7 Consensus

Philosophical and psychological discussions of coherence, including the ones in the previous six chapters, are generally concerned with coherence in the mind of a single person. But the achievement of coherent systems of representations is a social process as well as an individual cognitive one. In many reasoning tasks, from evaluating scientific theories to making ethical decisions, people often rely on information received from others. The effective functioning of many kinds of groups, from scientific research teams to corporate divisions, requires that their members reach consensus about what to believe and what to do.

This chapter presents a theory of consensus based on coherence and communication. It presumes that individuals reach their own conclusions by evaluating the relative coherence of competing positions, and that consensus arises in a group when communication ensures that the individuals in the group share approximately the same set of elements that contribute to coherence evaluation. Conferences and other social processes that serve to increase communication thereby help scientists and medical practitioners to reach common conclusions about what to believe and what to do.

This chapter presents a computational model of consensus formation that clarifies how coherence and communication can lead to agreement. The model is unavoidably a great simplification of consensus formation in real groups, but it serves to highlight some of the key factors in the achievement of consensus. After describing the model's application to arguments at recent medical consensus conferences concerning the causes and treatment of peptic ulcers, I discuss a second application to debates concerning the origin of the moon. The desired result of the model is increased appreciation of the epistemic contributions of medical consensus conferences, as well as a deeper understanding of the general process of consensus. At the end of the chapter I discuss why consensus is more difficult to achieve in ethics than in science.

1 CONSENSUS IN SCIENCE AND MEDICINE

Since 1977 the U.S. National Institutes of Health have held more than one hundred consensus-development conferences. The purpose of these conferences is to produce consensus statements on controversial issues in medicine that are important to health-care providers, patients, and the general public. Many countries besides the United States hold similar events to help establish effective medical practices based on the best evidence available. Typically, experts on a medical issue make presentations to a panel or jury, who weigh the evidence and produce a consensus report reflecting their evaluation.

In other areas of science, consensus formation takes place less formally. It is common for controversial issues to be debated at conferences, but without an official panel to report a consensus. Implicitly, the entire scientific community serves as a kind of jury to evaluate competing theories on the basis of available evidence. Consensus does not always arise, but especially in the natural sciences it is

not unusual for debate to give way to substantial agreement on issues that were previously controversial. For example, a consensus on the origin of the moon arose from a 1984 conference held in Kona, Hawaii. According to one of its organizers, G. Jeffrey Taylor:

> Given the tenacity with which scientists cling to their views, none of us suspected that one of the hypotheses of lunar origin would spring forth as a leading candidate above the others. Certainly none of us thought the postconference favorite would not be one of the three classic hypotheses. Each of these hypotheses had what some considered to be fatal flaws. Each also had ardent supporters. It is a testament to human persistence and imagination that so many scientists tried so hard to adapt their preferences to a growing list of facts. (1994, 41)

Thus in science and medicine, consensus can emerge from controversy.

2 A MODEL OF CONSENSUS

The proposed theory of consensus can be summarized in the following theses:

- People make inferences about what to believe and what to do on the basis of judgments of coherence (chaps. 2–3). In particular, scientists evaluate competing theories by their comparative explanatory coherence, and they evaluate alternative practical actions using deliberative coherence. Coherence can be construed as maximization of constraint satisfaction, and can be computed by connectionist (artificial neural network) and other algorithms.
- Disagreement exists when individuals reach different coherence-based conclusions about what to accept and what to reject. Consensus is achieved by a group when all members of the group accept and reject the same sets of

elements, which are representations that can include propositions, such as hypotheses and descriptions of evidence, as well as nonpropositional representations.

• Consensus arises when individuals in a group exchange information to a sufficient extent that they come to make the same coherence judgments about what to accept and what to reject. The information exchange involves both elements to be favored in a coherence evaluation (e.g., evidential propositions that describe the results of observation and experiment) and descriptions of the explanatory and other relations that hold between elements.

These theses are rather general and vague, but they can be made much more precise by describing a computational model that implements them and makes possible experimentation with different ways in which coherence-based consensus can develop.

The new consensus model, called CCC for "consensus = coherence + communication," builds on the computational models of coherence described in chapter 2. In all of these models, conclusions are reached by maximizing satisfaction of constraints among elements that represent aspects of the inferential situation. Hence we can understand the inferences of individual members of a group in terms of each of them reaching conclusions that try to maximize coherence of their own particular sets of elements and constraints. But how can agreement arise between individuals who accept and reject different elements because they assume different elements and different constraints? In scientific disputes, how can agreement arise between scientists who accept different theories based on evidence and explanations?

