that the network so constructed *represents* an external world beyond the given facts?

In response, Thagard might adopt a more normative stance. His model of explanatory coherence, he might claim, captures the normatively correct relationships among statements that determine rational acceptability. Moreover, there is a single, "correct" presentation of the data and rival hypotheses that reveals Lavoisier's theory to have greater explanatory coherence. The revolution took place because the scientists involved were rational agents who followed the norms of explanatory coherence. That this might ultimately be Thagard's position is suggested by his remarks in section 10.5 on "the descriptive and the normative," and by the more extended discussion in his book (1988, Chapter 7). In neither place does he provide any reason to believe that such norms are actually operative.

On the latter interpretation, Thagard's model functions like an inductive logic, though it is richer than the probabilistic logics developed by the logical empiricists. The strong similarity between these two approaches is largely due to the fact that both attempt to analyze scientific reasoning in terms of more or less formal relationships among statements, particularly statements representing "hypotheses" and "evidence." Although this assumption is common in the artificial intelligence community, it is less widely accepted in other areas of the cognitive sciences such as cognitive psychology (Tweney 1985) or the neurosciences (Churchland 1986).

My own view (Giere 1988) is that a genuinely "cognitive" approach to explaining science must get beneath the linguistic surface to the *nonlinguistic* representational mechanisms and judgmental strategies operative in individual cognitive agents. These mechanisms and strategies have representational significance because they incorporate active causal interaction with the world, especially through experimentation. Scientific revolutions emerge as the collective result of individual judgments by members of the relevant scientific community. There is no need for any normative principles of rationality. Indeed, one can allow a considerable role for "noncognitive" factors as well. The result is a more faithful account of science as it really is.

Are explanatory coherence and a connectionist model necessary?

Jerry R. Hobbs
SRI International, Menlo Park, CA
Electronic mail: *hobbs@ai.sri.com*

The general pattern for explanation is

H explains E,

where H and E are sets, or conjunctions, of propositions. The widely acknowledged criteria for determining the "goodness" of an explanation are that H should be as small as possible and E should be as big as possible. We want more bang for the buck, where E is the bang and H is the buck. These are the criteria of simplicity and consilience that Thagard has written about perceptively in previous papers and in parts of this one; his discussions of these criteria have been important in general and a significant influence on my own thinking.

The most simple-minded procedure for measuring the goodness of an explanation would be to count the propositions in E, count the propositions in H, and subtract. Pick the theory that has the highest such number and contains no contradictions. (Let's call this the Naive Method, and refer to the number as $\#E - \#H$.) There are at least two problems with this procedure–what is meant by "explains," and what are the individuating criteria for the propositions in H and E. They are problems for Thagard's method as well. He legitimately skirts the first of the

problems, and in section 7 he probably says all that can reasonably be said about the second in a short article.

In any case, the Naive Method needs to be replaced by something more sophisticated. Thagard proposes a more complex method for evaluating theories by defining a relation of explanatory coherence between propositions and then using those relations as the links in a connectionist model ECHO that computes the explanatory coherence of an entire explanation. It turns out, however, that for every single one of Thagard's examples, the Naive Method yields exactly the same result that ECHO yields. This gives rise to the disquieting suspicion that all of this connectionist architecture and the theory of explanatory coherence it rests upon amount to nothing more than a very complex and possibly inaccurate procedure for doing subtraction.

There are two issues that need to be examined more closely from the perspective of the Naive Method. The first is whether we should always prefer deeper theories even where we do not thereby expand the evidence explained. This is illustrated by the example in section 4.3. Here, for Thagard, {H3} is the best theory because in addition to explaining E1 and E2, it explains H1. Thagard's intuition is that this gives it a greater explanatory coherence. I'm not sure about that. If H1 is of no independent interest, should the fact that it is also explained make {H3} a better theory? In the Naive Method, if we follow Thagard, {H3} explains {H1,E1,E2} and $\#E - \#H = 3 - 1 = 2$. If we don't, {H3} explains only {E1,E2} and $\#E - \#H = 2 - 1 = 1$, the same score earned by the theories {H1} and {H2}. It is in the former case, when we follow Thagard's intuition, that the Naive Method matches ECHO's results. (In neither case do we include explained hypotheses in H.)

This consideration turns out to be significant in the Peyer case described in section 6.2. If we follow Thagard in valuing explained hypotheses and thus including them in E, then in the case for Peyer's guilt, $\#E - \#H = 9$, whereas in the case for his innocence, $\#E - \#H = 7$. This corresponds to the judgment of guilt that ECHO reached. On the other hand, if we don't include explained hypotheses in E, then $\#E - \#H = 6$ for both guilt and innocence. This, recall, is the case that resulted in a hung jury.

The second issue is the use of analogy in evaluating theories. Thagard views this as a separate and legitimate criterion, but it seems to me that it can be subsumed under simplicity and consilience. The existence of an analogy does not by itself enhance a theory's explanatory power. It does so only when that analogy can be seen to rest upon a deeper, underlying, perhaps very abstract principle from which the two analogous explanatory relations can be derived. It is not enough for "H1 explains E1" and "H2 explains E2" to be analogous. They must be so because they both instantiate some common abstract principle P. A particularly clear instance of this is Thagard's one concrete example of an analogy, in the Darwin case discussed in section 5.2. The two specific explanatory relations–that "human-directed selection can result in new breeds" and that "natural selection can result in new species"–are both instantiations of the more general principle that "selection can result in new varieties of living beings." This is true even if the general principle was not recognized until the analogy was constructed.

When a single mathematical theory is applied to seemingly different phenomena, the mathematical theory is the deeper, underlying, abstract principle that the analogy between the phenomena rests on.

It is hard to find examples in science of analogous explanations that do not rest on a common abstract principle–a fact that by itself supports my claim. But an analogy that, for me at least, does not have such an underpinning is the analogical picture of Thompson's model of the structure of the atom: Electrons are embedded in the nucleus of an atom as raisins are embedded in a pudding. This analogy may help us visualize Thompson's model, and the model itself provided an explanation of ionization, but the analogy does not, in my mind, draw on any underlying structural principle and is not in the least compelling as an argument for the model.

If an analogy between two explanatory relations, "H1 explains E1" and "H2 explains E2," to be convincing, must rest on a deeper underlying principle P, then explanation by analogy can be subsumed under the criteria of simplicity and consilience. The explanation is simpler and more consilient because a single general principle explains two, more specific principles, while the original hypotheses continue to support the original evidence. This observation translates directly into the operation of the Naive Method. Rather than explaining {E1} with {H1}, we explain {E1,E2,(EXPLAIN H1 E1), (EXPLAIN H2 E2)} with {H1, H2, P}, so that $\#E - \#H$ is increased by one. When analogy is treated in this way, in the examples of sections 4.7 and 5.2, the Naive Method yields the same results as ECHO.

I am sympathetic with the notion that the best theory is the most coherent one; and I am at least agnostic regarding connectionist models. But Thagard's target article does not, unfortunately, constitute an argument for either, because from the examples presented, we cannot be convinced that ECHO is more than an excessively complicated way of implementing an evaluation metric that involves neither and is surely far too simple.

Inference to the best explanation is basic

John R. Josephson

Laboratory for Artificial Intelligence Research and Department of Computer and Information Science, Ohio State University, Columbus, OH 43210
Electronic mail: *jj@cis.ohio-state.edu*

I am in full agreement with Thagard and others that there exists a powerful and ubiquitous form of inference that is built on explanatory relationships. Yet I believe that the explanatory-coherence account proposed by Harman, and given computational flesh by Thagard, is seriously but subtly flawed.

Harman (1965) argued that "inference to the best explanation" (IBE) is *the* basic form of nondeductive inference, subsuming "enumerative induction" and all other forms of nondeductive inference. He argued quite convincingly that IBE is a common and important pattern of inference and that it subsumes sample-to-population inferences, that is, inductive generalizations, as a special case. (This is my way of putting the matter.) The weakness of his overall argument was that other forms of nondeductive inference are not seemingly subsumed by IBE, most notably population-to-sample inferences, that is, predictions. The main problem is that the conclusion of a prediction does not seem to explain anything. (See Josephson, 1982, pp. 107–30 for more details.)

This last point, and others, were taken up by Ennis (1968). In Harman's reply to Ennis, instead of treating predictions as deductive or admitting them as a distinctive form of inference not reducible to IBE, Harman took the curious path of trying to absorb predictions, along with a quite reasonable idea of IBE, into the larger, vaguer, and less reasonable notion of "maximizing explanatory coherence" (Harman 1968). In this I think Harman made a big mistake, and Thagard has followed him in making it.

I think that there is a clear and basic form of inference that goes more or less as follows:

D is a collection of data (facts, observations, givens),
H explains D (would, if true, explain D),
No other hypothesis is able to explain D as well as H does.
Therefore, H is probably true.

The confidence in the conclusion should (and typically does) depend on the following considerations:

(1) how decisively H surpasses the alternatives;
(2) how good H is by itself, independent of the alternatives

(e.g., we will be cautious about accepting a hypothesis, even if it is clearly the best one we have, if it is not sufficiently plausible in itself);

(3) how thorough the search was for alternative explanations; and

(4) what are the pragmatic consequences, including the costs of being wrong and the benefits of being right;

(5) how strong the need is to come to a conclusion at all, especially considering the possibility of seeking further evidence before deciding.

This inferential pattern is the basic one, I contend, with an epistemic force and information-processing significance all its own (Josephson et al. 1987). The main goal of such an inference, to arrive at a confident explanation of something, is a reasonable one to pursue if we aim at understanding. But the reasons for wanting to maximize overall explanatory coherence are obscure. Moreover, IBE, as I have just described it, relies intimately on processing considerations not reflected in Thagard's model, such as the formation of hypotheses and the search for alternative explanations.

One sign of the weakness of Thagard's model is the symmetrical links, which make the model unable to accommodate logical implication or the asymmetry of cause and effect. A further sign of trouble is the "symmetry of explanation and prediction" built into the model by way of the symmetry in Principle 2(a) (note too the discussion in section 4.3 and the remark in section 4.4). At first appearance it may seem that a theory capable of explaining something is capable of predicting it, and conversely; yet this convenient relationship can be seen to break down rather quickly in realistic cases (see Scriven, 1962, for a classical criticism). Often we are in a position to predict a fact without thus being in a position to explain it (as for example when we trust someone else's prediction). Furthermore, we are often in position to explain a fact without thus being in position to predict it (namely, when our explanation is not complete, which is typical, or when the explanation does not posit a deterministic mechanism, which is also typical). Contrary to the thrust of Thagard's model, it is failed *predictions* that cause us to reject a theory, not facts which the theory could have explained if they had obtained, but which did not happen to obtain.

Because the nodes in Thagard's model are propositions, with the increase and decrease of activation levels corresponding to increased and decreased fitness for acceptance, we may expect that the spreading of activations from one node to the next reflects the communication of evidential support. Whereas one direction of the symmetrical Principle 2(a) corresponds reasonably well to IBE, and is thus a legitimate path of evidence, the other direction seems to reflect a big confusion between explanation, prediction, and a consideration of whether a proposed explanation is itself plausibly explainable. Thus I disagree with Thagard (and Harman) that, in general, an explanation conveys evidence for what is explained, and so I reject the symmetry of the central Principle 2(a).

ACKNOWLEDGMENTS

This work has been supported by the NIH, National Heart, Lung, and Blood Institute under grant no. 1 R01 HL38776–01, and by the Defense Advanced Research Projects Agency under RADC contract F30603–85–C-0010. This commentary has benefited from discussions with Susan Josephson and from the scholarly assistance of Lindley Darden and Andrew Dahl.

Does ECHO explain explanation? A psychological perspective

Joshua Klayman and Robin M. Hogarth

Center for Decision Research, Graduate School of Business, University of Chicago, Chicago, IL 60637

Like its author, ECHO has connections with philosophy, artificial intelligence, and psychology. The focus of our commentary is psychological. What is the status of ECHO as a descriptive model of explanation?

At the heart of ECHO lie seven basic principles specified in section 2.2. Indeed, it is hard to imagine any system that adhered to these principles and yet acted differently from ECHO in any significant way. Thagard skirts the issue of whether these are really meant as descriptive psychological principles, but almost all of them could be taken that way and would be interesting as such.

The ECHO analyses described by Thagard might be viewed as tests of these underlying principles. However, from a methodological viewpoint, they do not constitute good tests. Some of the cases are just too easy. For example, almost any system that tabulated arguments pro and con (e.g., Axelrod's "cognitive maps" [1976] or Franklin's "moral algebra" [Dawes 1988]) would conclude from Darwin's arguments that evolution was better than creationism. Other tests, it could be argued, are tougher, namely, the Peyer trial, which ended in a hung jury. But in this case, it is not clear how ECHO's conclusion should be evaluated or what an appropriate outcome would be (a hung model perhaps?). The descriptive adequacy of Thagard's principles could be tested, but this task would be better accomplished through direct psychological experimentation.

From a substantive viewpoint, ECHO does not model the process of thinking, but rather its end result. All of the examples presented come from prepared arguments, or from secondary accounts. Although ECHO tells us what one might conclude from reading Lavoisier's arguments, it is important to recall that Lavoisier had already established the intellectual agenda. Moreover, once the network has been specified, most of the interesting psychological judgments have been made either by the person being modeled or by the knowledge engineer. What evidence is relevant? Which hypotheses are supported or contradicted by it? Which hypotheses are mutually incompatible? What level of explanation is appropriate? Are two hypotheses really the same or different? Are those pieces of evidence redundant? That ECHO enters late in the process is demonstrated by how soon one can predict which hypotheses will be accepted. The oxygen/phlogiston fight is over by about round 10 (Figure 11), and even in the difficult Peyer case, the winners and losers are established by about the 15th cycle (Figure 20).

So is ECHO psychologically vacuous then? Not necessarily. ECHO might best serve as a model not of *how* people think, but of *what* they think. Here Thagard may have missed a useful analogy to the work of Pennington and Hastie (1986; 1988), which he cites. Their claim is not that the jurors' story structures lead them to think the way they do, but rather that the stories *represent* the way they think about the evidence and that their decisions more or less follow from the stories constructed.

As a representational model, ECHO could help us understand psychological processes. One intriguing possibility is to use ECHO to study the dynamics of belief formation and revision, looking at changes in the belief network as evidence is added or taken away, as new hypotheses are introduced, or as new links between hypotheses are suggested. The work by Ranney and Thagard (1988) is a promising effort in this vein. ECHO could also provide a framework for modeling and tracing a number of interesting psychological phenomena. For example:

(1) The order of information presentation can have a major effect on final beliefs (see Hogarth & Einhorn 1989). Pennington and Hastie (1986; 1988), for example, found that early informa-

tion has a strong impact on the way subsequent data are interpreted, and thus on the final representation. An ECHO-like analysis could help establish the locus and function of such effects.

(2) An important feature of ECHO is that hypotheses activate and deactivate data as well as vice versa. This clearly happens in the practice of science and may sometimes be normatively appropriate (Koehler 1989). On the other hand, it may induce inappropriate "belief perseverance" (Ross & Lepper 1980). ECHO might provide a framework for distinguishing the legitimate and illegitimate influences of hypotheses on the evaluation of data.

(3) In problem solving, sudden insights sometimes emerge from incremental changes in data and hypotheses ("aha" effects). ECHO could be used to elucidate the conditions that trigger or enable such restructuring of beliefs.

(4) Scientists (and other people) collect information to test hypotheses. How they do so can influence their beliefs (see Klayman & Ha 1987; 1989). ECHO might profitably be extended to model how hypothesis testing strategies affect beliefs and vice versa.

(5) Whereas Thagard treats probability as irrelevant, ECHO might be used to gain insight into the origins of subjective probabilities. In many ways, the activation states of the various hypotheses could be thought of as reflecting subjective "degrees of belief" even though they are not probabilities (cf. Gluck & Bower 1988).

(6) Whereas Thagard discusses sensitivity analyses concerning the parameters of activation, it may be more instructive to apply such tests to the structure of the network. What happens, for example, when redundant evidence or straw-man hypotheses are introduced? What if one changes the level of detail or the number of layers of explanations-of-explanations? The fact that ECHO doesn't specify how these aspects should be determined is a weakness, but one that provides an opportunity to test how such manipulations affect human thinking.

Finally, although ECHO's status as a descriptive or normative model is unclear, it may still have prescriptive value. In particular, ECHO could find a useful niche as a tool for promoting more effective problem solving and scientific exploration. If one could elaborate the belief network for an unsettled area of investigation, it might be possible to identify critical subquestions (e.g., "the whole thing hinges on whether H_8 is right or H_9"), thereby suggesting the more promising issues on which to focus future research. ECHO might also help in resolving conflicts between different investigators or schools of thought, clarifying critical differences or assumptions ("she says H_3 incoheres with H_4, but you don't").

In conclusion, ECHO is not a psychological process model and cannot provide good tests of its seven underlying principles. However, it succeeds in modeling complex situations with underlying processes that are simple, local, plausible, and few. That makes it a promising framework for analyzing and understanding human processes of hypothesis evaluation and belief revision.

Explanatory coherence in neural networks?

Daniel S. Levine

Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019

Electronic mail: *b344dsl@utarlg.bitnet*

Thagard's target article addresses a major issue for those interested in either biological or artificial intelligence. The hypothesis testing that he models falls into the class of reasoning or inferential processes that have thus far not been addressed adequately by connectionist (neural) networks. Some cognitive scientists, such as Fodor and Pylyshyn (1988), have argued that such processes *cannot*, in principle, be addressed within the connectionist framework. But our brains, made up of neurons and synapses, manage to perform such reasoning tasks (albeit fitfully!); hence there must be some way to understand mechanistically how we can do them. Work like Thagard's is a first step toward such a mechanistic understanding.

Understanding of any complex cognitive process is facilitated by breaking it up into simpler and more accessible processes. The major strength of Thagard's work is that it breaks up the testing of a theory into the implementation of several rules that incorporate simpler tasks. Coherence, incoherence, and analogy can then be represented by various excitatory and inhibitory links within a network, and the activation functions of nodes in the network can be computed over time. Moreover, some of the resulting dynamics are reminiscent of phenomena occurring in connectionist models. For example, feedback between "evidence" nodes and "hypothesis" nodes in Thagard's ECHO can lead to selective neglect of data (section 4.8), thus embodying a primitive form of the kind of selective attention arising from feedback between sensory and motivational nodes in neural networks (e.g., Grossberg & Levine 1987).

Yet significant gaps remain in Thagard's model at both the cognitive and the "neural" levels. At the cognitive level, the model simply suggests acceptance or rejection of a given theory. However, as new data are uncovered that conflict with the current form of a popular theory, the attempt is usually made to modify the theory rather than to abandon its entire structure. The ECHO model does not suggest a criterion for when and how to modify a theory within its fundamental structure or to synthesize parts of two conflicting theories.

For example, the Darwinian theory of evolution underwent modification within its originator's lifetime and continues to evolve (no pun intended) to the present day. In my opinion, Thagard's target article is incomplete in its treatment of the conflict that Darwin himself saw between the hypothesis that species have evolved and the relative lack of transitional forms in the fossil record. The ECHO simulations of this theory did not incorporate the incoherence between that hypothesis and the data, but simply combined the hypothesis with Darwin's own *ad hoc* assumption ("the fossil record is incomplete"). I am not well versed in the evolutionary biology literature, but my impression is that biologists are still, within the evolutionary framework, constructing less *ad hoc* explanations for the paucity of transitional forms. Such explanations include hypotheses that mutations are not purely random but are guided in some way by other mutations or by environmental events.

