CHAPTER 4

Explanatory Coherence

WHY did the oxygen theory of combustion supersede the phlogiston theory? Why is Darwin's theory of evolution by natural selection superior to creationism? This chapter 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. Both the theory and the program have been improved since their original presentation (Thagard 1989).

The theory of explanatory coherence, TEC for short, is central to the general theory of conceptual change in science. As we saw in the last chapter, conceptual revolutions require a mechanism that can lead people to abandon an old conceptual system and adopt a new one. The view of conceptual organization proposed in Chapters 2 and 3 renders implausible the claim that major conceptual replacements can take place in an incremental, evolutionary way. A more global mechanism that can install a new conceptual system is needed to account for such transitions, in which a new set of explanatory hypotheses intertwined with a new conceptual system replaces an old set. Whereas the last two chapters were concerned with how concepts are structured, this chapter describes how propositional systems are structured via relations of explanatory coherence.

The problem of inference to explanatory hypotheses has a long history in philosophy and a much shorter one in psychology and artificial intelligence. Scientists and philosophers have long considered the evaluation of theories on the basis of their explanatory power. In the late nineteenth century, C. S. 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 1988). 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 (Pople 1977; Peng and Reggia 1990; Josephson et al. 1987) and natural language interpretation (Hobbs, Stickel, Appelt, and Martin 1990; Charniak and McDermott 1985). 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 and Taylor 1984). 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 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 motive: 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 chapter presents a theory of explanatory coherence that is intended to account for a wide range of explanatory inferences. I shall propose 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: the statement of a theory, the description of an algorithm, and the application to diverse examples that show the feasibility of the algorithm and help to demonstrate the power of the theory. Finally, I consider a series of objections to the theory, TEC, and to the implementation, ECHO. Chapter 5 describes in more detail how explanatory coherence contributes to conceptual change and discusses related issues concerning rationality and explanation.

4.1 A THEORY OF EXPLANATORY COHERENCE

Suppose that a professor is found dead, crushed beneath a Sun Workstation. Detectives investigating the case are assigned the task of determining what happened. Suppose further that a student was seen lurking around the professor's office, and that the student's fingerprints are found on the workstation. The detectives will then likely hypothesize that the student murdered the professor, since that hypothesis explains the death, the lurking, and the finger-

prints. Of course, other hypotheses need to be taken into consideration: the murder might have been done by a dean, or it might have been an ingenious suicide; the fingerprints may be there because the student was doing some maintenance on the workstation. The detectives will naturally inquire whether the student had a motive, and they would be excited to find that the student had had a major quarrel with the professor about a course in which the student had been accused of submitting plagiarized work. The student's motive would explain the hypothesis that the student committed the murder, which then gains credibility from being explained as well as from explaining the evidence. Deciding whodunit is an exercise in explanatory coherence, requiring an assessment of the hypothesis that the student committed the murder on the basis of how well it fits with the evidence and other hypotheses. A theory of explanatory coherence should show how such assessments, in science as well as everyday life, can be made.

4.1.1 Coherence

Before presenting the theory, let me stress that I am not offering 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, which depends on relations of logical consistency and entailment among members of a set of propositions.

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

Semantic coherence, which 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 within a set of propositions.

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. Then 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.

Since 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.

In ordinary language, to cohere is to hold together, and explanatory coherence is 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. TEC takes "explanation" and "explain" as primitives (although see section 5.3), 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 competing explanations.

4.1.2 Principles of Explanatory Coherence

I now propose seven principles that establish relations of explanatory coherence and make possible an assessment of the acceptability of propositions in 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. Here are the principles.

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.[1]

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.

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. Competition.

If P and Q both explain a proposition P_i, and if P and Q are not explanatorily connected, then P and Q incohere. Here P and Q are explanatorily connected if any of the following conditions holds:

(a) P is part of the explanation of Q,

(b) Q is part of the explanation of P,

(c) P and Q are together part of the explanation of some proposition P_j.

Principle 7. Acceptability.

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

4.1.3 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 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

[1] In the original statement of TEC in Thagard (1989), Principle 3 included a second clause concerning disanalogies that is not included here because it lacks interesting scientific applications. The old Principle 7, system coherence, has similarly been deleted because it does little to illuminate actual scientific cases. A new Principle 6, competition, has been added to cover cases where noncontradictory hypotheses compete with each other. The old Principle 6, acceptability, becomes the new Principle 7.

will become clearer in the discussion of the connectionist implementation below.

Principle 2, Explanation, is by far the most important for assessing explanatory coherence, since 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 E, then P and E 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, since otherwise we could prove that any two propositions cohere. Unless we use a relevance logic (Anderson and 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 7, that the more a hypothesis explains, the more coherent and hence acceptable it is. Thus this principle subsumes the criterion of explanatory breadth (which William Whewell, 1967, called "consilience") that I have elsewhere claimed to be the most important for selecting the best explanation (Thagard 1978, 1988).

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 1963). I call two hypotheses that are used together in an explanation "cohypotheses." Again I assume that explanation is more restrictive than implication, since 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, 1988). 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. Harman et al. 1988; Foster and Martin 1966), 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. It is controversial whether analogy is of more than heuristic use, but scientists such as Charles Darwin have used analogies to defend their theories; his argument for evolution by natural selection is analyzed in Chapter 6. Principle 3 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 and Thagard 1989; Thagard et al. 1990).

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, only that it can stand on its own more successfully than a hypothesis whose sole justification is what it explains. A proposition Q may have some independent acceptability and still not be accepted if it is only coherent with propositions that are not themselves 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 nonexplanatory, based on background knowledge that observations of certain sorts are very likely to be true. From 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 different level, we have some confidence in the reliability of descriptions of experimental results in carefully refereed scientific journals. Observations may be "theory-laden," as Hanson (1958) urged, but they are far from being theory-determined. I count as data not just individual observations such as "the instrument dial reads 0.5," but also generalizations from such observations. See section 9.1 for a discussion of the distinction between theoretical hypotheses and empirical generalizations.

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 my earlier version of TEC, I tried to stretch "contradiction" to cover cases where hypotheses that are not strictly contradictory are nevertheless held to be incompatible, but such cases are better handled by the new Principle 6, Competition.

According to Principle 6, we should assume that *hypotheses that explain the same evidence compete with each other unless there is reason to believe otherwise*. Hence there need be no special relation between two hypotheses for them to be incoherent, since hypotheses that explain a piece of evidence are judged to incohere unless there are reasons to think that they cohere. Not all alternative hypotheses incohere, however, since 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. Normally, however, if hypotheses are proposed to explain the same evidence, they will be treated as competitors. For example, in the debate over dinosaur extinction (Thagard 1991b), scientists generally treat as contradictory the 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. According to Principle 6, they incohere because both are claimed to explain why dinosaurs became extinct and there is no explanatory relation between them.

Principle 7, Acceptability, proposes in part 7(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, since 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 7(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 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? 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 directly relevant, where similarity is a matter of dealing with phenomena of the same kind. Thus a theory of business cycles 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.

According to TEC, a new theory will replace an old one if its hypotheses possess greater explanatory coherence. But TEC is still too vague to show how this could work. To show how to compute the acceptability of competing hypotheses, I now describe a program that implements TEC.