Communication makes possible mutual coherence by enabling the transfer between individuals of both elements and constraints. Scientists, for example, can communicate

to each other information about the available evidence and about the explanatory relations that hold between hypotheses and evidence. This suggests the following straightforward process of consensus formation in science:

1. Start with a group of scientists who accept and reject different propositions because they reach different coherence judgements because of variations in evidence and explanations.

2. Exchange information between members of the group to change the coherence judgments made by the members.

3. Repeat (2) until the members have acquired sufficiently similar evidence and explanations so that all members accept and reject the same propositions; this is consensus.

The model CCC implements the process by representing each member of a group by a data structure:

Person

Name:

Favored elements:

Constraint input:

Accepts:

Rejects:

For simulations of scientific controversies involving explanatory coherence, the favored elements are propositions describing results of observation and experiments. Calling them "favored" does not mean that they cannot be rejected, only that their acceptance is encouraged in comparison with other elements representing hypotheses (see the discussion of discriminating coherentism in chapter 3). Even favored elements can be rejected if they fail to cohere optimally with other accepted elements. The constraint input includes statements of explanatory and contradictory relations. For example, in the ulcer controversy, the competing hypotheses included these:

AH1 Peptic ulcers are caused by excess acidity.

BH1 Peptic ulcers are caused by bacteria.

As well as other pieces of evidence about how people respond to different kinds of treatment, these hypotheses competed to explain the following primary piece of evidence:

E3 Some people get ulcers.

Constraint input can then include such information as the following:

(explain (AH1) E3)

(explain (BH1) E3)

The model CCC is implemented computationally in the programming language LISP, in which data and function calls are equally well written as lists, so these inputs that are part of the structure for a person can automatically be evaluated and produce new constraints (excitatory and inhibitory links). To evaluate coherence for a particular person, CCC uses the information about favored elements and inputs to create a network of units and links that can be used to spread activation to the units, and this results in the acceptance and rejection of units, which is recorded.

After performing a coherence calculation for all members of a given group, CCC checks for the presence of group consensus, which fails as soon as two members are found who differ in the propositions they accept or reject. Unless consensus already exists, communication begins, which enables members to acquire each other's elements and constraints. There are many ways in which communication might take place; here are the ones currently implemented:

Communication mode 1: random meetings Randomly pick two persons P_1 and P_2 to communicate with each other. Then transfer from P_1 to P_2 and vice versa is stochastic, in that whether a constraint input or favored element is transferred depends on a communication probability that ranges between 0 and 1. If communication probability in CCC is set high, then an element or input is more likely to be transferred than if it is set low.

Communication mode 2: lectures followed by random meetings A number of persons representing divergent opinions give "lectures," in which they are able to broadcast their elements and constraints to all other members of the group. Transfer of the information is still stochastic, in that the lecturer succeeds in transferring a favored element or input to a listener only with a certain probability. After the lectures, further communication continues by random meetings.

Although simple, this model can generate interesting experiments about the relative effects of variables such as group size and communication probability on the amount of time it takes to achieve consensus. The model differs dramatically from the only other formal model of consensus of which I am aware. Lehrer and Wagner (1981) present a mathematical means of finding a probability assignment that constitutes the best summary of the total information of a group. They do not address the processes, central to CCC, by which an individual reaches a coherence-based judgment about what to accept and reject, and by which individuals exchange information that affect each other's coherence judgments. On the other hand, their model incorporates an aspect not yet implemented in CCC: members of a group have opinions of the reliability of each other member of a group. A minor change to CCC could incorporate this aspect, which would make the transfer of information from one person to another a function not only of exchange probability but also of the degree of reliability that the receiver attributes to the sender. Because little information about such reliability judgments is available for the cases to which CCC has so far been applied, this important aspect of communication has not yet been implemented. Full implementation of reliability assessments would involve judgments of the trustworthiness of other members of the group, and hence require all the coherence-based inferences described in my discussion of trust in chapter 6. The next section describes experiments done with a more limited simulation of consensus formation in the ulcer controversy.

3 CONSENSUS AND THE CAUSES OF ULCERS

When Barry Marshall and Robin Warren proposed in 1984 that most peptic (gastric and duodenal) ulcers are caused by infection by a newly discovered bacterium, the medical community was highly skeptical. But by 1994 the evidence for their hypothesis had accumulated to such an extent that an NIH Medical Consensus Conference recommended that antibiotics be used to treat duodenal ulcers. It is now standard practice among gastroenterologists to test ulcer patients for the presence of *Helicobacter pylori* infection, whose eradication usually brings about a permanent cure. Thagard (1999) analyzed the cognitive and social processes that contributed to the dramatic shift in medical belief and practice.