At the "neural" level, Thagard, of course, assumed that the various belief and knowledge representations in the network were "atoms" without giving an explanation for how they might arise from lower-order processes. This is not a criticism of his work but an expression of a challenge to connectionist theorists. Neural network theories of categorization and of segmentation of the perceptual environment are already available (see Edelman 1987; Grossberg 1988; and Levine, in press, for summaries of recent work). Going from categorization and segmentation to constructs, knowledge, and beliefs should take only a few more steps (though, as Neil Armstrong would say, they are likely to be giant steps).

Although the work that still needs to be done is vast, Thagard should be commended for building a bridge between several different cognitive outlooks. The target article should be used as a source for others attempting to build realistic theories of knowledge representation.

Explanationism, ECHO, and the connectionist paradigm

William G. Lycan

Department of Philosophy, University of North Carolina, Chapel Hill, NC 27599-3125

Explanationism, in epistemology or in the philosophy of science, is the view that an inference is warranted/justified/rational/reasonable/legitimate/. . . when it increases the "explanatory coherence" of a subject's total belief set–that is, when the resulting belief set exhibits greater coherence than did the subject's initial, preinferential belief set. Explanationism admits of a weaker and a much stronger version: A "weak" explanationist holds just that coherence increase *can* per se justify an inference; a "strong" explanationist maintains that coherence increase is the *only* thing that can ever justify an inference. As yet there are very few straightforward strong explanationists (Harman 1986; Lycan 1988); even weak explanationism has been hotly contested (Cartwright 1983; Hacking 1982; van Fraassen 1980). Thagard accepts the weak but rejects the strong variety (section 10.4).

A persistent embarrassment to explanationist epistemology is that the notion of "coherence" itself has remained airily vague. The only immediately obvious candidate as a specific element of coherence is self-consistency or logical coherence, the bare absence of self-contradiction. That feature alone gives epistemologists little to go on, and precious little else has been said on the topic of what makes for coherence.[1] And interestingly, even that feature is forgone or at least deemphasized by Thagard, who thinks consistency is highly desirable but only in its proper place. As his own centerpiece of coherence, Thagard suggests that propositions cohere when they explain, when they are explained, or when they join with other propositions in explaining.

Though plausible, those suggestions in themselves are no more specific or testable than any other explanationist slogans to date have been. It is notoriously hard to think of any realistic way to implement such slogans. But Thagard's project is precisely to implement them.

He begins by taking coherence to be a binary relation on individual propositions. In light of the total-belief-set holism espoused by explanationists under the original influence of Sellars (1963) and Quine (1953; 1960), that choice seems pathetically preliminary and unworkable. Worse, Thagard stipulates that binary coherence is symmetric; for example, a proposition coheres exactly as much by being explained as by explaining, other relations being equal. (Though explanation itself is not symmetric, Thagard argues that the more general notion of coherence is.) Yet Thagard relaxes neither assumption. And, surprisingly, his results seem none the worse for his deliberately naive treatment. ECHO's overall analyses so far square well with the (crude) history of science and with present-day intuitive judgment. As Thagard admits, in any given case study, ECHO gets a lot for free: particularly (1) the data, (2) the initial credibility of the data, (3) the "explaining" relation taken as primitive, and (4) the partitioning of data and of *explanantia* into distinct atomic propositions. But even so, ECHO does remarkably well at its nontrivial subsequent job of theory ranking. And if ECHO continues to do well at more subtle and complicated theory-ranking tasks, whether or not it also takes over responsibility for some of the presently gratis (1)–(4), that will be fairly big news for explanationists, indicating that coherence is not only quantifiable but can usefully be taken as binary.

My only further question at this point concerns Thagard's allegiance to connectionism as a format for implementing his explanationist model. He calls ECHO "a straightforward application of connectionist algorithms to the problem of explanatory coherence." True, symmetric pairwise coherence of propositions is easily depicted as a mutually excitatory link between proposition-representing units in a connectionist network (though the symmetry of excitation/inhibition strengths as well is unusual in connectionist modeling). And the coherence of the network as a whole can be represented by the connectionist measure H (aptly respelled by Thagard for the explanationist tradition as "harmany"), which we want to maximize. But these notational facts show neither that connectionism lends support to explanationism nor the reverse. For pairwise coherence can be equally well represented (on whatever machine) by a linear numerical function, and likewise global "network" coherence by an agglomerative arithmetical function on unit "activation" values; connectionist architecture has no particular advantage over von Neumann architecture in the implementing of Thagard's explanationist device. Architecture does not distinguish ECHO from classical probability theory, standard or nonstandard confirmation theory, or any other known calculus of proposition credibility; all of them assign credibility values to propositions as a function of the values of other propositions. Thus, what Thagard calls "ECHO's connectionist character" is not strongly marked.

In saying that, I mean no criticism of Thagard, who is commendably modest in his claims and in particular disavows any attempt at "neural plausibility." My point is a general one: It is increasingly fashionable to formulate one's epistemological/psychological theory or device in "connectionist" terms, ostensibly as opposed to proof-theoretic or other good old-fashioned AI terms. But it often unclear what advantage is being secured. A model such as Thagard's (or Goldman's 1986), whose "units" represent whole propositions, bears no relation at all to neurophysiology, and can derive no glory from some connectionists' early claims to be engaged in neural modeling. Its benefit must be some more general computational advantage of parallel processors over traditional artificial intelligence programs. But seldom are we shown such an advantage. Typically, the models in question could have been implemented just as easily on the same hardware using traditional architecture. "Connectionist models" of this and "connectionist approaches" to that are often not essentially or even notably connectionist at all.

NOTE

1. See, however, the papers featured in a special issue of *Linguistics and Philosophy* (February 1984; 7[1]) on "Coherence," edited by Douglas F. Stalker.

New science for old

Bruce Mangan and Stephen Palmer

Department of Psychology, University of California at Berkeley, Berkeley, CA 94720

Electronic mail: *palmer@cogsci.berkeley.edu*

Thagard's target article embodies a paradox. On the one hand, his theoretical view of the nature of science is progressive: He is at home with Kuhn, Lakatos, Quine, and Duhem, with holistic explanation and Gestalt shifts. His examples of scientific thinking are of the paradigm type, with classic examples drawn from scientific revolutions rather than from the more prosaic realms of "normal science." And of course the model into which Thagard puts his analysis of coherent explanation incorporates one of the newest fields in cognitive theory and computer simulation: connectionism.

On the other hand, the actual structure of Thagard's simulation looks much closer to Kant, with a tincture of Bacon. There is nothing wrong with Bacon or Kant. As philosophers of science they are a bit out of style, but that does not make them less important or less potentially valuable for current thinking; much current thinking is built on them. However, if one were given the exercise of putting some of the more recent ideas about the

nature of science and scientific explanation into connectionist terms, an architecture rather different from Thagard's would probably emerge, one which would take advantage of more of the resources of connectionism. For purposes of comparison, we will later sketch an example of this sort. But for the most part, we will consider some of the less "progressive" components of Thagard's model and see how they contrast with the theoretical ideas Thagard seems to believe he incorporated in ECHO.

Holistic approaches to the philosophy of science go back at least as far as Leibniz (see especially his *New Essays* 1765/1981) and underlie much of Kant's work, the most influential being the *Critique of Pure Reason* (1787/1963) and the *Critique of Judgement* (1790/1951). Perhaps the fundamental difference between Kant's holistic philosophy of science and the holism of the later twentieth century involves the degree to which the underlying principles of cognition are thought to change. For Kant these cognitive principles are a priori and absolutely fixed.

The Principles of Explanatory Coherence in Thagard's model function very much as if they were a priori principles. They are prior to any hypothesis or data and remain invariant from case to case; they serve to connect every hypothesis with a set of particular data. This complex is then further integrated into a single, maximally coherent whole, jointly constrained by a set of particular facts, and a set of unchanging principles of analysis and explanation. Kant's approach has many similarities. The Categories, for example, though analytically distinct, were understood to operate simultaneously in any cognitive or perceptual act. The final aim of cognition was the "synthesis of the manifold." The German-speaking focus on the unity of Gestalten stems from Kant, and much of Kant's work aimed to explain the cognitive process behind scientific thinking, with Newton's method of analysis and synthesis (see Mackinnon 1978) as the great exemplar. But for Kant, as for Thagard, there was no way the data, or any particular cognition or hypothesis, could ever modify the basic principles that structure the system.

The Duhem/Quine holism has a very different flavor. For Quine (1961) in particular, as Thagard points out, there was no absolute distinction between analytic and synthetic propositions, and propositions were organized into a "corporate body." The first position would have horrified Kant; the second, applied to cognitive processes, would have passed as a truism. But Quine also held in *Two Dogmas of Empiricism* (1951/1961) that all such principles could be conditioned and modified by experience. The corporate body was not fixed. This is the significant modern twist to holism, but it is not reflected in Thagard's model. The principles of explanation, as they operate in ECHO, are *outside* the model and thus cannot be changed except from the outside. The principles used by ECHO condition the analysis in advance, but are unaffected by any outcome of that analysis.

Thagard has a similar problem vis-à-vis Kuhn (1970). For Kuhn and related thinkers, the fundamental principles are also malleable. A scientific revolution means a shift in basic principles of explanation. For example, the movement from Aristotelian to Galilean physics was in large part a shift in what would count as an explanation (see Feyerabend 1975). The notion of a "natural place" ceased to make explanatory sense, and other notions such as mathematically specified prediction came to the fore. Thagard's analysis of Darwin provides a very good example of the importance of introducing, or emphasizing, new explanatory principles and not just new empirical findings or hypotheses. As Thagard himself points out, the use of analogy became an important explanatory device for Darwin. Although an argument by analogy is generally weak and usually avoided in modern science, Darwin was able to integrate it into his battery of explanatory principles because no better alternative existed and useful theoretical work could be done if it were accepted. So for Darwin we may say that the explanatory principle of analogy from the observed to the unobserved was in a sense recruited by Darwin's more specific hypothesis. Although hypothesis and evidence can interact in ECHO, the dynamic role of explanatory principles at the heart of Darwin's work in particular and paradigm shifts in general currently stands outside Thagard's model.

It is therefore not correct to think of ECHO as modeling a paradigm shift. A paradigm shift involves a basic change in the mode of analysis, and nothing like this happens in ECHO. Any impression that ECHO does model something especially germane to the process of scientific revolution is mistaken. If Thagard's aim is simply to model the general structure of scientific thinking, then any specimen of scientific thinking should do. Choosing examples solely from revolutionary moments in science is misleading, as it invites the inference that paradigm shift is the process being modelled. Scientific revolutions may involve a Gestalt shift, but not all Gestalt shifts that occur in the process of doing science are harbingers of a scientific revolution. One can suddenly "see the point" while doing quite ordinary research *within* a given paradigm. Indeed, ECHO looks much more like Kuhn's model of "normal science," in that Thagard's explanatory principles do function as a kind of paradigm, but a paradigm that cannot shift. We will return to this point below.

Thagard's model also has an "inductive" quality that, in effect, deemphasizes the role of hypotheses relative to modern thinking in philosophy of science. Even some neopositivists recognize the importance of hypotheses as the organizing entity that activates and focuses scientific work. The standard contrast is with Bacon's (1620/1960) idea that science was to be scrupulously inductive. Darwin again provides a good example, in this case of the fundamental organizing role of his hypothesis. *The Origin of Species*, as he once wrote to Lyell, involved "inventing a theory and seeing how many classes of facts the theory would explain" (Himmelfarb 1962, p. 157).

ECHO's architecture, however, looks inductive in at least two ways. The first is harmless but suggestive: Activation enters from the evidence units and can only then move on to the various hypotheses. Because the activation can circulate back to the evidence units, this may have little real effect on hypothesis choice. So although there is the form of evidentiary priority, it is probably without great substance.

The second way in which an inductive tendency affects ECHO's operation is more significant, because it may have driven a wedge between Thagard's official controlling idea–System Coherence–and ECHO's actual method of selecting a hypothesis. System coherence, also known as goodness, harmony, and so on, is a metric that characterizes the global or holistic *degrees of consistency* within the entire system. As with virtually any connectionist network, ECHO must settle into a state of maximum goodness or coherence to work at all. But except for the fact that activation at any given node will stabilize as the result of this process, hypothesis choice in ECHO cannot be directly equated with system coherence at all. Hypothesis choice in ECHO is determined simply by comparing the discrete activations of a few hypothesis units with one another.

In contrast, consider a more "holistic" and "deductive" way of choosing between hypotheses, but one that is still roughly within the ECHO format: Activation enters the system, not through evidence units, but via a given *hypothesis* unit that is "clamped" on. This hypothesis unit then evokes its own best-fitting configuration of activation in the network. The hypothesis is then deactivated, the next hypothesis unit is clamped on, and the process is repeated for each remaining hypothesis. The winning hypothesis is the one that creates the most coherent network. Note that in this case we are choosing the best hypothesis by observing its direct effect on the network as a whole and not by using any indirect measure such as the relative activation of the hypothesis units compared in isolation. Further changes in ECHO's architecture would probably be necessary to implement this idea, but the general point should be clear: The present proposal attempts simultaneously (1) to bring ECHO closer to the modern view of hypotheses as central organizing

devices and (2) to use system coherence directly to evaluate explanations in Thagard's sense.

A further step in ECHO's development requires a much bigger conceptual change. If a connectionist network can be made to represent explanatory principles of ECHO's general type, it might be possible to move ECHO squarely into the later twentieth century and provide it with mechanisms that will simulate paradigm changes. The essential innovation is somehow to incorporate the paradigm *within* the network rather than having it stand outside. What needs to be accomplished is to represent explanatory principles themselves as units in the net in such a way that (a) they functionally implement the excitatory (cohering) and inhibitory (incohering) weights between pairs of data and hypothesis units, (b) they are selectively recruited in fitting a hypothesis to data, and (c) they allow the system to learn through feedback which explanatory principles are useful in achieving maximum network coherence.

Although we have not worked out all the details, one way to accomplish this might be to model each explanatory principle as a multiplicative "gating" unit (Hinton 1981) that modulates the excitatory or inhibitory connection between pairs of Thagard's present units related by the corresponding explanatory principle. Thus, if two propositions cohere due to the "analogy" principle, for example, the link between them will be gated by the "analogy" gating unit such that their mutual excitation will occur only if the "analogy" unit is also active (see Figure 1A). Similarly, if two propositions incohere due to some explanatory principle, the link between them will be gated such that their mutual inhibition will occur only if the relevant gating unit is active (see Figure 1B). In this way, the links that represent coherent and incoherent relations (a) can be effectively "labeled" by their explanatory principle and (b) can be selectively turned on and off depending on whether the relevant explanatory principle is "recruited" by the relations among relevant units when the to-be-evaluated hypothesis unit is clamped on. The recruiting is accomplished naturally by Hinton's gating units because of how the three-way multiplicative connections work: The product of each pair of units is transmitted to the third. This means not only that the explanatory unit will influence the activations of the datum and hypothesis units, but also that the activations of the hypothesis and datum units will influence the activation of the explanatory unit in the appropriate way. The latter operation has the desired effect of selectively turning on the explanatory units as needed, thus dynamically recruiting explanatory principles in the process of evaluating the network's coherence vis à vis the clamped hypothesis.

Although such a network is based on Thagard's ECHO model, it has distinct advantages for modelling automated hypothesis evaluation within a dynamic paradigm. First, explanatory principles are contained within the model itself–in the form of the explanatory units–and thus play a direct and crucial role in evaluating explanatory coherence. This has the desirable feature of allowing that, with a host of more complex reasoning procedures, the system could actually *figure out* what relations hold between its network units and could make the necessary adjustments to represent these relations. Thagard's present network cannot possibly do this, because it does not contain the principles of explanation in any explicit form. Second, additional mechanisms could be incorporated that would amplify or attenuate the activation of specific explanatory units to reflect whether the corresponding modes of explanation are in or out of favor within the current paradigm. This could be modeled by weights between another "special" unit that is always on and the explanatory units, exciting some and inhibiting others. Third, and most important, learning mechanisms could be added that would automatically adjust the amplification/attenuation of paradigmatic explanatory units in keeping with feedback about which kinds of explanations have proven useful in previous analyses. This would allow the network to change the basis of the paradigm over time as evidence accrues that certain modes of explanation are valuable in evaluating the coherence of scientific hypotheses. In some small percentage of cases, these changes in the underlying network of explanatory units might be sufficiently synergistic that an analogue of true Kuhnian "paradigm shift" would emerge.

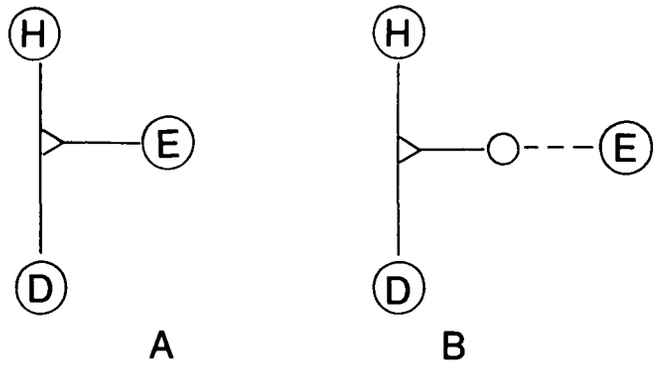

Figure 1 (Mangan and Palmer). Modeling explanatory principles as "gating" units in a connectionist network. Triangular symbols represent special connections in which the product of each pair of units gets transmitted to the third. Figure 1A shows how an *excitatory* connection between a datum unit (D) and a hypothesis unit (H) can be implemented by an explanatory gating unit (E), and Figure 1B shows how an *inhibitory* connection can be implemented by means of an intermediate inhibitory unit. (The dotted line represents an inhibitory connection.)

In summary, if this modified architecture works, it should model some additional features of scientific cognition not captured by ECHO. Among these are: Paradigm or explanatory units will manifest various levels of salience by virtue of their weight differences, thus operating as an intrinsic part of the system rather than as a discrete set of external principles; a hypothesis unit will recruit its most compatible explanatory principles as it recruits its data; the paradigm subsystem will have the property of stability without sacrificing the ability to change substantially under, say, data pressure. In other words, if this model (call it PAN for Paradigm Analogue Network) is given data and hypotheses sufficiently different from those on which it was trained, the weights connecting the paradigm units should slowly change. This would, of course, not only change the character of the paradigm subsystem and coherence of the data and the hypotheses, but the principles of explanation would simultaneously reconfigure. In this way PAN, if it could work, would move closer to modeling paradigm shift in its normal sense. We want to emphasize, however, that PAN is only meant to illustrate how an ECHO-like network might conform more closely to current thinking about the process of scientific evaluation of hypotheses and so support Thagard's original intuition–namely, that connectionism may prove useful in probing the nature of science itself.

Acceptability, analogy, and the acceptability of analogies

Robert N. McCauley

Department of Philosophy, Emory University, Atlanta, GA 30322

Thagard proposes a model of explanatory coherence for the evaluation of competing explanatory hypotheses in which the acceptability of propositions can depend, at least in part, on analogical relationships that might exist between a promising hypothesis and a successful one as well as between their explananda (as summarized in his principle 3). Thagard's enthusiasm

about the contribution of such analogies is considerable. He repeatedly (and quite justifiably) cites as one of its outstanding advantages ECHO's ability to factor such analogical relationships into the assessments it makes. More important, Thagard sets the default value of "analogy impact" at 1, which insures that "the links connecting analogous hypotheses are just as strong as those set up by simple explanations." Although it will typically constitute only one of many factors affecting ECHO's judgment, Thagard acknowledges that at this setting "analogy can have a very strong effect."

In light of Thagard's aspirations concerning ECHO's psychological plausibility, it may be important that in controversies about explanatory power, analogy frequently does not (and probably should not) have such strong effects in people's deliberations, because the strength of its effects turns on the theoretical commitments of the reasoner ECHO models. Who that reasoner should be is not completely clear. Thagard becomes more inclusive as the paper progresses:

(1) In the scientific cases, Thagard reconstructs the arguments from the perspectives of the winners. Note that Table 1 includes only one of the evidential propositions (viz., E2) that the losers had advanced uniquely, and Table 3 has none.

(2) In the legal cases, by contrast, Thagard seems to reconstruct plausible outcomes of jurors' deliberations that take into consideration (sometimes conflicting) evidential propositions from both the prosecution and the defense.

(3) At the end of the paper, Thagard claims finally that ECHO is out "to capture both what people generally do and what they ought to do."

These comments raise the question of whose reasoning processes guide the programmer and at what stage in the debate they do so. The issue, in short, is how in any particular case the programmer decides what gets coded as ECHO's input. (I am suggesting that there is an important *dis*analogy between the scientific and legal cases to which Thagard applies his model.)

Analogies are not the sorts of things that programmers can code neutrally–nor, for that matter, are summaries of evidence or even contradictions. The problem with analogies cuts more deeply than the other two for Thagard's model (a) because it is usually not too difficult to get disputants to agree that they are disputing, or even to agree about the foci of their disputes, and (b) because Thagard presumes from the outset (illicitly, I fear) the ability of inquirers (regardless of their theoretical orientations) to recognize when any proposition explains another. (As he states repeatedly, the point of this project is *not* to offer a theory of explanation.)

The problems with explicating analogical reasoning are legion, because virtually anything (from one standpoint or another) can be analogous (on one count or another) to virtually anything else. The specific problem here concerns precisely the fact that there is no such thing as analogy *simpliciter*. Not only the importance but even the mere possibility of an analogy is in the eye of the (theoretically influenced) beholder. Analogies make sense only from some (often implicit) theoretical standpoint or other. Again, the problem for Thagard is whose standpoint reigns in defining ECHO's input and, in particular, its input about analogies. The problem is especially clear if ECHO is to apply to cases of scientific reasoning (as with the cases of legal reasoning that Thagard discusses) in the midst of the debates, that is, before the case is settled.

Darwin's attempted analogy between artificial and natural selection, for example, does not move the nineteenth-century creationist. The salient point for the creationist is that artificial selection has never resulted in a new species. The traditional Darwinian replies that speciation requires much more time. Unfortunately, the nineteenth-century creationist (at least) remains unimpressed. If speciation is impossible, as creationism maintains, then additional time is irrelevant. Furthermore, most Victorian creationists had confidence, if not in the pronouncements of Bishop Ussher, then certainly in those of Lord Kelvin, who offered assurance that there was not nearly time enough.

The point of this is not to defend creationism ("devil's advocacy" is an inappropriate turn of phrase here on two counts), but rather to emphasize that during theoretical disputes the acceptability of proposed analogies is precisely one of the points at issue. It is exactly when disagreements about the relative merits of competing explanatory hypotheses arise that determinations about the acceptability of theoretically inspired analogies is up for grabs. It is only after the resolution of such debates, when one of the competing explanatory theories emerges triumphant, that we confidently pronounce on the value of various analogies. Consider Francesco Sizi's argument against Galileo that there must be seven planets (and therefore no moons around Jupiter) because human beings by nature have seven holes in their heads (Hempel 1966)!

If ECHO offers normative guidance about the role of analogy in explanatory reasoning, then, concerning any particular analogical proposal, it only does so, at best, after the fact. But that could well be the place to which epistemology has come.

Optimization and connectionism are two different things

Drew McDermott

Computer Science Department, Yale University, New Haven, CT 06520
Electronic mail: *mcdermott@cs.yale.edu*

My main objection to Thagard's target article concerns its emphasis. The word "connectionism" is really out of place in it. The whole idea of connectionism is that mental function ought to be modeled by devices consisting of large numbers of smallish units operating in parallel and communicating via fixed links–that is, the way the brain presumably does it. Thagard's paper proposes a model of explanatory coherence based on minimizing a certain energy function. The independent variables are the activation levels of various propositions. The objective function is a sum, H, of terms that express the support and inhibition relationships between these propositions. (See equation 1.) It is not clear how to judge whether H is a good measure of explanatory coherence, but the use of connectionist "settling" techniques is a distraction. They are probably unnecessary, because the number of independent variables is quite small. I am no expert, but I would guess that standard numerical-optimization techniques (e.g., conjugate-gradient descent) would do better than simulating a network of "units"; and they might focus attention better on the properties of the H function.

Connectionist techniques are distracting in another way, because their use inevitably suggests that the author is hypothesizing the existence of fixed locations for various hypotheses. He isn't, of course. His program must wire up the network anew for every problem. (Many connectionist papers speak as if the inevitable software simulation were a stopgap until the hardware arrives, but Thagard can't do that: His whole approach is based on software.) The apology for the use of connectionism in section 8.1 is really quite puzzling. There is no more connectionism in this algorithm than there is in GPS.

With this confusion cleared away, we can examine the real issue raised by the target article, which is whether the H function is a good measure of explanatory coherence. My conjecture is that it's completely adequate. As with all measures of success in nonmonotonic inference problems, the details of the measure function are swamped by the properties of the algorithm that generates things to measure. That's because if this algorithm overlooks something important, any fine-tuning of the combination of the remaining factors is futile. Unfortunately, we know very little about explanation generation. So it's too early to tell whether, for instance, it is an asset or a liability that

Thagard's algorithm gives weight to links between propositions based purely on the structure of the problem, and not on the content of those propositions.

Coherence and abduction

Paul O'Rorke

Department of information & Computer Science, University of California, Irvine, CA 92717
Electronic mail: *ororke@ics.uci.edu*

The theory of explanatory coherence presented by Thagard focuses on the problem of selecting a particular explanation from among given competing alternatives. He presents an interesting set of principles designed to capture the notion of explanatory coherence and provides a connectionist method for evaluating competing explanations. Thagard's ideas are stimulating and worthy of further study. At present, however, I see three major problems with his approach. First, he appears to take a modular, sequential approach to the construction and evaluation of explanations. Second, coherence seems to be the only criterion used to decide whether to accept or reject explanations. Third, there seems to be no distinction between passive, unconscious acceptance processes and active, conscious evaluations of competing explanations.

What's wrong with a modular, sequential approach to constructing and evaluating explanations? Thagard's system is given explanations, but ECHO does not address the problem of how they can be generated. At the University of California at Irvine, my students and I have built a number of computer programs for automating abduction. Initially, we tried to maintain a separation between our models of the processes responsible for constructing explanations on the one hand and our evaluations of explanations on the other. However, in computational experiments involving physical and psychological explanations, our initial systems sank in seas of explanations, most of which were totally implausible. We were forced to introduce evaluation into the construction process in order to control the search and reduce the number of explanations generated. We now believe that construction and evaluation must be integrated, and coherence must play a role in the generation of explanations.

What's wrong with coherence as the sole criterion for deciding whether to accept or reject a theory? Unfortunately, coherence is not the only factor that plays a role in constructing and evaluating explanations. We have found that an agent's goals and priorities play important roles in evaluation. For example, in diagnosis one typically constructs explanations of abnormal behaviors of devices or systems. In diagnoses that occur in diverse application areas such as medicine, space technology, and so on, the consequences of explanations turning out to be correct or incorrect play important roles in evaluating the explanation. Engineers working on diagnostic systems for the space station, for example, are explicitly directed by NASA to develop systems that attend not only to the plausibility of explanations but also to the associated risks. A flaw in critical life support, even if it is considered highly implausible, should be attended to sooner than a more plausible, but less dangerous flaw because we generally give high priority to staying alive.

In Thagard's example of the Peyer trial, and in legal examples in general, the fact that a decision against the defendant entails undesirable consequences for him should play a role in the decision-making process. When a jury decides a case, they not only reason about what happened; they also reason about what will happen as a result of their actions. They worry about the possibility that they will unjustly let a criminal go unpunished, or punish an innocent. In our society, a strong desire not to punish innocents has been institutionalized in the concept of guilt "beyond a reasonable doubt." Thagard speculates that this sort of bias might be implemented by tweaking ECHO parameters, but it is likely that machinery for reasoning about goals, goal priorities, and consequences of actions with respect to goals will also be necessary.

Actions and conscious reasoning play an important role in evaluation. My final major worry is that Thagard's principles of coherence and connectionist implementation are best suited to modeling passive, unconscious sorts of acceptance processes. One can imagine this sort of evaluation and adoption of explanations taking place subconsciously–for example, during natural language processing. Waltz and Pollack (1985) describe a connectionist model similar to ECHO that parses "semantic garden-path" sentences so as to produce activation histories that seem to simulate our own subjective experiences with these sentences. Given a sentence such as "The astronomer married the star," their system goes through a sequence of patterns of activation representing different interpretations of the sentence, just as people seem to realize unconsciously that the celestial object meaning of the word "star" cannot be the object of marriage so the astronomer must be married to an actress.

The evaluation of explanations associated with complex diagnoses, legal arguments, and (especially) scientific theories seems to require much more active reasoning and decision-making processes. In some situations it is not necessary or even prudent to accept one of a competing set of hypothetical explanations. Instead, it may be necessary to try to gather more information. In addition, it is often useful to distinguish between explanations that can and cannot be acted on, because the former tend to be more useful. In diagnosis, an explanation that pins down the source of a fault to particular malfunctioning components is more useful in that it suggests the obvious repair plan of replacing the bad parts. Similarly, in science, theories that make predictions and suggest actions that can be taken to verify or falsify the predictions are preferred over theories with no observable consequences.

This argument suggests that a full account of the difficult examples chosen by Thagard will probably have to include computational models of rational deliberation and planning. This is one reason why some artificial intelligence (AI) researchers believe that abduction (at least as AI researchers use the term) is probably "AI-complete." Rather than taking this as an indication that modeling abduction is impossible or that Thagard has taken on an intractable problem, I take this as a sign that he is working out a view of a piece of a problem of fundamental importance. It will be interesting to see his ideas combine and compete with ideas proposed by AI researchers and others attempting to discover principles underlying cognitive processes associated with explanations.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under grant number IRI–8813048.

Probability and normativity

David Papineau

Department of History and Philosophy of Science, University of Cambridge, Cambridge CB2 3RH, England

ECHO is an elegant and impressive program, but I have doubts about some of the philosophical and psychological claims made on its behalf.

Thagard argues that the model of theory evaluation embodied in ECHO is superior to probabilistic models. In particular, he doubts the availability of the various probability judgments that probabilistic models need as input (sections 8.2 and 10.3).

However, ECHO itself assumes independently given "explanation" and "contradiction statements" as input. Often these

seem little different from probability judgments. For example, in section 2.3 (para. 8), Thagard says that alternative explanations are treated as contradictions because "their conjunction is unlikely." More generally, ECHO deals with choices between competing explanatory hypotheses, but not with which hypotheses are allowed to enter the competition. This, too, arguably presupposes judgments of prior (im)probability.

It is true that ECHO, at least as so far applied (but see section 4.1), starts with nonquantitative "explanation statements," rather than numerical probability inputs, and correspondingly yields "on-off" conclusions, rather than numerical probability outputs. Thagard takes this to enhance the psychological realism of ECHO (sections 3, 8.2, 10.2). However, even if we grant Thagard this psychological realism for the moment, we can still have doubts about the *normative* significance of ECHO. For even if our natural psychological inclinations are qualitative, surely it would be better, in both scientific and legal contexts, to be sensitive to the prior probabilities of explanatory hypotheses and the varying degrees to which they render the evidence unsurprising, and to have a wider repertoire of responses than a simple "yes" or "no."

Moreover, given this normative point, we can then ask further questions about ECHO's psychological realism. For, after all, theory choice in both law and science is a highly self-conscious enterprise, where practitioners are quite capable of recognizing that it is better to reason probabilistically. And I would argue that practitioners in both areas *have* widely recognized this, and so often *do* reason probabilistically. (For example, in civil cases juries are explicitly required to decide "on the balance of probabilities," in contrast with criminal cases whose special circumstances require guilty verdicts to be "beyond a reasonable doubt.")

Even if simple qualitative evaluations of explanatory hypotheses are in some sense "natural" to human beings, self-conscious reflection can nevertheless lead to different people in different contexts opting for more sophisticated quantitative ways of evaluating hypotheses. This seems to me to make it doubtful that "hypothesis choice" is a well-defined psychological category in the first place.

There is a general moral here: Model-builders who are after psychological realism should concentrate on mental processes like visual pattern recognition or speech processing and should shun the kind of mental process traditionally discussed by philosophers. For the latter processes are precisely of the kind whose identities are constantly being transformed and fragmented by self-conscious normative reflection.

Explanatory coherence in understanding persons, interactions, and relationships

Stephen J. Read[a] and Lynn C. Miller[b]

[a]Department of Psychology, University of Southern California, Los Angeles, CA 90089-1061 and [b]Department of Psychology, Scripps College, Claremont, CA 91711

Electronic mail: *[a]read@uscvm.bitnet*

We examine the implications of Thagard's model of explanatory coherence for two major issues in social psychology and the psychology of personality: (1) the role of coherence in a recently proposed theory of attribution, and (2) the role of coherence in individuals' models of themselves and others.

Coherence in attribution theory. Recently, several authors (e.g., Lalljee & Abelson 1983; Read 1987) have presented a theory of attribution based on Schank and Abelson's (1977) knowledge structure approach. According to this theory, the typical attributional problem is to explain a sequence of actions involving one or more individuals. To understand such sequences, people use detailed social and physical knowledge to construct a causal scenario that characterizes how the actions of the individual(s) hang together to form a plan aimed at the attainment of some goal(s) (Read 1987). Thus, social explanation is akin to creating a story of how actions by individuals go together.

To understand a sequence of behaviors, people must often characterize it in terms of higher-order structures such as goals, themes, scripts, or plans, which describe the relations among the actions and go beyond the individual actions (Abelson & Black 1986). For example, going beyond the causal relations of the individual actions and recognizing that a sequence of events is a drug arrest suggests that the participants are acting as they do because of their roles as drug dealers or undercover police officers. Or realizing that a sequence of events is an assault rather than two lovers playfully wrestling leads to very different explanations of the behavior.

How do we decide whether a particular structure is appropriate? One major criterion is how coherent it would be with the actions (Read 1987; Wilensky 1983). Thagard's model provides an elegant approach to understanding how people might choose among alternative knowledge structures as characterizations of action sequences.

Different scenarios can be constructed out of the same set of facts, using different knowledge structures. Which knowledge structures are chosen and which scenario is constructed depends on which is more coherent. For example, if two different structures, lovers' wrestling and assault, were potentially applicable, we should prefer the one that requires fewer assumptions (simplicity) and is able to handle more of the sequence (breadth). In addition, we might prefer structures that are consistent with previous interpretations of similar events (analogy; e.g., we recently observed a couple roughhousing in a park). The idea that a hypothesis will be more coherent if explained by other hypotheses further suggests that a characterization of an event sequence would be more coherent, and thus more likely to be selected, if it could be explained by other features of the persons involved, such as personal characteristics, goals, or abilities. Finally, Thagard's model suggests that we should be unsatisfied with the application of a structure to a sequence if it leaves many of the facts and events unaccounted for.

Coherence of models of personality and persons. How do the various behaviors, beliefs, and motives of a particular person "fit together" to form a coherent system? This question is an old and important one for personality theory (Allport 1964; Read & Miller 1989a; 1989b), but one that is particularly troublesome methodologically.

Let us consider how Thagard's simulation may help us explore such idiographic coherence quantitatively. First, Thagard argues that

> a system S will tend to have more global coherence than another if (1) S has more data in it; (2) S has more internal explanatory links between propositions that cohere because of explanations and analogies; and (3) S succeeds in separating coherent subsystems of propositions from conflicting subsystems. (sect. 2.3, para. 12)

Individuals are likely to differ in the extent to which their systems (e.g., belief systems, self-system) cohere. Thagard's model suggests, for example, that more coherent self-systems would be those in which a given individual has more accessible self-relevant data, those in which there are more "internal explanatory links" between beliefs, behaviors, and self-conceptions, and those that "succeed in separating coherent subsystems" of beliefs, behaviors, and self conceptions "from conflicting subsystems." Thus, an individual might understand that one subset of behaviors, self-conceptions, and beliefs coheres under some circumstances (e.g., with a close friend to whom one can make intimate disclosures) but a different set of beliefs, self-conceptions, and behaviors would cohere under different circumstances (e.g., with strangers). Such conflicting subsystems might then cohere at a higher level for people because they recognize a higher-order explanation that links

them (e.g., avoiding rejection is usually an important goal but is deactivated when with accepting friends).

We could also assess *how* the system coheres–that is, we could examine why behavioral observations and beliefs cohere for an individual the way they do. How does the individual weigh beliefs and behaviors in selecting hypotheses and what hypotheses and analogies support this belief system? What would happen if different alternative hypotheses about the self were introduced? What would it take to change the activation of leading hypotheses for a particular individual?

Thagard's approach also allows us to examine why different individuals' models of the same person differ. Presumably, individuals who have more data about the person being examined (e.g., close friends) will have more coherent representations. Also, the order in which we are exposed to various pieces of information can affect the likelihood that an individual will select a given hypothesis and retain it (even in the face of counterinformation and equally plausible alternative hypotheses). Analogies to past relationships and preexisting knowledge structures (e.g., stereotypes) may also bias the process of building models of persons in the current relationship.

In addition, for models of self, interactions, others, and relationships, how does the coherence of the models change as new information is added to the system? Could we examine changes in such systems developmentally, during therapy, or during the development and dissolution of relationships? These are all exciting questions, and Thagard's model may prove an important step in providing a methodology to address them idiographically–that is, at the level of the unique individual or relationship.