The generally accepted view in 1983 that peptic ulcers are caused by excess acidity, and the dominant view in 1994 that bacterial infection accounts for most ulcers, can be represented by the following inputs to ECHO.

Dominant View in 1983

Evidence

(proposition E1 "Association between bacteria and ulcers.")

(proposition E2 "Warren observed stomach bacteria.")

(proposition E3 "Some people have stomach ulcers.")

(proposition E4 "Antacids heal ulcers.")

(proposition E5 "Previous researchers found no bacteria.")

Bacteria hypotheses

(proposition BH1 "Bacteria cause ulcers.")

(proposition BH2 "Stomach contains bacteria.")

Acid hypotheses

(proposition AH1 "Excess acidity causes ulcers.")

(proposition AH2 "Stomach is sterile.")

(proposition AH3 "Bacterial samples are contaminated.")

Bacteria explanations

(explain (BH1 BH2) E1)

(explain (BH2) E2)

(explain (BH1 BH2) E3)

Acid explanations

(explain (AH1 AH2 AH3) E1)

(explain (AH1 AH2 AH3) E2)

(explain (AH1) E3)

(explain (AH1) E4)

(explain (AH2) E5)

(data (E1 E2 E3 E4 E5))

There is no need for an explicit statement of which hypotheses contradict or compete with each other (e.g., AH1 and BH1), because the program ECHO automatically identifies hypotheses from different theories that compete to explain the same evidence (Thagard 1992b). ECHO then sets up inhibitory links between units representing pairs of competing hypotheses. When ECHO is run on this input, it reaches the same conclusion that most medical researchers did in 1983: the bacterial theory of ulcers should be rejected.

In contrast, the following input yields acceptance of the bacterial theory:

Dominant View in 1994

Evidence

(proposition E1 "Association between bacteria and ulcers.")

(proposition E2 "Many have observed stomach bacteria.")

(proposition E3 "Some people have stomach ulcers.")

(proposition E4 "Antacids heal ulcers.")

(proposition E6 "Marshall's 1988 study that antibiotics cure ulcers.")

(proposition E7 "Graham's 1992 study that antibiotics cure ulcers.")

(proposition E8 "Several other cure studies.")

(proposition E9 "Bacteria/acid study.")

Bacteria hypotheses

(proposition BH1 "Bacteria cause ulcers.")

(proposition BH2 "Stomach contains bacteria.")

(proposition BH3 "Bacteria produce acid.")

(proposition BH4 "Eradicating bacteria cures ulcers.")

Acid hypothesis

(proposition AH1 "Excess acidity causes ulcers.")

Bacteria explanations

(explain (BH1 BH2) E1)

(explain (BH2) E2)

(explain (BH1 BH2) E3)

(explain (BH1 BH2) BH4)

(explain (BH1 BH3) E4)

(explain (BH3) E9)

(explain (BH4) E6)

(explain (BH4) E7)

(explain (BH4) E8)

Acid explanations

(explain (AH1) E3)

(explain (AH1) E4)

(data (E1 E2 E3 E4 E6 E7 E8 E9))

It is evident, and ECHO simulations confirm, that explanatory coherence based on this information supports accepting the bacterial theory in 1994 even though it was widely rejected earlier.

The consensus problem here is, How did the medical community come to achieve consensus that most peptic ulcers are caused by bacteria? CCC can be used to model consensus formation in this case by creating a population of scientists that includes proponents of the 1983 view and proponents of the 1994 view. Communication in which evidence and explanations are transferred between scientists gradually leads to general agreement. We would expect that the time required for consensus to be reached would be affected by a number of factors, including these:

- The number of members of the scientific community
- The probability of exchange of information on a given encounter

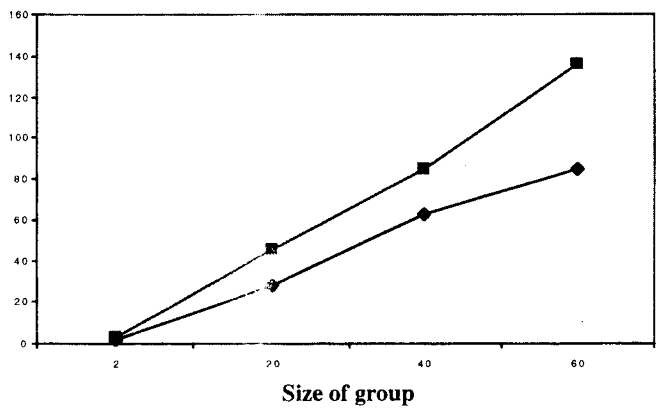

Figure 7.1
Time to consensus in the ulcer simulation, measured by number of meetings before full agreement was reached. The lower line shows the results of simulations that began with lectures. Exchange probability is held constant at 0.5. Results are the mean of five different simulations.