Measuring the plausibility of explanatory hypotheses

James A. Reggia

Departments of Computer Science and Neurology, University of Maryland, College Park, MD 20742

Electronic mail: *reggia@mimsy.umd.edu*

Thagard's theory of explanatory coherence (TEC) provides a broad and useful framework for considering the plausibility of explanatory hypotheses. Because TEC is intended to apply to "reasoning in everyday life," it seems appropriate to compare it with a related but less general framework, parsimonious covering theory (PCT), which also provides an application-independent theory of explanatory coherence (Reggia et al. 1983; 1985). PCT differs from TEC in that it is restricted to consideration of explanatory hypotheses in general diagnostic problem solving, although it has been adopted for a number of nondiagnostic applications. Because of its restricted applicability, it does not address some issues of TEC (e.g., analogy). However, like TEC, PCT precisely defines the notion of explanatory hypotheses and what makes them plausible, has been applied to specific applications, and has been formulated as a connectionist model (Peng & Reggia, in press; Wald et al., in press). Because of space limitations, I will restrict my attention to comparing TEC's measure of "degree of coherence" (Principle 2c) to the related notions of plausibility and probability of explanatory hypotheses in the simplest version of PCT. Our experience with PCT and diagnosis suggests that counting propositions (TEC Principle 2c) is an inadequate measure of "coherence" or plausibility. (Principles 6b and 7 also appear to conflict with PCT, but are not considered here.)

In the simplest form of PCT, there is a set of disorders, D, and a set of manifestations ("symptoms"), M. For each disorder, d_i, there is a connection (association) between d_i and each manifestation, m_j, that can be caused by d_i. A subset of M, denoted M^+. A set of disorders, D_I, is called a cover of the given M^+ when the disorders in D_I can cause all of the manifestations in M^+. A set of disorders, D_I, is an *explanatory hypothesis* if (1) D_I is a cover of M^+, and (2) D_I is parsimonious. Roughly speaking, asserting the presence/absence of manifestation m_j or disorder d_i in PCT corresponds to a proposition in TEC, and a parsimonious cover represents specification of the function defining system coherence (TEC Principle 7).

A difficult problem in diagnostic reasoning theories in general, and in PCT in particular, has been how to define precisely what is meant by the "best," "most plausible," "simplest," or "most parsimonious" explanation for a given set of facts (deKleer & Williams 1986; Josephson et al. 1987; Peng & Reggia 1987; Pople 1973; Reggia et al. 1983; 1985; Reiter 1987). Previous notions of plausibility have largely been based on *subjective* criteria; we consider two of these here.

An early criterion of plausibility used in PCT and by others is similar to TEC Principle 2c. It is called *minimal cardinality*: Explanatory hypotheses with the smallest number of hypothesized components are preferable. In applying PCT to specific diagnostic problems, it quickly became evident that minimal cardinality is an inadequate measure of plausibility. For example, in medical diagnosis two common diseases are often more plausible than a single rare disease in explaining a given set of symptoms (Reggia et al. 1985); and in electronic diagnosis analogous examples exist (Reiter 1987). For this reason, PCT as well as other models of diagnostic inference have adopted a more relaxed criterion of plausibility called *irredundancy*: A set of disorders, D_I, that covers (causes all of) the manifestations in M^+ is irredundant if it has no proper subsets that also cover M^+. Although it does not favor the smallest set of propositions (as does TEC Principle 2c), irredundancy is a preferable criterion because it handles cases like the medical and electronics examples referenced above while still constraining the number of disorders in a hypothesis. However, irredundancy has the problem that in larger applications it may identify many implausible hypotheses as well as the plausible ones; and as indicated below, in some cases it may still fail to identify the most reasonable hypothesis.

The criteria used in most theories of explanatory plausibility, including those of TEC and PCT, are *subjective*. An important question is whether one might devise *objective* measures of plausibility and then ask under what conditions various subjective criteria would work or fail according to the objective criterion. We have recently generalized Bayes's Theorem to apply to a restricted class of diagnostic problems formulated in PCT (Peng & Reggia 1987). Each disorder, d_i, is associated with its prior probability, p_i. Each casual link is associated with a number, c_{ij}, the causal strength from d_i to m_j representing how frequently d_i causes m_j. Under assumptions less restrictive than those traditionally made with Bayesian classification, the relative likelihood $L(D_I, M^+)$ of any potential explanatory hypothesis D_I given the presence of M^+ can be calculated using relevant p_i and c_{ij} values. Using the objective, albeit limited, measure $L(D_I, M^+)$, one can ask under what conditions various plausibility criteria such as minimal cardinality, irredundancy, and others would be guaranteed to identify the most probable hypothesis.

Analytical treatment of this question leads to a number of interesting results (Peng & Reggia 1987). For example, minimal cardinality is an appropriate criterion only when, for all disorders, d_i, the prior probabilities are very small and about equal, and the c_{ij} are fairly large in general. Otherwise, it may be that the most probable explanation does not have minimal cardinality, supporting the conclusion above that counting is not sufficient.

Thagard points out (section 8.2) correctly that in some nondiagnostic domains the probabilities do not exist. They do not

really exist in diagnostic applications either. However, because TEC and PCT are intended to be theories that encompass diagnostic reasoning, they cannot ignore measures of likelihood that go beyond counting, be they numeric probabilities or other nonnumeric, subjective measures. Some measure of "prior plausibility" or "intrinsic merit" and "conditional plausibility" of causation is essential in diagnosis and seems to me to be just as important in scientific and legal reasoning (Thagard may agree with this to some extent; see section 4.1 on ECHO). Basing coherence on counting propositions as in TEC Principle 2c would therefore appear to need revision, at least to encompass diagnostic inference.

TEC provides a broad and useful framework for considering these and related issues. Although one can always argue about specifics, as I have done here, the overall thrust of Thagard's work strikes me as being in the right direction, and it will be very interesting to follow its evolution.

ECHO and STAHL: On the theory of combustion

Herbert A. Simon

Department of Psychology, Carnegie-Mellon University, Pittsburgh, PA 15213

Electronic mail: *has@cs.cmu.edu and has@a.gp.cs.cmu.edu*

Thagard takes the theory of combustion as his first example of how ECHO uses "explanatory coherence" to evaluate scientific theories: in this case to choose between the competing oxygen and phlogiston theories. A similar task was undertaken a few years ago, using an entirely different computational architecture, the STAHL program of Langley et al. (1987). We have here a valuable opportunity to compare connectionist and symbolic solutions to problems of this kind. What tasks do the two systems perform, respectively? What kinds of information and assumptions have to be provided them? What heuristics do they use in their reasoning? What kinds of conclusions can they reach?

The tasks. ECHO's task is to compare the relative compatibility of two theories with a body of evidence. STAHL's task is to derive a theoretical explanation of a body of evidence. Although the tasks of ECHO and STAHL seem very similar, we shall see that the two programs differ drastically in terms of the information that must be provided to them by the user, and in terms of how they use the empirical evidence to reach their conclusions. STAHL requires far fewer and weaker givens than ECHO and makes its own inferences about logical relations among the propositions; these relations must be posited by the user of ECHO.

The givens. ECHO must be supplied with sets of propositions to (1) represent the empirical evidence, (2) represent the two sets of competing propositions, and (3) designate which propositions support or contradict which other propositions (see Thagard's Tables 1 and 2). In addition, ECHO is provided with signed weights for the members of (3), whose magnitudes are adjustable system parameters.

ECHO makes no use of the semantic or syntactic structure of any of the propositions of (1) and (2), but only their names as these appear in the expressions of (3). Hence, the logical connections among the propositions are not inferred from them but are posited by the user, as are the initial "strengths" of these connections.

In contrast, STAHL is supplied only with propositions that correspond to ECHO's evidence propositions (1). STAHL infers additional propositions by using a small set of heuristics to reason from the evidence. The logical connections among propositions emerge from the structure of the propositions themselves, without requiring the input of either explanatory or contradictory propositions like those exhibited in Thagard's Table 2. Moreover, STAHL has no parameters to represent "strengths" of connections.

All of STAHL's inputs, and its inferred propositions as well, are (qualitative) descriptions of chemical reactions in terms of inputs and outputs. For example, given the inputs:

(reacts inputs {charcoal air} outputs {phlogiston ash air})
(reacts inputs {calx-of-iron charcoal air} outputs {iron ash air})

STAHL infers:

(components of {charcoal} are {phlogiston ash})
(components of {iron} are {calx-of-iron phlogiston})

The latter two statements can be recognized as a standard form of the phlogiston theory. When the input reactions are changed–for example, to specify that input of red-calx-of-mercury yields the outputs mercury and oxygen–STAHL arrives at Lavoisier's oxygen theory.

The heuristics. ECHO uses a connectionist scheme for assigning weights to (the names of) hypotheses on the basis of weights exogenously assigned to (the names of) evidential propositions and linkages. ECHO has no way of determining endogenously the logical relations among propositions, whether supportive or contradictory.

STAHL's reasoning is based on five heuristics for inferring the components of substances from the inputs and outputs of chemical reactions involving these substances (Langley et al. 1987, pp. 228–34). For example, the IDENTIFY-COMPOUNDS heuristic reads: "If A is composed of C and D, and B is composed of C and D, and neither A contains B nor B contains A, then identify A with B." The other heuristics are called INFER-COMPONENTS, REDUCE, SUBSTITUTE, and IDENTIFY-COMPONENTS. The inferences STAHL makes depend on the order in which reactions are presented to it; it can back off from reasoning that produces contradictions and try alternative analyses (Langley et al. 1987, pp. 242–45). It cannot be stressed too strongly that STAHL does carry out actual reasoning on the basis of heuristic inference rules drawn from the practice of chemistry.

The conclusions. STAHL's reasoning is limited to chemistry. To operate in another domain, it has to be provided with a knowledge representation and heuristic inference rules for that domain. ECHO is quite general, but only because all of the domain-specific knowledge is provided to it by the user in each application and it is oblivious to the content of its propositions. Moreover, STAHL invents its own hypotheses, whereas ECHO must be provided with them. ECHO therefore operates at a much more superficial level than STAHL.

Both STAHL and ECHO will corroborate the oxygen theory of combustion if given Lavoisier's "facts," and the phlogiston theory if given STAHL's "facts." The difference in interpretation depends on whether one attends to the oxygen input and carbon dioxide output to the combustion reaction, or to the output of flame and smoke (caloric), respectively. Contrary to popular accounts, the advance to the better theory did not depend on Lavoisier's quantitative measurements, but on the growing awareness of the participation in the reaction of the enveloping gases, and the striking of heat and flame from the list of "substances." STAHL, which uses no quantitative information, makes the switch in interpretations quite readily (Langley et al. 1987, pp. 248–51).

Summary. A comparison of ECHO with STAHL in application to the theory of combustion shows that the latter provides a far deeper account of the development of theory than the former, handling endogenously many of the elements that must be provided to ECHO as givens. STAHL carries out genuine reasoning; ECHO does not.

Theory autonomy and future promise

Matti Sintonen

The Academy of Finland and Department of Philosophy, University of Helsinki, 00170 Helsinki, Finland
Electronic mail: *msintonen@cc.helsinki.fi*

I much admire Thagard's way of baking a variety of virtues into explanatory coherence. My queries and suggestions center on the notion of promise. Take Darwin's theory. Although Darwin thought that its greatest asset was its capacity to group and explain *classes* of phenomena, few details were available. The theory was more akin to a research program, with DH2–DH3 as basic tools, and E1–E15 as problem areas (subdomains or applications; see Thagard 1978). I understand that singular evidential propositions have been left out for expository reasons, but there is also an issue of principle. The choice seems to be between a promising project and already articulated rivals, and this involves pitching possible future unification against less unified but established breadth.

I wonder how ECHO handles promise. Although the notion is needed (see section 4.4), ECHO looks back to certified explanatory propositions and not forward to future prospects. ECHO inputs are tied to an instant in time, and so are the verdicts for global coherence in the connectionist formula (see section 4.9). Does ECHO, unlike Thagard's earlier account (1978) of dynamic consilience, gesture towards instant rationality (see Lakatos 1970)?

This cannot be the intent, and indeed ECHO is prepared to give a hypothesis a new hearing if new explanatory sentences are added (section 4.2). More dynamics can be brought in if not only rival theories but also consecutive versions of a theory can be brought to trial, so that H(t') may be higher than H(t), for t' later than t. But note that dynamic tinkering with theory presupposes identity through change and a clear notion of theoretical commitments.

Take theory identity first. There is of course the Darwinian core DH2–DH3, and to ECHO's credit, structure emerges *as a result* of its operation, crystallizing in high connectivity. (Note, though, that to explain DH2, DH4 should be appended with the cohypotheses that some variation is relevant to survival and reproduction, and that properties are largely inherited.) How helpful this is is hard to say, for the onus of deciding what explains what is still on the programmer. PI (processes of induction; section 7) seems helpful in tracing origins, but PI rules hover on the same conceptual level as explanatory propositions. The principal reason for being suspicious about automatic input is that a formal-logical inference can be given explanatory and nonexplanatory interpretations, with no formal way to distinguish between them (see Gärdenfors 1976).

What ECHO could do is acknowledge more clearly a hierarchy with core hypotheses on top and auxiliaries under it: The former carry the banner while the latter reach toward so far unconquered subdomains. The division is visible in Table 3, but the very idea of promise indicates that the two are on different levels. The latter are not all at hand when the core is proposed, and they are not as central. Actually, matters are more complicated: Some auxiliaries span entire subdomains, whereas others are very short-lived and needed for singular explanations. A hierarchy of levels "harmanizes" (section 4.9) with ECHO: It allows for layers of explanation (cf. section 6.1), avoids crude holism (section 10.1), and organizes constraints. For unification has to do with subdomains brought under one umbrella, whereas simplicity is a constraint on auxiliaries within a subdomain.

Next, a remark about pragmatic reference to "explanations and hypotheses actually proposed by scientists" (section 2.3). The text gives the impression that these are needed to discredit explanations with irrelevant cohypotheses: ECHO is discharged of the obligation to consider P1&P2 (where P2 is irrelevant), because it doesn't surface as a serious option. But ECHO does better than this: Irrelevant premises decrease coherence, and therefore overt pragmatic reference here is otiose. Principles 2(b)–(c) explain why scientists actually do not flirt with irrelevant cohypotheses (and why explanation is more restrictive than implication). But this may be the intended reading anyway.

With these pragmatic hurdles behind, there is the tough one of theory claim ahead. Choosing between networks presupposes shared data. But theories have some *autonomy* in selecting the territories they claim and in slicing them into subdomains. This problem should worry whoever feeds ECHO, for it conspires with the problem of promise, threatening to let autonomy run rampant. Actual theories often promise to carry you through thick and thin, but have relatively scant justification in terms of detailed results. Moreover, the secured results may concentrate on a few subdomains, and the claim of a theory to handle others may be frustrated, either because the main hypotheses are unsuitable or because carrying out the detailed program flouts other constraints. Thus prospected unification or simplicity may crumble on closer inspection.

One reaction is to downplay troubles, as in cognitive dissonance elsewhere. ECHO acknowledges that not all data are treated equally. Theory autonomy explains why data excitation values differ, suggesting another interpretation for numerical parameters in explanatory statements like (EXPLAIN (H1) E1.9). A weakened link between H1 and E1 could reflect either E1's dubious epistemic status (Thagard's proposal) or, equally well, its dissimilarity with unproblematic exemplars within a class that H1 should address. E5 in Figure 8 could be epistemically impeccable but deactivated because it is marginal to H1's concerns.

As to autonomy, there is no way to force a problem on a theory unless it commits itself to a set of domains and paradigm explananda in these domains. Commitment to main hypotheses and a set of domains would thus explicate relevance and explain how data, analogy, and simplicity can have differing weights. Thagard observes (section 4.10) that such contextual features are learned from established coworkers in the field. The same sociopsychological peer pressure sets limits to tolerance and skepticism.

Note, then, that a project is a gamble. Traditional decision-making models acknowledge this through expected cognitive value. True, precise probabilities are not easy to come by, but ECHO could add a dynamic feather to its hat by allotting expected coherence a role, however modest.

Let me conclude with a note on broader vistas. Thagard rejects the normative/descriptive dichotomy, yet (rightly) insists on the cognitive nature of explanatory virtues (Thagard & Nowak 1988). But consider the legal examples, in which the prosecution and defense advocate incompatible ways of explaining the evidence. Clearly, more than "harmany" or truth is at stake. A juror may think that "guilty of first-degree murder" *maximizes "harmany," but he may think twice before speaking* out. Thagard surmises that jurors hesitate because hypotheses of innocence receive special activation (or require a high tolerance level). I agree, and suggest a reason: apart from cognitive considerations, jurors may (without, perhaps, being aware of this) keep an eye on the practical consequences that follow if the judge acts on a chosen cognitive verdict (as he usually must). The most "harmanious" option may have dramatic consequences for the defendant, which is why the moral and legal principle of safeguarding the innocent exists. I suggest that this principle brings in a noncognitive constraint that should operate on what one says, not on what one thinks. Thagard's examples suggest that being established "beyond reasonable doubt" runs these two aspects together.

Psychology, or sociology of science?

N. E. Wetherick

Department of Psychology, University of Aberdeen, King's College, Old Aberdeen AB9 2UB, Scotland

My problem with Thagard's target article is that it purports to present a connectionist theory of "the acceptance and rejection of scientific hypotheses" as an individual psychological (or perhaps neurophysiological) process, but cites only examples involving the acceptance or rejection of hypotheses by a community of scientists–a sociological process. It is not obvious to me that the same theory could apply to individual psychological processes as well as to sociological ones.

Thagard's examples come from the sociology of science. The phlogiston controversy seems to have been settled over what is a very short period if one considers the slowness of scientific communication at the time. It is natural to assume from appearances that something is lost by substances when they burn; smoke is given off, along with soot particles. But when a controlled experiment is done, it turns out that weight is gained by the burned substance exactly equivalent to the weight lost by the air in which it is burned. Something passes from the air to the substance, but phlogiston theory predicts the opposite. Though Lamarck, for example, and Goethe never ceased to advocate a "qualitative" science concerned with "the very nature" of things, the period was one in which the idea of a "quantitative" science was gaining ground prior to its triumph in the nineteenth century; that movement of thought was sufficient to ensure a preference for theories consistent with the quantitative evidence. ECHO picks up this effect (in which variations between individual scientists in the perceived relevance of different parts of the argument cancel out over the whole).

The controversy over Darwin's argument for natural selection mirrors the psychological process more closely–creationists are still active among us, but not phlogiston theorists. If someone accepts that species were created by God because that is what he tells us in the Bible, then no amount of geological or biological evidence will oblige him to believe in natural selection! God may have incorporated evidence suggesting natural selection (over a period of millions of years) when he created the world in 4004 B.C., just to test our faith in his word. This argument was seriously advanced in the nineteenth century, but in the nineteenth century there was also a movement of thought favouring arguments that did not depend on revelation but made a rational case for alternative naturalistic explanations of phenomena. ECHO picks up this effect too.

Thagard's examples evaluate hypotheses against evidence as if in a mind free of prejudice and preconceptions. No real individual mind qualifies, and I suspect that even the "mind" of the scientific community only appears to qualify because in the examples chosen the "true" hypothesis was the one consistent with a movement of thought that was, in any case, beginning to be accepted on much more general grounds. If Thagard modelled the mind of the scientific community of the 1920s on the subject of Polanyi's theory of the adsorption of gases on the surface of a solid, he would show that Polanyi was wrong, though in fact Polanyi was right. (Langmuir got the Nobel prize for being wrong!) At that time, the "movement of thought" was against the type of theory advocated by Polanyi (Polanyi 1963).

Thagard may argue that ECHO can perfectly well model a prejudiced mind, and so it can, but I question the value of a connectionist model of this particular type of mental activity. Perceived explanatory coherence is always determined by conscious symbolic processing, usually accompanied by discussion with other perceivers interested in the same problem–which would be impossible without symbolic processes. This is recognised in ECHO by the fact that the degree of relevance of evidence to hypotheses (and of analogy between hypotheses) has to be input by higher (symbolic?) mental processes. When we as individuals consider alternative hypotheses as explanations of a given set of phenomena, we may "err" (as shown above) because we attach exceptional weight to a particular hypothesis for reasons extraneous to science. Or we may be unaware of some piece of evidence that, we agree as soon as we hear of it, determines the issue in favour of one of the hypotheses. Or we may temporarily have forgotten some such piece of evidence and be willing to change our opinion as soon as we are reminded of it. In any case, symbolic processing will be involved, and Thagard's model contains no hint as to how the transition to this is to be achieved; it can account for no more than, for example, "incubation" in problem solving.

I conclude that nothing is to be gained at present by constructing a connectionist account of a psychological process that must involve symbolic processes. Thagard's model could be an account of the sociological process by which a community of scientists comes to adopt a common view on some theoretical issue, but his references to "excitation" and "inhibition" would then be entirely metaphorical, and that does not seem to be what he intended.

Testing ECHO on historical data

Jan M. Zytkow*

Department of Computer Science, George Mason University, Fairfax, VA 22030-4444

Electronic mail: *zytkow@gmuvas.gmu.edu*

A number of interesting phenomena related to the choice between competing theories were reproduced in ECHO on toy problems in section 4 of Thagard's target article. This makes ECHO an interesting framework for further analysis. Such a limited validation is usually an easy first step for any framework, however. The four cases examined in sections 5 and 6 were selected by Thagard to play the role of much more substantial examples. How convincing are they? I will concentrate on the first one, representing Lavoisier's 1783 critique of phlogiston. I will argue that this example says little about how one of the competing theories is superseded by the other. Then I will discuss the possibility of better tests.

ECHO *applied to Lavoisier's arguments against phlogiston.* Consider the oxygen/phlogiston example. Does ECHO answer why the oxygen theory of combustion superseded the phlogiston theory? I do not think so. Running on data that Thagard reconstructed from Lavoisier's 1783 paper (Lavoisier 1862), ECHO concludes that the phlogiston theory is less coherent than the oxygen theory. This supports Thagard's descriptive claim about ECHO by confirming Lavoisier's conclusion based on Lavoisier's data. However, in order to understand the shift from the phlogiston theory to the oxygen theory, it is not as important to understand why Lavoisier became convinced that the phlogiston theory is inferior as it is to understand why the phlogistians gave up. To understand this, we must take the strongest, not the weakest, accounts of their theory. In the example considered by Thagard, Lavoisier criticizes some of the old phlogistic claims made prior to the discovery of oxygen. In the early 1780s, phlogiston theory, improved by Cavendish, Kirwan, and Priestley, could explain evidence E3, E4, E5, E6, and E7 in Thagard's Table 1 (Musgrave 1976, pp. 193–94; Partington 1962, p. 255; Zytkow & Lewenstam 1982, pp. 45–46), contrary to Lavoisier's claim.

According to the improved phlogiston theory, during calcination the phlogiston disengages from metal, forming a compound with dephlogisticated air (oxygen). That compound in turn combines with the calx that remains from metal. This schema, extended to other substances, was able to explain the decom-

position of calx of mercury into mercury and oxygen, and solved the nagging anomaly of the increase in weight during calcination.

In selecting Lavoisier's 1783 paper, Thagard considers a phlogiston theory no longer entertained by the leading phlogistians–hence it is not strange that they remained unconvinced by Lavoisier's criticism. So even if ECHO demonstrates the superiority of Lavoisier's theory, the relevance of the example does not go beyond a description of the particular reasoning of a particular scientist against a dead or imaginary opponent. It might have had an impact only on some outsiders in chemistry, as it has had on some historians.

The input given to ECHO, listed in Tables 1 and 2, raises further doubts. Why is OH4 – "Oxygen has weight" – treated as a hypothesis in Table 1, not as evidence? Examining Table 2, I do not see why some hypotheses are relevant, nor why some evidence is explained. For instance, why is E1 explained by hypotheses OH1, OH2, and OH3? Neither hypothesis tells anything about any substance being given off, so how can we conclude that heat and light are given off? I do not see why OH1 is relevant to the explanation of E3: Thagard treats his explanatory relation as an undefined primitive. Because similar doubts apply to most entries in Table 2 and to the input for the remaining three examples, they suggest a more explicit treatment of explanation. Thagard himself mentions the input preparation problem in section 7. But his problem has been solved to a large extent elsewhere. The generation of hypotheses and explanatory links can be automated by a combination of two existing computer discovery systems, STAHL and GLAUBER, developed several years ago (Langley et al. 1983; 1987, Chapters 6 and 7; Rose & Langley 1986; Zytkow & Simon 1986). Jointly, these systems can construct both the hypotheses in Table 1 and explanations similar to those in Table 2 by using the evidence in Table 1 as well as some additional observations. STAHL and GLAUBER have been tested on many historical episodes. Their simple and explicit operators allow for detailed examination of the explanatory process.

***Better tests for* ECHO.** Can ECHO describe the shift from the phlogistic view to Lavoisier's view? To test this problem we need cases of phlogistians, who, knowing both theories, decide in favor of Lavoisier. To pass the test, ECHO should be able not only to mimic this performance by selecting a stable state corresponding to the oxygen theory, but also demonstrate that a particular, historically valid input caused the shift from the previously held stable state corresponding to the phlogistic theory. Unfortunately, little is known about these episodes. If they were available, however, I would not expect them to confirm Thagard's conjecture about the descriptive capability of ECHO. Leading phlogistians conducted very incisive analyses of Lavoisier's claims and produced excellent accounts of his theory, but they did not feel that their theories were less coherent (Cavendish 1785; Kirwan 1789). In my view, a plausible explanation of the paradigm shift in chemistry at the end of eighteenth century cannot be explained by abstracting from the contents of both theories.

From the descriptive point of view, validating ECHO requires many historical episodes as data points. One can find many other candidates for test cases which are perhaps not as spectacular as the transition from the phlogiston theory to the oxygen theory, but which are much better documented. Leading theories of eighteenth-century chemistry underwent many changes following the discovery of hydrogen, oxygen, decomposition of water, and so forth. Because each discovery led to different responses by different chemists as described in their writings, many test cases are readily available.

Conclusions. ECHO provides an interesting, uniform, domain-independent mechanism for coherence testing. It is poorly supported, however, and the impact of this work is unclear. In each domain of application, ECHO should be tested on a number of carefully selected cases. In the domain of eighteenth-century chemistry, ECHO may be coupled with STAHL and GLAUBER, because they are able to generate most of ECHO's input.

NOTE

*On leave from Wichita State University and the University of Warsaw.

Author's Response

Extending explanatory coherence

Paul Thagard

Cognitive Science Laboratory, Princeton University, Princeton, NJ 08542
Electronic mail: *pault@confidence.princeton.edu*

The commentators have raised numerous important questions about my account of explanatory coherence. In reply, I will first address the most general issues about the kind of approach to understanding inference I have taken, answering queries about the philosophical, psychological, and computational nature of this project and about the role that connectionist ideas play. I will then address theoretical questions about explanation, simplicity, analogy, probability, and conceptual change, and will subsequently look at problems concerning the ECHO model. (Adopting **Reggia**'s acronym, I distinguish between TEC, the theory of explanatory coherence expressed in the seven principles in section 2.2 of the target article, and ECHO, the computational implementation of those principles.) Finally, I will discuss problems pertaining to the adequacy of TEC and ECHO for characterizing human thinking.

1. The general approach

1.1. Philosophy, psychology, and artificial intelligence. TEC and ECHO are intended simultaneously to contribute to philosophy of science, cognitive psychology, and artificial intelligence (AI). **Dietrich** seems puzzled about whether TEC is a theory in the philosophy of science or a psychological theory. My intention is for it to be both, and I acknowledge the possibility that it could fail to be adequate in both respects. That he sees a tension between these two interpretations is not surprising given the logical positivist tradition in the philosophy of science that tried to separate logic from psychology. But postpositivist philosophy of science should be psychologistic, not in the strong sense that supposes that however scientists think is rational, but in the weak sense that judgments of rationality take actual thought processes as their starting points. The investigation of those processes then becomes part of the philosophy of science. The best current method for psychological theorizing comes from computational modeling. From this perspective, philosophy becomes part of cognitive science; it should not seem odd to find a computer program described as part of a theory in the philosophy of science. The point of ECHO is to show that a much more detailed and applicable account can be provided of explanatory coherence and theory evaluation

Table 1. *The format of this response*

1. The general approach
   1.1. Philosophy, psychology, and artificial intelligence
   Dietrich, Wetherick
   1.2. Connectionism
   Dietrich, Lycan, Cheng & Keane, Wetherick, Giere, Levine
2. Theoretical issues
   2.1. Explanation and hypothesis evaluation
   Achinstein, O'Rorke, Sintonen, Josephson
   2.2. Simplicity
   Reggia
   2.3. Analogy
   McCauley, Gabrys & Lesgold, Hobbs
   2.4. Conceptual change
   Giere, Mangan & Palmer
   2.5. Logic and probability
   Feldman, Cohen, Papineau, Lycan, Dawes, Bereiter & Scardamalia
3. Problems with the ECHO model
   McCauley, Mangan & Palmer, Dietrich, Zytkow, McDermott, Hobbs, Bereiter & Scardamalia, Simon, Zytkow
4. Psychological adequacy
   Klayman & Hogarth, Earle, Cheng & Keane, Chi, Bereiter & Scardamalia, Read & Miller

Note: The commentaries discussed in each category are listed in order of appearance.

than philosophers have given so far. Computational philosophy of science (Thagard 1988a) fits within the center of the interdisciplinary field of cognitive science, attempting an integrated assault on problems common to philosophy, psychology, and AI. Dietrich's suggestion that ECHO is in the logical positivist tradition ignores the fact that TEC and ECHO are neither logical (in the narrow sense) nor positivist. They are not positivist because the emphasis is on high-level theories, not on observation, and data can be rejected; and the principles of explanatory coherence go well beyond formal logic.

Perhaps it would be useful to coin a new term to describe an approach that is intended to be both descriptive and prescriptive. I shall say that a model is *biscriptive* if it describes how people make inferences in accord with the best practices compatible with their cognitive capacities. Unlike a purely prescriptive approach, a biscriptive approach does not offer a theory of God's cognitive performance, but is intimately related to actual human performance. Unlike a purely descriptive approach, biscriptive models can be used to criticize and improve human performance.

Whereas my project is intended to be philosophical, psychological, and computational, **Wetherick** sees TEC and ECHO as sociological on the grounds that my examples involve acceptance or rejection by a community of scientists. I was explicitly modeling Lavoisier and Darwin, however, not communities of chemists or biologists. My examples come from the history of science, not its sociology. Nor do I pretend to model minds free of prejudice and preconceptions. As described in section 10.4 of the target article, ECHO does display a degree of conservatism in how it deals with new evidence, and other sorts of preconceptions could be modeled using the mechanisms for analogy. Wetherick simply asserts that explanatory coherence is always determined by conscious symbolic processing, but people often appreciate the coherence of a new point of view only after they have stopped consciously arguing about it.

1.2. Connectionism. The computational side of my account of explanatory coherence draws heavily on connectionist ideas. Yet **Dietrich** and others see ECHO as almost peripheral to the theory, TEC. There are two responses to this, one biographical and the other methodological. Although the organization of the target article suggests that TEC came first and ECHO followed, I in fact got the idea for ECHO by analogy with the ACME program for analogical mapping that Keith Holyoak and I were developing (Holyoak & Thagard, in press). Thinking in terms of connectionist algorithms for simultaneously satisfying multiple constraints had enabled us to reconceptualize the problem of how the components of two analogs can be put in correspondence with each other, and it struck me that a similar approach might work for the problem of hypothesis evaluation. General ideas about inference to the best explanation and parallel constraint satisfaction led to ECHO, which led to TEC, and ECHO and TEC thereafter evolved together. As usual in cognitive science, there was considerable interplay of theory and model, with ideas about how to improve ECHO suggesting improvements in TEC and vice versa. The connectionist model thus played a crucial role in theory development, but it has also been instrumental in evaluating the theory. A typical theory in the philosophy of science is defended with a brief discussion of a couple of examples. ECHO makes possible and necessary the development of very detailed simulations that simultaneously lend credence to claims about the scope of ECHO and the scope of TEC. I therefore see connectionist ideas about parallel constraint satisfaction as integral to both the generation and the evaluation of a theory of explanatory coherence.

Lycan claims that my emphasis on connectionism is misleading, for he seems to consider distributed representations and neurological aspirations central to connectionism. Yet the more careful connectionists have made it clear that the similarity between the brain and current connectionist models, including distributed ones, is superficial at best. Rather than viewing connectionism as a new "paradigm" that obviates traditional AI, I see connectionist ideas as a very useful supplement to traditional ideas in AI. A convincing argument for the redundancy of the connectionist approach would require the development and general application of a nonconnectionist version of ECHO. I have produced the rule-based version of ECHO described in section 7 of the target article, and it has the conjectured limitations. Although it duplicates ECHO's performance in quite a few cases, there are numerous other cases where it lacks ECHO's subtlety and generates different, less appropriate, conclusions.

I agree with **Cheng & Keane** about the importance of developing a psychological account of explanation, but I want to challenge their dichotomy between "conventional symbolic" models and connectionist ones. **Wetherick** also erroneously contrasts my account of explanatory coherence with symbolic approaches. Both proponents

and critics of connectionism have exaggerated the difference between connectionist and traditional approaches. For one thing, even connectionist models with distributed representations have a substantial symbolic component involved in their inputs and outputs. Both in this work on explanatory coherence and in research on analogy (Holyoak & Thagard, in press), I favor a hybrid approach in which symbolic reasoning is used to create a network and connectionist algorithms are used to do parallel constraint satisfaction. There is nothing "subsymbolic" about this approach at all. I also like the way that distributed representations work, but doubt that the symbolic/subsymbolic distinction is useful there either. What is needed in cognitive science today is work that integrates many different approaches.

Whereas several commentators have upbraided me for being too connectionist, other readers will undoubtedly think that I have not been connectionist enough, relying too heavily on symbolic input and not using distributed representations. I have no doubt that the understanding of cognitive processes will require the nonlinguistic representational mechanisms and judgmental strategies that **Giere** advocates. I conjecture, however, that explanation and theory evaluation are heavily influenced by our ability to use language, so that linguistic representations of varying degrees of complexity will be central to understanding the highly verbal practices of scientists.

Levine is of course right that nothing in the ECHO model addresses the question of how to modify a theory or to synthesize parts of two conflicting theories. These are important issues for future research. I also very much like his challenge to connectionist and neural theories to indicate how higher-order processes of explanation and coherence might emerge from lower-order processes. Some extreme proponents of the neurophysiological approach might be tempted to argue that the coherence relations I discuss are mere epiphenomena and will prove superfluous once neuroscience really gets rolling. I think such proponents are roughly in the position of a fifteenth-century scientist trying to practice molecular biology before much was known about whole organisms. My recommended strategy for cognitive science is to have many people working simultaneously at many levels, with researchers at each level keeping appreciative eyes open for relevant work at the other levels. Section 8.1 of the target article listed eight different approaches, all of which I think are worth pursuing. My conjecture is that one of the major sources of scientific progress in the near future will come from the *interpenetration* of approaches – for example between neuroscience and experimental cognitive psychology, and between connectionist models and traditional AI models. Fortunately, this seems to be exactly what is happening, despite the jeremiads of some researchers who insist that only their favorite approach is worthwhile.

2. Theoretical issues

2.1. Explanation and hypothesis evaluation. Now let us turn to issues central to TEC, the theory of explanatory coherence. **Achinstein** contends that I need to show some intrinsic connection between explanation and acceptability. He suggests that I would not want to say that the *Book of Genesis* explains the origin of the universe, in a sense of "explains" that has any connection with acceptability. On the contrary, I see no difficulty in saying that *Genesis* explains the origin of the universe and would even allow that there was a time when it (and equivalent theological views) provided the best available explanation of the universe. The reason that the *Genesis* account is no longer acceptable is that we have accumulated masses of data about the development of the universe that are explained much more comprehensively and simply by new theories such as the Big Bang theory. We would need a connection between explanation and acceptability only if we were arguing directly from "H explains the evidence" to "H is acceptable," but no one advocates that. Instead, we ought to make sure that before H is accepted we have sought alternative explanatory hypotheses and done a comparative explanation. The inference from "H is the best explanation of the evidence" to "H is acceptable" does not require any special relation between explanation and acceptability. The use of inference to the best explanation can be justified on the basis of my methodology for going from the descriptive to the normative (Thagard 1988a, Chap. 8).

I have regrettably not offered a theory of explanation, but I do see glimmers of what such a theory might look like. It would not resemble the attempts by philosophers to give an *analysis* of the concept of explanation, that is, to give a set of necessary and sufficient conditions for something being an explanation. Concepts in science and ordinary life are rarely susceptible to such definitions. At best, we can hope to describe features characteristic of typical explanations. Although not all explanations are deductive, many are at least quasideductive in that they place what is to be explained in the context of general laws, doing at least a rough derivation. Many are causal, in that they invoke causal mechanisms as part of the derivation of what is to be explained. Many are oriented toward answering particular questions that have been posed. Many involve fitting schemas to a situation to generate a contextual understanding of events. A theory of explanation should integrate these quasideductive, causal, question-oriented, and schematic components to account for the explanatory practices of scientists. The theory would be computational in that it would be precise enough to be implemented in a computer program, but it would be clearly distinguished from the program that implements it. Finally, the theory of explanation would make contact with experimental studies of how people generate and use explanations. A cognitive theory of explanation would be a substantial contribution to philosophy, psychology, and AI.

O'Rorke accurately characterizes aspects of the scientific reasoning process that TEC and ECHO do not address. He suggests that it is necessary to introduce evaluation into the construction process in order to reduce the number of explanations generated. Implicitly, this is true of the abduction mechanism in the system PI ("Process of Induction," Thagard 1988a, Chap. 4), because a hypothesis has to explain at least one fact for it to be generated. It will be interesting to see what other constraints are embodied in O'Rorke's programs and to try to integrate explanatory coherence considerations with programs that make decisions about when to collect more information. O'Rorke also suggests that an agent's goals and priorities

play important roles in evaluation. Now I can certainly see the relevance of goals and priorities for the *generation* of hypotheses. Holland et al. (1986) emphasized that induction should be constrained by problem-solving contexts. But the question of whether to accept a hypothesis is separable from the question of where to focus attention and the decisions that may be based on the hypothesis once it is accepted. Eventually I plan to develop a modified version of ECHO, MOTIV-ECHO, that is capable of conflating hypothesis evaluation and decision making, which is also naturally understood to be a parallel constraint-satisfaction process. Motivated ECHO will reach conclusions on the basis of how well beliefs satisfy its goals, as well as on the basis of how much explanatory coherence they have (cf. Kunda 1987; Thagard & Kunda 1987). But Kunda's data, although they support the view that people make motivated inferences, also suggest that the way this works is much more subtle than merely believing what one wants to believe. Motivation affects memory search for evidence that is then selectively applied in support of desired conclusions. People do not just believe what they want, although they attempt to find evidence for what they want to believe.

Sintonen raises the important question of whether ECHO can deal with promise, suggesting that scientists adopt a theory in part because they think it has the potential for growth. Undoubtedly this occurs. I am sure that the reason so many graduate students are working on connectionist models, in some cases to the chagrin of their supervisors, is partly that there are many open questions to be investigated; in contrast, many techniques that have been very important for AI, such as rule-based systems, have already been well explored. I think, however, that we can distinguish the decision to work on a project from the judgment, based on explanatory coherence, that the theory underlying the project is the best one available. The category of "promise" blurs into wishful thinking, opening up the possibility that scientists believe a theory because it has the potential to make them successful rather than because of its coherence. MOTIV-ECHO might model such inferences. Similarly, Sintonen suggests that jurors take into account the consequences of their decisions for the punishment of the accused, not just the explanatory coherence of the competing accounts. For scientists, however, I think it is more common to reverse this inference and think that a theory will make them successful because of its explanatory coherence. Sintonen is not suggesting that ECHO be broadened to a fully motivated ECHO, only that a component of "expected coherence" be added. If there were some reasonable way to assess the expectation, this might be appropriate, but I do not see any natural way to add this constraint.

Sintonen has a legitimate concern about the identity criteria of theories and suggests that I distinguish more carefully between core hypotheses and auxiliary hypotheses. I would prefer to have that distinction emerge from the model: Core hypotheses have many explanatory connections to evidence and other hypotheses, whereas the auxiliary hypotheses are very sparsely connected. I grant Sintonen's claim that theories have some autonomy in selecting the territories they claim, but I believe that once two theories have territories that overlap, each of them should pay attention to the territory of the other. That does not mean that the best explanation has to explain everything the other theory does, only that it normally tries. The system PI incorporates an algorithm for accumulating alternative hypotheses and relevant evidence (Thagard 1988a, p. 207).

Josephson challenges TEC's assumption, derived from Harman (1973), that a hypothesis becomes more acceptable if it is explained by a higher-level hypothesis. Yet in every domain in which explanatory inference is used, higher-level explanations can add to the acceptability of a hypothesis. Darwin, for example, thought that the fact that evolution was explained by natural selection was a crucial part of the evidence for evolution. In murder trials, questions of motive play a major role, because we naturally want an explanation of why the suspect committed the crime as well as evidence that is explained by the hypothesis that the suspect did it. In Josephson's own favorite domain, medical diagnosis, I expect that doctors normally find the hypothesis that cirrhosis of the liver is the cause of a patient's symptoms more convincing if they can also explain why the patient got cirrhosis from being a heavy drinker. ECHO shows that incorporating this element of explanatory coherence into a computational model does not create any intractable problems. Unlike Harman, however, I do not subsume enumerative induction under inference to the best explanation, but treat it as an independent form of inference (Holland et al. 1986, Chap. 8; Thagard 1988a). Contrary to Josephson's suggestions, nothing in TEC involves a position on the philosophical question of the symmetry of explanation and prediction.

2.2. Simplicity. The important parsimonious covering theory suggested by **Reggia** provides a serious alternative to TEC, although TEC's notion of simplicity is not as simple as Reggia suggests. He interprets TEC's Principle 2(c) (section 2.2) as a principle of minimal cardinality: Explanatory hypotheses with the smallest number of hypothesized components are preferable. But 2(c) does not have this consequence. Consider a theory, T1, consisting of three hypotheses, H1, H2, and H3. The alternative theory, T2, consists of H4, H5, H6, and H7, where H1 and H4 are contradictory. Now suppose that there are two pieces of evidence, E1 and E2, and that H1 and H2 explain E1, and H1 and H3 explain E2. On the competing side, suppose that H4 explains E1, and H4, H5, and H6 explain E2, and moreover that H7 explains H4, H5, and H6. When ECHO is run on this example, the hypotheses in T2 are all accepted and H1 is rejected, even though T2 has more hypotheses than T1. An unlimited number of similar examples would show that what matters is not just the sheer number of hypotheses but also their configuration. That being said, I have much sympathy for Reggia's general approach, and would be ready to use Bayesian methods when frequencies and prior probabilities are available. In the examples to which ECHO has been applied, they generally are not. The closest one could come is perhaps to use analogies to indicate prior plausibilities, favoring hypotheses that figure in explanations that are similar to ones already used. In the medical and engineering domains where Reggia's theory has been successful, probabilities based on frequencies are more obtainable and sensible than in the wide-open scientific and legal domains to which ECHO has been applied.

2.3. Analogy. Principle 3 of TEC is challenged by **McCauley** on the grounds that theories cannot be evaluated on the basis of analogies, because what counts as an analogy is partly determined by theory. Although granting that theories influence analogies, I think that the process of analogy recognition is sufficiently independent of theorizing to leave Principle 3 intact. Our theory of analogical mapping (Holyoak & Thagard, in press) shows how structural, semantic, and pragmatic constraints affect how the components of two analogs can be placed in correspondence with each other. Defenders of different theories may well have different goals in the use of an analogy, and this will affect their use of the analogy; our theory accommodates this in its admission of a pragmatic constraint. But that is only one constraint among many, and if the structural (syntactic) and semantic (meaning-related) aspects of the two analogs are similar, as I think they usually are, then proponents of different views can reach some agreement about the nature of the analogy. McCauley is of course right that the analogy between artificial and natural selection did not in itself move the nineteenth-century creationist, but it was only one component of the whole picture. Darwin had to argue that species were more like breeds than creationists had allowed, and this was very controversial. But creationists could nevertheless appreciate the structure of the analogy that said that having nature select and produce species was something like having breeders select and produce breeds. Similarly, although I have challenged the general usefulness of the analogy between biological evolution and the growth of scientific knowledge (Thagard 1988a, Chap. 6), I have no difficulty in seeing the relations that constitute the analogy.

Gabrys & Lesgold rightly point out that jury reasoning, such as in the cases modeled by ECHO, is very different from jurisprudential reasoning, in which judges and lawyers apply the law. I am in complete agreement that case-based (analogical) reasoning is important in law, but I do not understand why these commentators view this as being incompatible with constraint-satisfaction models. If our account is correct (Holyoak & Thagard, in press), then analogical reasoning, which subsumes case-based reasoning restricted to a single domain, is very much a matter of simultaneous satisfaction of structural, semantic, and pragmatic constraints. Just as ECHO uses connectionist algorithms to integrate various considerations for evaluating hypotheses, the analogy program ACME uses such algorithms to integrate constraints about how analogs can be put into correspondence with each other. ACME is of course different from ECHO in the way it constructs constraint networks, but it is similar in the way it uses relaxation techniques to calculate how to satisfy constraints.

Hobbs argues that the existence of an analogy enhances a theory's explanatory power only when it rests on a deeper abstract principle. Similarly, some AI researchers have claimed that analogies always involve some kind of abstraction. In the view of Holyoak and Thagard (in press), however, analogies can be recognized independent of such abstract principles; in fact, it is often the recognition of the analogy that prompts the formation of the abstract principle. Once the relevance of analogies A1 and A2 has been noticed, it becomes possible to abstract from them a schema that incorporates the relevant features of both (Holland et al. 1986, Chap. 10). In the Darwin case, for example, I suspect that the abstraction "Selection can result in new varieties of living beings" came about only after Darwin's theory and the analogy he had cited were accepted. So analogy remains independent of simplicity and consilience (explanatory breadth).

2.4. Conceptual change. TEC is intended to play a role in accounting for conceptual change in science. But **Giere** contends that I have given a model of scientists' arguments, not of their reasoning. To be sure, my historical cases are based on published arguments, not on protocols collected from scientists in the heat of reasoning. In contrast, Ranney and Thagard (1988) report ECHO analyses based on subjects' verbal reports of the stages of their reasoning. Whether similar techniques could be used with practicing scientists to generate evidence concerning scientists' use of ECHO-style considerations is an open empirical question. I hope the experiments are done with practicing scientists. I like the fact that what might appear to be a philosophical disagreement between Giere and me about the applicability of TEC is amenable to empirical investigation.

Giere questions whether TEC can explain the transition from the phlogiston to the oxygen theory, because we would expect proponents of the earlier theory simply to mount their own explanations. This is undoubtedly what happens initially, but in the course of argument scientists start to understand the evidence and explanatory relations of their opponents, so that they acquire an enhanced picture of the explanatory relations. Over time (it typically took a couple of years in the chemical revolution), the expanded explanatory network can lead to the adoption of the new theory. I therefore do not want to retreat to the purely normative stance that Giere suggests. I have no objection to Giere's proposal that we look at the full range of representational mechanisms and judgmental strategies that operate in individual cognitive agents, but I contend that psychological experiments may show explanatory coherence considerations to be paramount. TEC is thus intended to be much more psychological and no less normative than inductive logics.

Mangan & Palmer find the philosophy of science implicit in TEC insufficiently up to date, assailing its Kantian flavor in contrast to the more relativist views of Kuhn and Feyerabend. But the principles of TEC are not claimed to be synthetic a priori like Kant's fundamental principles. Elsewhere (Thagard 1988a, Chap. 7) I offer an account of how to develop normative principles from descriptive considerations; I would want that account to apply to TEC as well. The main support for TEC comes from its application to numerous cases in the history of science, not from some a priori deduction. Methodological theories have to cohere with inferential practice, although they can do this in part by invoking psychological and sociological background knowledge to explain deviations from the principles. Kuhn and Feyerabend overestimate, I would argue, the degree of variability of methodological principles in the history of science.

The real issue between TEC and the Kuhn/Feyerabend view of scientific change concerns the degree to which principles of explanatory coherence change as part of "paradigm shifts." **Mangan & Palmer**'s uncritical acceptance of Kuhnian dogmas presupposes that Kuhn got the

history of science right, successfully using it to refute his positivist predecessors. But in many respects, Kuhn's description of the nature and magnitude of scientific change is not historically accurate (Donovan et al. 1988; Thagard, in press b). TEC would indeed be historically inadequate if it turned out that the shift from one major theory to another introduced new principles of explanatory coherence and rejected old ones, but the extent to which this has occurred has been exaggerated. To take one example, Mangan & Palmer attribute to me the view that the use of analogy "became" an important explanatory device for Darwin. But Darwin certainly did not originate it. In fact, we find analogy prominent in the writings (much admired by Darwin) of William Paley, one of the leading scientific creationists. Analogy also figured in arguments used for the wave theory of light by Huygens and Fresnel. So it was not the case that one aspect of the Darwinian revolution was the introduction of a new principle of explanatory coherence.

Although I reject for philosophical and historical reasons the drift of **Mangan & Palmer**'s approach, I am intrigued by the architecture they propose for using gating units to modify the impact of explanations and analogies. I hope they will explore the kind of structure they describe for adjusting the impact of explanation, simplicity, and analogy, possibly learning it from feedback. I would argue, however, that understanding the major kinds of conceptual change that take place in scientific revolutions should pay more attention to questions of conceptual structure (Thagard, in press b; Thagard & Nowak, in press) than to questions of change in principles of explanatory coherence.

2.5. Logic and probability. Several commentators compare ECHO unfavorably with probability theory. **Feldman** expresses regret that more attention was not paid to *foundational* questions. He contrasts my theory of explanatory coherence with logic and probability theory, each of which is said to have a relatively clean and well-understood formal semantics. He wants a similar interpretation for the weights and activity levels in ECHO. My guess is that such a semantic foundation for explanatory coherence is going to be very difficult to find. In fact, logic and probability theory do not have much of a foundation either. The appearance that they have a clearly understood semantics dissipates when one looks closely at basic cases. Consider logic in its best understood form, first-order predicate calculus, whose Tarskian formal semantics consists of giving a recursive truth definition for progressively more complex formulas. The simplest formulas are atomic propositions such as "Fa," which is semantically interpreted as saying that the object that provides the interpretation of the constant "a" is in the set of objects that provides the interpretation of the predicate "F." But this account begs a host of foundational questions, particularly what makes a set the interpretation of "F." Surely it is the meaning of "F" that determines what objects fall under it, so that the Tarskian interpretation dodges the central semantic question. The situation gets even worse when one moves beyond first-order logic to consider modal notions that are crucial for understanding scientific discourse. Matters of causality and explanation require conditionals (if-then statements) that go well beyond those found in standard logics, for example, to permit semantic evaluation of counterfactual conditionals such as "If Bush had not been elected President, then the economy would be stronger." Formal semantics for such conditionals generally use the notion of possible worlds, which are far from being "clean and well understood."

Similarly, how can one say that probability theory has a clean, formal semantics when philosophical debates about the interpretation of probability still rage? Cohen (1977), for example, rejects some of the standard axioms of probability theory [see Cohen "Can Human Irrationality be Experimentally Demonstrated?" *BBS* 4(3) 1981.], and even philosophers who accept the axioms debate whether probabilities should be interpreted as frequencies, propensities, or subjective degrees of belief. The apparent superiority of logic and probability theory derives more from their familiar syntax than from any foundational advantage.

Cohen raises two substantial challenges to my account of explanatory coherence. The first is based on the intuition that a hypothesis that is used to predict a piece of evidence gains more confirmation from it than does a hypothesis that explains it after the fact. I have argued previously that the apparent importance of prediction is really a matter of simplicity, in the sense used in TEC:

> I contend that the major reason why prediction of new phenomena appears so important is that such predictions are likely to be a sign of simple explanations. In making a prediction, one does not have the opportunity to adjust the theory to an already-known outcome by means of auxiliary hypotheses. Using only the theory and already familiar auxiliary assumptions, a future outcome is predicted with no opportunity for adjustments that are local to the prediction. In contrast, explanation after the fact can make many special assumptions to derive the outcome from the theory . . . Successful predictions are to be valued as signs of the simplicity of a theory, showing that its explanations do not require post hoc additions. (Thagard 1988a, pp. 84–85)

If this account is right, then TEC does not need any ad hoc additions to account for people's preference for hypotheses that make predictions. The account could be challenged, however, by providing evidence from the history of science or controlled psychological experiments that display people preferring hypotheses that make predictions over ones that provide post hoc explanations with the same number of auxiliary assumptions.

Cohen's second point is that neither TEC nor ECHO provides a way to determine the acceptability of a conjunction based on the acceptability of the conjuncts. This would be a grave problem if TEC were intended to be a general theory of inference, but it is not. As I stated in the target article (section 10.4), I view inference to explanatory hypotheses as only one of a battery of inferential mechanisms. We still lack a general theory of how to combine explanatory inferences with deduction, generalization, specialization, analogy, and statistical reasoning. Both Bayesian and Baconian analyses seem to me to lack the requisite generality. I therefore view the problem of the acceptability of conjunctions as unsolved and beyond the scope of TEC. As I pointed out in the target article (section 10.2), however, the problem in real cases is not solved by probability theory either, because cal-

culating the probability of a conjunction requires knowing the degree of dependence of the conjuncts, which is often indeterminate.

Papineau challenges my arguments against probability theory largely on the grounds that people can learn to reason better probabilistically. I agree with his general point and would encourage every effort to improve probabilistic reasoning. Psychologists, who originally reached very pessimistic conclusions about people's ability to reason statistically, have more recently been investigating how to *teach* the use of inferential rules (Holland et al. 1986, Chap. 9). My view, from what **Lycan** correctly labels a "weak explanationist" perspective, is that we do not make all our inferences on the grounds of explanatory coherence but that we should exploit probablistic knowledge whenever it is available. Nothing that Papineau says, however, overcomes my basic point, which is that such knowledge is rarely available in the qualitative evaluation of scientific theories. My recommendation is to use statistical inference and probabilistic reasoning whenever possible, but not to pretend it is always possible. There is no reason to accept Papineau's prohibition of computational models of processes that can be transformed by normative reflection. Understanding through use of such models how these processes might work can be an important part of bringing about the transformation. Contrary to the suggestion of **Dawes,** I do not follow **Cohen** in supposing that what people actually do is normatively correct; the real is not always rational. I like **Bereiter & Scardamalia**'s suggestions about teaching people to make better judgments of explanatory coherence; I doubt that probability theory will be of much help there, however useful it is for other sorts of problems.

The major challenge laid down in **Dawes**'s commentary is to show that ECHO can deal with Simpson's paradox. This paradox arises when a hypothesis, H1, gains acceptability from one piece of evidence, E1, and also from another pieces of evidence, E2, but becomes less acceptable given the conjunction of E1 and E2. TEC and ECHO can handle this naturally if E1 and E2 together support some alternative hypothesis that affects the acceptability of H1. Here is an example that I think is clearer than the ones Dawes presents: Suppose that Mike is charged with committing a murder in New York. The acceptability of the hypothesis that Mike is innocent is increased by the piece of evidence that his friend Sam says Mike was in Philadelphia at the time of the murder. Taken alone, the acceptability of the innocence view would also be enhanced by the evidence that another friend, Fred, says that Mike was in Boston at the crucial time; but we find Mike's innocence less plausible if Sam and Fred both furnish incompatible alibis. All this is naturally understood in terms of explanatory coherence, as shown by the input for ECHO listed in Tables 2 and 3. When ECHO is provided simply with Sam's testimony, which is explained by the hypothesis that Mike really was in Philadelphia, then the hypothesis that Mike is guilty is defeated. But in the more complicated case where there are contradictory alibis, the hypothesis that Mike is guilty is accepted, so the claim that he is innocent is rejected. I conjecture that other cases of Simpson's paradox can similarly be dealt with by attending to the full complexity of the networks of competing explanatory hypotheses involved in the cases. In Dawes's case of the drunk in the bar, we can explain his being drunk before the crime as an attempt to get his courage up to commit the murder, and we can explain his being drunk after the crime as an attempt to overcome the stress of committing the crime, but the best explanation of his being in the bar both before and after the crime is that he spent the whole time in the bar drinking.

Table 2. *Input to* ECHO *for Mike's simple alibi*

```
(proposition 'G0  "Mike committed the murder in New York.")
(proposition 'G1  "Sam is lying to protect Mike.")
(proposition 'I1  "Mike was in Philadelphia.")

(proposition 'E1  "Sam says that Mike was in Philadelphia.")

(explain '(G0) 'G1)
(explain '(G1) 'E1)
(explain '(I1) 'E1)

(contradict 'G0 'I1)
(contradict 'G1 'I1)

(data '(E1))
```

3. Problems with the ECHO model

An important question about the arbitrariness of the inputs to ECHO is raised by **McCauley.** He is skeptical about how much agreement might be found between disputants about what constitutes an explanation and what constitutes an analogy. Skepticism would certainly be warranted if the Kuhnian view, defended by **Mangan & Palmer,** were correct. We would then expect what counts as an explanation or analogy to vary considerably from scientist to scientist; but I think this possibility is

Table 3. *Input to* ECHO *for Mike's contradictory alibis*

```
(proposition 'G0  "Mike committed the murder in New York.")
(proposition 'G1  "Sam is lying to protect Mike.")
(proposition 'G2  "Fred is lying to protect Mike.")
(proposition 'I1  "Mike was in Philadelphia.")
(proposition 'I2  "Mike was in Boston.")

(proposition 'E1  "Sam says that Mike was in Philadelphia.")
(proposition 'E2  "Fred says that Mike was in Boston.")

(explain '(G0) 'G1)
(explain '(G0) 'G2)
(explain '(G1) 'E1)
(explain '(G2) 'E2)
(explain '(I1) 'E1)
(explain '(I2) 'E2)

(contradict 'I1 'I2)
(contradict 'G0 'I1)
(contradict 'G0 'I2)
(contradict 'G1 'I1)
(contradict 'G2 'I2)

(data '(E1 E2 ))
```

exaggerated. I am surprised that **Dietrich** had difficulty distinguishing between hypotheses and evidence in the case of Poincare's explanation of the Eureka Phenomenon. My collaborators and I have found the distinction unproblematic. A proposition is evidence if it describes the result of observation or experimentation. Hypotheses, in contrast, explain such results or other hypotheses. **Zytkow** asks why OH4, "Oxygen has weight," is treated as a hypothesis, not as evidence. This was obviously a hypothesis for Lavoisier, because oxygen as such is not observable. The explanatory connection in Table 1 of the target article between the hypotheses OH1, OH2, and OH3 and the evidence E1 is there because of the background assumption that it was the heat and light *from the oxygen* that was produced by the reaction. The explanation here is not at the level of a deductive derivation, but at the level of the discourse at which scientists normally operate.

McDermott sees as central to my model of explanatory coherence the minimizing of the energy function H defined by equation (1) in section 4.9. Perhaps he was misled by my assertion that ECHO stands for "Explanatory Coherence by Harmany Optimization," which was only an attempt to combine a catchy name with a pun. The H function strikes me as peripheral to the whole model, whose main function is to show how explanatory hypotheses can be evaluated in complex ways. The reason for taking the connectionist route is simply that networks provide a very useful way of simultaneously representing a host of evidential relations, and the numerical relaxation algorithms standardly used in connectionist models are a very natural way to accomplish parallel satisfaction of the numerous constraints implicit in the networks that are created.

Hobbs mounts a substantial challenge: Why bother with all the apparatus of ECHO when it might appear that a "naive method" that simply counts propositions does just as well? The naive method evaluates a theory by subtracting the number of hypotheses it uses from the number of pieces of evidence it explains; in Hobb's notation, this is #E − #H. There are, however, an unlimited number of cases in which ECHO yields a conclusion different from the naive method. To take one of the simplest, consider a theory, T1, consisting of hypotheses H1 and H2, which are both used together to explain evidence E1 and E2; that is, H1 and H2 together explain E1, and together explain E2. The alternative explanations are H3 and H4, but H3 explains E1 alone and H4 explains E2 alone. Suppose that H1 contradicts H3 and H2 contradicts H4. T1 is more unified than the other singleton hypotheses, and ECHO indeed prefers them, despite the fact that the naive method calculates #E − #H as 0 in both cases. So even independent of questions of being explained and analogy, the naive method is not equivalent to ECHO. The divergence is even clearer in the examples discussed in section 4.3 of the target article and in relation to **Reggia** above. I have already responded to attempts by Hobbs and Josephson to downplay the significance of analogy and of hypotheses being explained by other hypotheses.

Bereiter & Scardamalia notice an important problem that arises in ECHO when several hypotheses compete against each other: A hypothesis, H1, can get activation just by virtue of contradicting another hypothesis, H2, that gets negative activation because it is contradicted by a better hypothesis, H3. Thus H1 illicitly gets help from H3, because H3 drives H2 down, which activates H1. There are two ways of dealing with this problem, one by a trivial technical adjustment and the other by enriching the input. One can often expand the input to notice contradictions that were previously omitted. I do not see why Bereiter & Scardamalia's Satanic hypothesis ("The Devil is responsible for differences") cannot be construed as contradicting the hypotheses of evolution and natural selection, once the purported explanatory role of the Satanic hypothesis is further spelled out. On the technical side, the Common LISP version of ECHO allows one to set an output threshold so that once the activation of a unit drops below that threshold, it no longer has any effect on the activations of other units. In the runs of ECHO we have done so far, the threshold is effectively ignored by setting it at −1, the minimum activation level, but in simulations using the analogy-mapping program ACME (Holyoak & Thagard, in press), we routinely set the threshold at 0. This is because in ACME, as in Bereiter & Scardamalia's example, there are often multiple competing hypotheses, and it is necessary to prevent a poor hypothesis from getting accepted just because it contradicts one that is even worse. So in cases where there are more than two competing hypotheses, one can set the output threshold at 0, preventing the badness of one hypothesis from helping others that contradict it.

Simon compares ECHO unfavorably with STAHL, an impressive program that infers chemical components from descriptions of reactions in terms of inputs and outputs. The major distinction between ECHO and STAHL is obviously that STAHL is a discovery program whereas ECHO models evaluation. As I have frequently stressed, ECHO does not generate hypotheses; a more appropriate comparison might be STAHL versus the tag team of ECHO and PI (Thagard 1988a), which does some simple forms of abduction, although PI has not been applied to chemical examples. Independent of discovery, it should be clear that ECHO does more complicated kinds of theory evaluation than STAHL. STAHL only considers pieces of evidence based on the inputs and outputs of reactions, and even here it is historically limited. As I have summarized elsewhere (Thagard, in press b), Lavoisier's development of the oxygen theory took place over several years and clearly involved processes that go well beyond what STAHL is capable of. For example, the first input statement listed by Simon is

(reacts inputs {charcoal air} outputs {phlogiston ash air}).

This is clearly not a statement of evidence, because phlogiston is not observable (because, we would now say, it does not exist). This statement should really be treated as a hypothesis to be evaluated on the basis of what is in fact observed. Lavoisier's own writings show that he was dealing with data that went well beyond simple descriptions of inputs and outputs. Many of these involved quantitative relations – for example, that things gain weight when they undergo combustion or calcination. ECHO is undoubtedly inferior to STAHL in not considering the content of propositions, a consideration that is crucial for generating explanations and hypotheses, but ECHO is superior in that it is not restricted to a single domain or a simple method of hypothesis evaluation. It would be presumptuous to say of either ECHO or STAHL that it

"carries out genuine reasoning." STAHL has its strengths, but ECHO is a much more comprehensive model of the process of theory evaluation.

Zytkow, Simon's collaborator on STAHL, appropriately suggests that STAHL and the similar program GLAUBER can be used to generate input for ECHO. These programs can then be used to generate the hypotheses that ECHO evaluates, although they seem to me too limited to substantiate Zytkow's claim to generate "most of" this input. Still, the general account is right, that discovery programs like STAHL should be combined with evaluation programs like ECHO.

Zytkow also raises the question of whether TEC and ECHO could explain not just why Lavoisier thought his oxygen theory was best but why proponents of the phlogiston theory changed over to the oxygen theory. As he points out, the phlogiston theory was not static and was modified in the face of such discoveries as oxygen (dephlogisticated air) and hydrogen, which some theorists identified with phlogiston. I do not know whether the historical record is rich enough to trace the development of some of these phlogistinians, but I do not see any reason why ECHO could not be used to chart their development toward the oxygen theory as they gradually came to see it as more and more coherent.

4. Psychological adequacy

I agree with **Klayman & Hogarth** that the ECHO analyses presented in the target article do not constitute good tests of the psychological validity of TEC and ECHO. Such tests will have to be provided by controlled psychological experiments. In addition, efforts should be made to see whether ECHO naturally models some of the psychological effects that Klayman & Hogarth mention. Ranney and Thagard (1988) is just the beginning of what I hope will be a series of experiments pinning the empirical side of ECHO down more effectively.

Earle also notes some of the problems involved in testing ECHO's psychological validity. Researchers who try to test ECHO will have to be sensitive to the problems of input representation and free parameters. The latter is probably most easily dealt with, because the default parameters in ECHO have been applied to such a wide range of cases that it would seem fair to expect them to apply also to the results of new psychological experiments. Divining the belief systems and explanatory coherence relations of subjects will of course be difficult. But there is at least the promise of a series of psychological experiments by different researchers based on TEC and ECHO.

Cheng & Keane suggest two modifications they deem essential if TEC and ECHO are to be psychologically adequate. First, my account seems to them too holistic and parallel; this view is based on the grounds that people approach problems of theory evaluation in a much more piecemeal fashion. It is undoubtedly true that people do not consciously consider all the hypotheses and evidence simultaneously, and probably could not do so because of limitations of short-term memory. My assumption, however, is that evaluations of the explanatory coherence of a set of propositions occurs unconsciously, and at this level there is no reason to assume that it cannot be fully parallel. People make implicit judgments that something "makes sense" to them, based, I would argue, on this sort of holistic judgment. A full cognitive model would integrate ECHO with the processes of attention and conscious deliberation that Cheng & Keane rightly point to, but I see no serious problems in accomplishing the integration.

Chi raises the intriguing possibility that TEC and ECHO could provide a "transition mechanism" to explain conceptual change in ordinary people, particularly the radical restructuring that some psychologists have attributed to children. ECHO does help to understand how revolutionary conceptual change can take place in science (Thagard, in press b; Thagard & Nowak, in press), but the jury is still out concerning how similar such change is to what children undergo. By looking at scientific examples, I have been able to specify the transformations in conceptual and explanatory structures that have taken place in several scientific revolutions, whereas most of the discussions in the literature in developmental psychology are vaguer. Over the next few years, I expect that much progress will be made in determining the similarities and differences between scientists' and children's conceptual changes, once both are described more fully than has so far occurred.

Chi raises a number of important problems for my account as a full psychological model. Aspects of conceptual change that may be crucial include the realization that a particular hypothesis explains a particular piece of evidence. Chi suggests that someone who disagrees with a hypothesis may well resist encoding a piece of evidence as explained by the hypothesis. This is a psychological phenomenon that goes well beyond TEC and ECHO, and research is very much needed to see whether it can be modeled and to what extent it interferes with the application of considerations of explanatory coherence. Perhaps, as **Bereiter & Scardamalia** suggest, people can be taught not to resist alternative explanations and even to seek them out, just as graduate students are taught to eschew dogmatism. The empirical question concerns not just whether people are much worse than ECHO in integrating multiple pieces of evidence, but also whether they can be taught to be better at it.

I am excited that **Bereiter & Scardamalia** have had some success with 11-year-olds using ECHO, and am not surprised at the difficulties that arise. I agree that applying ECHO is problematic in domains where much of what is at issue is whether the evidence is any good. Our attempt to apply ECHO to the debate about parapsychology faltered because most of the issues there concern the quality of the experiments rather than the explanatory coherence of competing theories. TEC does not purport to be a general theory of inference, and in particular it does not apply to the statistical and methodological inferences that underlie data analysis. What Bereiter & Scardamalia call "contextual facts" figure in some of my examples, but not to the same degree as in theirs. One can easily imagine a con artist weaving a ridiculous hypothesis into a blanket of undisputed facts in such a way that a person fails to evaluate the hypothesis, merely seeing it as making sense with respect to the rest of the information. People may well be susceptible to this kind of strategy to an extent that would undermine their use of evidence and considerations of explanatory coherence. But we know that people can learn to get better at statistical reasoning,

and Bereiter & Scardamalia give us reason to hope that even children can be taught to evaluate hypotheses more effectively.

Read & Miller propose very inviting avenues for exploring the application of TEC and ECHO to social phenomena. Although I am enthusiastic about the research they propose, I want to offer a few words of caution concerning potential applications of explanatory coherence ideas to the phenomena they consider. First, investigators should be careful to distinguish explanatory from other notions of coherence. It would be illegitimate to give TEC or ECHO credit for accounting for the results of experiments that tapped into coherence phenomena that were independent of explanation. Second, as several commentators have suggested, more work needs to be done by researchers in psychology as well as in philosophy and AI concerning what explanations are. The knowledge structure approach advocated by Read will probably not constitute a full account of explanation. Nevertheless, I look forward to the results of Read & Miller's experiments, which I hope will suggest interesting extensions and revisions of the ECHO model.

I sum, I see several appealing avenues for continuing work on explanatory coherence. The most wide-open road is psychological experimentation to evaluate the adequacy of TEC and ECHO as accounts of human cognition. Theoretical development is also highly desirable, particularly in relation to the construction of a cognitive theory of explanation. Theory development should occur in the context of an attempt to develop a fully integrated computational model of the generation as well as the evaluation of explanatory hypotheses. Perhaps someday an ECHO analysis of TEC will show that explanatory coherence theory performs well by its own standards.

References

Letters a and r appearing before authors' initials refer to target article and response respectively.

Abelson, R. P. & Black, J. B. (1986) Introduction. In: *Knowledge structures*, ed. J. A. Galambos, R. P. Abelson & J. B. Black. Erlbaum. [SJR]

Achinstein, P. (1983) *The nature of explanation*. Oxford University Press. [aPT, PA]

(forthcoming) Hypotheses, probability, and waves. *British Journal for the Philosophy of Science*. [PA]

Ackermann, R. J. (1985) *Data, instruments and theory*. Princeton University Press. [PC-HC]

Alcock, J. E. (1987) Parapsychology: Science of the anomalous or search for the soul? *Behavioral and Brain Sciences* 10:553–65. [aPT]

Allen, R. J. (1986) A reconceptualization of civil trials. *Boston University Law Review* 66:401–37. [LJC]

Allport, G. W. (1964) The open system in personality theory. In: *Varieties of personality theory*, ed. H. H. Ruitenbeek. E. P. Dutton. [SJR]

Anderson, A. & Belnap, N. (1975) *Entailment*. Princeton University Press. [aPT]

Ashley, K. D. (1988) Arguing by analogy in law: A case-based model. In: *Analogical reasoning: Perspectives of artificial intelligence, cognitive science, and philosophy*, ed. D. H. Helman. Reidel. [GG]

Axelrod, R., ed. (1976) *Structure of decision: The cognitive maps of political elites*. Princeton University Press. [JK]

Bacon, F. (1859) Novum organum (first published 1620). In: *The works of Francis Bacon*, vol. I, ed. J. Spedding, R. Ellis & D. N. Heath. Longmans. [LJC, BM]

Bartlett, F. (1958) *Thinking: An experimental and social study*. Allen and Unwin. [CB]

BonJour, L. (1985) *The structure of empirical knowledge*. Harvard University Press. [aPT]

Bradley, F. H. (1914) *Essays on truth and reality*. Clarendon Press. [aPT]

Buchanan, B. & Shortliffe, E., eds. (1984) *Rule-based expert systems*. Addison Wesley. [aPT]

Carey, S. (1985) *Conceptual change in childhood*. MIT Press. [aPT]

Carnap, R. (1950) *Logical foundations of probability*. University of Chicago Press. [aPT]

Carter, L. (1984) *Reason in law*. Little, Brown. [aPT]

Cartwright, N. (1983) *How the laws of physics lie*. Oxford University Press. [WGL]

Cavendish, H. (1785) Experiments on air. *Philosophical Transactions* 75:372–84. [JMZ]

Charniak, E. & McDermott, D. (1985) *Introduction to artificial intelligence*. Addison-Wesley. [aPT]

Churchland, P. S. (1986) *Neurophilosophy*. MIT Press. [RNG]

Clark, H. & Lucy, P. (1975) Understanding what is meant from what is said: A study in conversationally conveyed requests. *Journal of Verbal Learning and Verbal Behavior* 14:56–72. [aPT]

Cohen, L. J. (1977) *The probable and the provable*. Oxford University Press. [rPT, LJC]

(1978) The coherence theory of truth. *Philosophical Studies* 34:351–60. [aPT]

(1981) Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences* 4:317–70. [RMD]

(1989) *An introduction to the philosophy of induction and probability*. Oxford University Press. [LJC]

Cohen R. (1983) A computational model for the analysis of arguments. Technical report CSRG-151. Department of Computer Science, University of Toronto. [aPT]

Darden, L. (1983) Artificial intelligence and philosophy of science: Reasoning by analogy in theory construction. In: *PSA 1982*, vol. 2, ed. P. Asquith & T. Nickles. Philosophy of Science Association. [RNG]

Darwin, C. (1962) *On the origin of species* (text of sixth edition of 1872). Macmillan. [aPT, BM]

Davis, P. J. & Hersh, P. (1981) *The mathematical experience*. Birkhauser. [RMD]

Dawes, R. M. (1971) A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist* 26:180–88. [MTHC]

(1988) *Rational choice in an uncertain world*. Harcourt, Brace, Javanovich. [JK]

Dawid, A. P. (1987) The difficulty about conjunction. *The Statistician* 36:91–97. [LJC]

DeJong, G. & Mooney, R. (1986) Explanation-based learning: An alternative view. *Machine Learning* 1:145–76. [aPT]

deKleer, J. & Williams, B. (1986) Reasoning about multiple faults. *Proceedings of the Fifth National Conference on Artificial Intelligence*. American Association for Artificial Intelligence [JAR]

Dietrich, E. (in press a) Programs in the search for intelligent machines: The mistaken foundation of AL. In: *The foundations of artificial intelligence: A source book*, ed. D. Patridge & Y. Wilks. Cambridge University Press. [ED]

(in press b) Computationalism. *Social Epistemology*. [ED]

Donovan, A., Laudan, L. & Laudan, R., eds. (1988) *Scrutinizing science: Empirical studies of scientific change*. Kluwer. [rPT]

Duhem, P. (1954) *The aim and structure of physical theory*, trans. P. Wiener (first published 1914). Princeton University Press. [aPT]

Edelman, G. (1987) *Neural Darwinism*. Basic Books. [DSL]

Eggleston, R. (1983) *Evidence, proof and probability*, 2nd ed. Weidenfeld and Nicholson. [LJC]

Einhorn, H. E. (1972) Expert measurement and mechanical combination. *Organizational Behavior and Human Performance* 7:86–106. [MTHC]

Einstein, A. (1952) On the electrodynamics of moving bodies. In: *The principle of relativity*, ed. H. A. Lorentz, A. Einstein, H. Minkowski & H. Weyl. Dover (originally published in 1905). [aPT]

Ennis, R. (1968) Enumerative induction and best explanation. *Journal of Philosophy* 65(18):523–29. [JRJ]

Falk, R. & Bar-Hillel, M. (1983) Probabilistic dependence between events. *Two Year College Mathematics Journal* 14(3):240–47. [RMD]

Falkenhainer, B., & Rajamoney, S. (1988) The interdependencies of theory formation, revision, and experimentation. In: *Proceedings of the Fifth International Conference on Machine Learning*, ed. J. Laird. Morgan Kaufmann. [aPT]

Feldman, J. & Ballard, D. (1982) Connectionist models and their properties. *Cognitive Science* 6:205–54. [aPT]

Feyerabend, P. (1975) *Against method*. London: Venso. [BM]

Fiske, S. & Taylor, S. (1984) *Social cognition*. Random House. [aPT]

Fodor, J. (1983) *The modularity of mind*. MIT Press. [aPT]

Fodor, J. A., & Pylyshyn, Z. W. (1988) Connectionism and cognitive

architecture: A critical analysis. In: *Connections and symbols*, ed. S. Pinker & J. Mehler. MIT Press. [JK]
Foster, M. & Martin, M., eds. (1966) *Probability, confirmation, and simplicity*. Odyssey Press. [aPT]
Franklin, A. (1986) *The neglect of experiment*. Cambridge University Press. [PC-HC]
Galison, P. (1987) *How experiments end*. University of Chicago Press. [PC-HC]
Gärdenfors, P. (1976) Relevance and redundancy in deductive explanations. *Philosophy of Science* 43:420–31. [MS]
Gardner, A. (1987) *An artificial intelligence approach to legal reasoning*. MIT Press/Bradford Books. [aPT]
Gibbs, R. (1984) Literal meaning and psychological theory. *Cognitive Science* 8:275–304. [aPT]
Giere, R. N. (1988) *Explaining science: A cognitive approach*. University of Chicago Press. [RNG]
Gluck, M. A., & Bower, G. H. (1988) From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General* 117:227–47. [JK]
Glymour, C. (1980) *Theory and evidence*. Princeton University Press. [aPT]
Glymour, C., Scheines, R., Spirtes, P. &/Kelly, K. (1987) *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modelling*. Academic Press. [RNG]
Goel, A., Ramanujam, J. & Sadayappan, P. (1988) Towards a neural architecture for abductive reasoning. Proceedings of the Second IEEE International Conference on Neural Networks, vol. 1, San Diego. [aPT]
Golding, M. (1984) *Legal reasoning*. Knopf. [aPT]
Goldman, A. (1986) *Epistemology and cognition*. Harvard University Press. [aPT, WGL]
Grossberg, S., ed. (1988) *Neural networks and natural intelligence*. MIT Press. [JK]
Grossberg, S. & Levine, D. S. (1987) Neural dynamics of attentionally modulated Pavlovian conditioning: Blocking, interstimulus interval, and secondary reinforcement. *Applied Optics* 26:5015–30. [DSL]
Hacking, I. (1982) Experimentation and scientific realism. *Philosophical Topics* 13:71–87. [WGL]
(1983) *Representing and intervening*. Cambridge University Press. [PC-HC]
Hanen, M. (1987) Unpublished manuscript. University of Calgary. [aPT]
Harman, G. (1965) The inference to the best explanation. *Philosophical Review* 74:88–95. [JRJ]
(1968) Enumerative induction as inference to the best explanation. *Journal of Philosophy* 65(18):529–33. [JRJ]
(1973) *Thought*. Princeton University Press. [arPT, ED]
(1986)*Change in view: Principles of reasoning*. MIT Press/Bradford Books. [aPT, WGL]
Harman, G., Ranney, M., Salem, K., Doring, F., Epstein, J. & Jaworksa, A. (1988) A theory of simplicity. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Erlbaum. [aPT]
Hedges, L. (1987) How hard is hard science, how soft is soft science? *American Psychologist* 42:443–55. [aPT]
Hegel, G. (1967) *The phenomenology of mind*, trans. J. Baillie (first published 1807). Harper & Row. [aPT]
Hempel, C. (1965) *Aspects of scientific explanation*. Free Press. [aPT, PA]
(1966) *Philosophy of natural science*. Prentice-Hall. [RNM]
Hendler, J. (1987) Marker-passing and microfeatures. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann. [aPT]
Himmelfarb, G. (1962) *Darwin and the Darwinian revolution*. Doubleday. [BM]
Hinton, G. E. (1981) A parallel computation that assigns canonical object-based frames of reference. *Proceedings of the International Joint Conference on Artificial Intelligence*, Vancouver, Canada. [BM]
Hobbs, J., Stickel, M., Martin, P. & Edwards, D. (1988) Interpretation as abduction. *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*. [aPT]
Hogarth, R. M. & Einhorn, H. J. (1989) *Order effects in belief updating: The belief-adjustment model*. Working paper. Center for Decision Research, University of Chicago. [JK]
Holland, J., Holyoak, K., Nisbett, R. & Thagard, P. (1986) *Induction: Processes of inference, learning, and discovery*. MIT Press/Bradford Books. [arPT]
Holyoak, K. & Thagard, P. (in press) Analogical mapping by constraint satisfaction. *Cognitive Science*. [arPT]
Jones, E. & Davis, K. (1965) From acts to dispositions: The attribution process in person perception. In: *Advances in experimental social psychology*, vol. 2, ed. L. Berkowitz. Academic Press. [aPT]
Josephson, J. (1982) *Explanation and induction*. Ph.D. thesis, Ohio State University. (University Microfilms DEO 82-22107) [JRJ]
Josephson, J., Chandrasekaran, B., Smith, J. & Tanner, M. (1987) A mechanism for forming composite explanatory hypotheses. *IEEE Transactions on Systems, Man, and Cybernetics* 17:445–54. [aPT, JRJ, JAR]
Kamin, L. (1974) *The science and politics of IQ*. Halstead Press. [CB]
Kant, I. (1787/1963) *Critique of pure reason*. N. Kemp Smith, translator, 2nd edition. Macmillan. [BM]
(1790/1951) *Critique of judgement*. J. H. Bertrand, translator. Collier-Macmillan. [BM]
Kaye, D. H. (1986) Do we need a calculus of weight to understand proof beyond reasonable doubt? *Boston University Law Review* 66:657–72. [LJC]
Kintsch, W. (1988) The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review* 95:163–82. [aPT]
Kirwan, R. (1789/1968) *An essay on pholgiston and the constitution of acids* (new impression of second English edition). Cass. [aPT]
Kirwan, R. (1789) *An essay on phlogiston and the constitution of acids*. London: J. Johnson. [JMZ]
Kitcher, P. (1981) Explanatory unification. *Philosophy of Science* 48:507–31. [aPT, ED]
Klayman, J. & Ha., Y.-W. (1987) Confirmation, disconfirmation, and information in hypothesis-testing. *Psychological Review* 94(2):211–28. [JK]
(1989) Hypothesis testing in rule discovery: Strategy, structure and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15:596–604. [JK]
Koehler, J. J. (1989) *Judgments of evidence quality among scientists as a function of prior beliefs and commitments*. Unpublished doctoral dissertation, University of Chicago, [JK]
Kuhn, T. S. (1970) *Structure of scientific revolutions* (2nd ed., first published 1962). University of Chicago Press. [aPT, BM]
(1977) *The essential tension*. University of Chicago Press. [ED]
Kunda, Z. (1987) Motivation and inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology* 53:636–47. [arPT]
Lakatos, I. (1970) Falsification and the methodology of scientific research programs. In: *Criticism and the growth of knowledge*, ed. I. Lakatos & A. Musgrave. Cambridge University Press. [aPT, LJC, MS]
Lalljee, M. & Abelson, R. P. (1983) The organization of explanations. In: *Attribution theory: Social and functional extensions*, ed. M. Hewstone. Blackwell. [SJR]
Langley, P., Simon, H. A., Bradshaw, G. L. & Zytkow, J. (1987) *Scientific discovery: Computational explorations of the creative process*. MIT Press. [HAS, JMZ]
Langley, P., Zytkow, J. M., Simon, H. A. & Bradshaw, G. (1983) Mechanisms for qualitative and quantitative discovery. In: *Proceedings of the International Machine Learning Workshop*, ed. R. S. Michalski. University of Illinois at Urbana-Champaign. [JMZ]
Laudan, L. (1976) Two dogmas of methodology. *Philosophy of Science* 43:585–97. [aPT]
Lavoisier, A. (1862) *Oeuvres* (6 vols.). Paris: Imprimerie Impériale. [aPT, JMZ]
Leake, D. (1988) Evaluating explanations. *Proceedings of the Seventh National Conference on Artificial Intelligence*. Morgan Kaufmann. [aPT]
Lehnert, W. (1987) Case-based problem solving with a large knowledge base of learned cases. *Proceedings of the Sixth National Conference on Artificial Intelligence*. Morgan Kaufmann. [aPT]
Lehrer, K. (1974) *Knowledge*. Clarendon Press. [aPT]
Leibniz, G. W. (1765/1981) *New essays on human understanding*, ed. P. Remnant & J. Bennett. Cambridge University Press. [BM]
(1865) Letter to Conring, 19 March 1678. In: *Die philosophische Schriften von Gottfried Wilhelm Leibniz*, vol. 1, ed. C. I. Gerhardt. Wiedmannsche Buchhandlung. [LJC]
Levi, I. (1967) *Gambling with truth*. Knopf. [PA]
(1980) *The enterprise of knowledge*. MIT Press. [aPT]
Levine, D. S. (in press) *Introduction to neural and cognitive modeling*. Erlbaum. [JK]
Lycan, W. (1988) *Judgment and justification*. Cambridge University Press. [aPT, WGL]
MacKinnon, E. (1978) The development of Kant's conception of scientific explanation. Proceedings of the 1978 Biennial Meeting of the Philosophy of Science Association. [BM]
Marr, D. (1982) *Vision*. Freeman. [aPT]
Messick, D. M. & van de Geer, J. P. (1981) A reversal paradox. *Psychological Bulletin* 90:582–93. [RMD]

Mitchell, T., Keller, R. & Kedar-Cabelli, S. (1986) Explanation-based generalization: A unifying view. *Machine Learning* 1:47–80. [aPT]

Musgrave, A. (1976) Why did oxygen supplant phlogiston? Research programmes in the chemical revolution. In: *Method and appraisal in the physical sciences*, ed. C. Howson. Cambridge University Press. [JMZ]

Neisser, U. (1976) *Cognition and Reality*. Freeman. [RMD]

Nersessian, N. J. (1984) *Faraday to Einstein: Constructing meaning in scientific theories*. Nijhoff. [RNG]

Nisbett, R. E. & Ross, L. (1980) *Human inference: Strategies and shortcomings of social judgement*. Prentice-Hall. [aPT]

Nisbett, R. E. & Wilson, T. D. (1977) Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84:231–59. [DCE]

Nowak, G. & Thagard, P. (forthcoming) Copernicus, Newton, and explanatory coherence. *Minnesota Studies in the Philosophy of Science*. [aPT]

O'Rorke, P., Aha, D. & Sage, S. (1988) A deductive-nomological model of abduction with learning. Unpublished manuscript. University of California at Irvine. [aPT]

Partington, J. R. (1962) *A history of chemistry*, vol. 3. Macmillan/St. Martin's Press. [JMZ]

Pearl, J. (1986) Fusion, propagation, and structuring in belief networks. *Artificial Intelligence* 29:241–88. [aPT]

(1987) Distributed revision of composite beliefs. *Artificial Intelligence* 33:173–215. [aPT]

Peirce, C. S. (1931–1958) *Collected papers* (8 vols.), ed. C. Hartshorne, P. Weiss & A. Burks. Harvard University Press. [aPT]

Peng, Y. & Reggia, J. (1987) A probabilistic causal model for dignostic problem solving. *IEEE Transactions on Systems, Man, and Cybernetics* 17:146–62; 395–406. [JAR]

(in press) A connectionist model for diagnostic problem solving. *IEEE Transactions on Systems, Man and Cybernetics*. [aPT, JAR]

Pennington, N. & Hastie, R. (1986) Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology* 51:242–58. [aPT, JK]

(1987) Explanation-based decision making. *Proceedings of the Ninth Annual Meeting of the Cognitive Science Society*. Erlbaum. [aPT, RMD]

(1988) Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14(3):521–33. [RMD, JK]

Perrin, C. (1988) The chemical revolution: Shifts in guiding assumptions. In: *Scrutinizing science: Empirical studies of scientific change*, ed. A. Donovan, L. Laudan & R. Laudan. Kluwer. [aPT]

Poincare, H. (1952) *Science and method*. Dover Publications. [ED]

Polanyi, M. (1963) The potential theory of adsorption. *Science* 141:1010–13. [NEW]

Pople, H. (1973) On the mechanization of abductive logic. *Proceedings of the Third International Joint Conference on Artificial Intelligence*. [JAR]

(1977) The formation of composite hypotheses in diagnostic problem solving. In: *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann. [aPT]

Popper, K. (1959) *The logic of scientific discovery*. Hutchinson. [aPT]

Quine, W. V. (1953) *From a logical point of view*. Harper & Row. [WGL]

(1960)*Word and object*. MIT Press. [WGL]

(1961) *From a logical point of view* (2nd ed.). Harper Torch-books. [aPT, BM]

Rajamoney, S. & DeJong, G. (1988) Active explanation reduction: An approach to the multiple explanations problem. In: *Proceedings of the Fifth International Conference on Machine Learning*, ed. J. Laird. Morgan Kaufmann. [aPT]

Ranney, M. (1987) Changing naive conceptions of motion. Doctoral dissertation, Learning, Research, and Development Center, University of Pittsburgh. [aPT]

Ranney, M. & Thagard, P. (1988) Explanatory coherence and belief revision in naive physics. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Erlbaum. [arPT, MTHC, JK]

Rao, K. R. & Palmer, J. (1987) The anomaly called psi: Recent research and criticism. *Behavioral and Brain Sciences* 10:539–51. [aPT]

Read, S. J. (1987) Constructing causal scenarios: A knowledge structure approach to causal reasoning. *Journal of Personality and Social Psychology* 52:288–302. [SJR]

Read, S. J. & Miller, L. C. (1989a) Inter-personalism: Toward a goal-based theory of persons in relationships. In: *Goal concepts in personality and social psychology*, ed. L. Pervin. Erlbaum. [SJR]

(1989b) The importance of goals in personality: Toward a coherent model of persons. In: *Social intelligence*, vol. 2, ed. R. S. Wyer, Jr. & T. K. Srull. Erlbaum. [SJR]

Reggia, J., Nau, D. & Wang, P. (1983) Diagnostic expert systems based on a set covering model. *International Journal of Man-Machine Studies* 19:437–60. [aPT, JAR]

Reggia, J., Nau, D., Wang, P. & Peng, Y. (1985) A formal model of diagnostic inference. *Information Sciences* 37:227–85. [JAR]

Reiter, R. (1987) A theory of diagnosis from first principles. *Artificial Intelligence* 32:57–95. [JAR]

Rescher, N. (1973) *The coherence theory of truth*. Clarendon Press. [aPT]

Rose, D. & Langley, P. (1986) Chemical discovery as belief revision. *Machine Learning* 1:423–52. [JMZ]

Ross, L. Lepper, M. R. (1980) The perserverance of beliefs: Empirical and normative considerations. In: *New directions for methodology of social and behavioral science: Fallible judgement in behavioral research*, vol. 4, ed. R. Shweder & D. W. Fiske. Jossey-Bass. [JK]

Rumelhart, D. E., McClelland, J. R. & the PDP Research Group (1986) *Parallel distributed processing: Explorations in the microstructure of cognition*, 2 vols. MIT Press. [aPT]

Salmon, W. (1966) *The foundations of scientific inference*. University of Pittsburgh Press. [aPT, ED]

(1970) Statistical explanation. In: *The nature and function of scientific theories*, ed. R. Colodny. University of Pittsburgh Press. [aPT]

(1984)*Scientific explanation and the causal structure of the world*. Princeton University Press. [aPT]

Schank, R. (1986) *Explanation patterns*. Erlbaum. [aPT]

Schank, R. C. & Abelson, R. P. (1977) *Scripts, plans, goals, and understanding*. Erlbaum. [SJR]

Schneider, W. & Detweiler, M. (1987) A connectionist/control architecture for working memory. In: *The psychology of learning and motivation*, vol. 21, ed. G. H. Bower. Academic Press. [GG]

Schoeman, F. (1986) Cohen on inductive probability and the law of evidence. *Philosophy of Science* 54:76–91. [LJC]

Schum, D. (1986) Probability and the processes of discovery, proof, and choice. *Boston University Law Review* 66:825–76. [LJC]

Scriven, M. (1962) Explanations, predictions, and laws. In: *Minnesota studies in the philosophy of science*, vol. 3, ed. H. Feigl & G. Maxwell. University of Minnesota Press. [JRJ]

Sellars, W. (1963) *Science, perception, and reality*. Routledge and Kegan Paul. [WGL]

Stalker, D. F., ed. (1984) *Linguistics and Philosophy* 7(1). [Special issue on "coherence."] [WGL]

Taubman, B. (1988) *The preppy murder trial*. St. Martin's Press. [aPT]

Thagard, P. (1978) The best explanation: Criteria for theory choice. *Journal of Philosophy* 75:76–92. [aPT, MS]

(1986) Parallel computation and the mind-body problem. *Cognitive Science* 10:301–18. [aPT]

(1988a) *Computational philosophy of science*. MIT Press/Bradford Books. [arPT, RNG]

(1988b) The dinosaur debate: Application of a connectionist model of theory evaluation. Unpublished manuscript, Princeton University. [aPT]

(in press a) Connectionism and epistemology: Goldman on winner-take-all networks. *Philosophia*. [aPT]

(in press b) The conceptual structure of the chemical revolution. *Philosophy of Science*. [arPT]

Thagard, P., Holyoak, K., Nelson, G. & Gochfeld, D. (1989) Analogical retrieval by constraint satisfaction. Unpublished manuscript, Princeton University. [aPT]

Thagard, P. & Kunda, Z. (1987) Hot cognition: Mechanisms of motivated inference. In: *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, ed. E. Hunt. Erlbaum. [rPT]

Thagard, P. & Nowak, G. (1988) The explanatory coherence of continental drift. In: *PSA 1988*, vol. 1, ed. A. Fine & J. Leplin. Philosophy of Science Association. [aPT, MS]

(in press) The conceptual structure of the geological revolution. In: *Computational models of discovery and theory formation*, ed. J. Shrager & P. Langley. Erlbaum. [arPT]

Trabasso, T., Secco, T. & van den Broek, P. (1984) Causal cohesion and story coherence. In: *Learning and comprehension of texts*, ed. H. Mandl, N. Stein & T. Trabasso. Erlbaum. [aPT]

Tribe, L. H. (1971) Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review* 84(6):1329–93. [RMD]

Tweney, R. D. (1985) Faraday's discovery of induction: A cognitive approach. In: *Faraday rediscovered*, ed. D. Gooding & F. James. Stockton. [RNG]

Urbach, P. (1974) Progress and degeneration in the "IQ debate." *British Journal of the Philosophy of Science* 25:99–135; 235–59. [CB]

van Fraassen, B. (1980) *The scientific image*. Clarendon Press. [aPT, WGL]

Wald, J., Farach, M., Tagamets, M. & Reggia, J. (in press) Generating

plausible diagnostic hypotheses with self-processing causal networks. *Journal of Experimental and Theoretical Artificial Intelligence.* [JAR]

Waltz, D. L. & Pollack, J. B. (1985) Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science* 9(1):51–74. [PO]

Whewell, W. (1967) *The philosophy of the inductive sciences* (first published 1840). Johnson Reprint. [aPT]

Wilensky, R. (1983) *Planning and understanding: A computational approach to human reasoning.* Addison-Wesley. [SJR]

Williams, G. (1979) The mathematics of proof. *Criminal Law Review* 297:308; 340–54. [LJC]

Zytkow, J. M. & Lewenstam, A. (1982) Czy tlenowa teoria Lavoisiera byla lepsza od teorii flogistonowej? (Was the oxygen theory of Lavoisier better than the phlogiston theory?) *Studia Filozoficzne* 202–203: 39–65. [JMZ]

Zytkow, J. M. & Simon, H. A. (1986) A theory of historical discovery: The construction of componential models. *Machine Learning* 1:107–36. [JMZ]