- The occurrence of lectures in which scientists can communicate simultaneously with a large number of other scientists
- The extent to which a superior view is initially distributed in the community

A series of computational experiments with CCC found that each of these factors influence the time to consensus.

Experiment 1 varied the size of the group of scientists seeking consensus, with groups of 2, 20, 40, and 60 members; half of the members started with the dominant view of ulcer causation in 1983, and half started with the dominant view in 1994. Figure 7.1 shows that the time required for consensus to be reached is a function of group size: the larger the group, the greater the number of meetings required to achieve consensus. Figure 7.1 also shows that, regardless of group size, lectures speed up the achievement of consensus. Computational experiment 2 held group size constant and varied exchange probability, which yielded the expected result that higher exchange probabilities produce consensus faster, in both the lecture and no-lecture conditions.

Both experiments 1 and 2 were unrealistic in that they began with half the members of the group of scientists holding each of the two competing theories. Historically, the bacteria theory of ulcers began with Marshall and Warren and then spread only very gradually through the community of gastroenterologists. Accordingly, computational experiment 3, which held group size constant at 40 and exchange probability constant at 0.5, varied proportions of the scientists beginning with the eventually dominant 1994 bacterial theory. In the toughest situation, starting with only 1 proponent of the bacterial theory of ulcers, it takes a long time for opinion to shift, an average of more than 250 meetings. Acceptance of the bacterial theory by the group is initially very slow, but accelerates rapidly as the theory spreads. When the simulation starts with 5 or 10 representatives of the bacterial theory, it reaches consensus much more rapidly, in around 100 meetings. The first simulation, starting with only 1 advocate of the bacterial theory, models much more closely the spread of the theory through the community of gastroenterologists.

The three computational experiments just described show that the CCC model displays some of the consensus behavior that one might expect of a scientific community. Consensus takes longer to achieve when group sizes are larger, when exchange probability is lower, when there are fewer members beginning with the dominant position, and when there are no lectures to jump-start communication. Similar results occur when CCC is applied to a different case, discussed in the next section.

It is rather artificial to have only two positions in the simulations, the 1983 rejection of the bacterial theory of ulcers and the 1994 acceptance. A full simulation of the case would have numerous individuals with many different starting points, arriving at agreement with the eventual consensus at different times. A more detailed account of the developments during the decade would explain more incrementally how beliefs such as AH2, "The stomach is sterile," could drop out of the picture by 1994. Despite the oversimplifications of the computational experiments so far accomplished, CCC provides the start of a model of how scientific consensus can arise through coherence and communication.

Can CCC account for cases where a scientific community fails to reach consensus? The computer simulations of the ulcer case allow exchange of information to be repeated until consensus is reached, but in the real world there are limits on the time and social opportunities for such exchange. Hence a community may not achieve consensus simply because it has not had enough instances of information exchanges with high enough exchange probabilities. If the simulations in figure 7.1 with 60 scientists had been stopped after only 40 interactions, then consensus would not have been achieved. More problematically, there may be communication barriers between scientists that prevent them from receiving each other's evidence and hypotheses, so that the exchange probability for some information drops to 0. Then consensus would never be reached, because the scientists would never end up making the same coherence calculations. I know of no cases in the history of science, however, where such complete communication breakdown has occurred: even the most major scientific revolutions have involved a high degree of comparability of competing theories (Thagard 1992b). Yet when two theories are conceptually very different,

scientists may have difficulty understanding the hypotheses proposed by their opponents, and they may have little trust in the evidence adduced by the other side. In such cases, the exchange probability would be very low, so the scientific community and CCC would take a long time to reach consensus.

4 CONSENSUS AND THE ORIGIN OF THE MOON

To run CCC on the dispute concerning the origin of the moon, I encoded the key evidence and hypotheses as input to the explanatory coherence program ECHO largely according to the analysis of the debate by Wood (1986; see also Hartmann, Philips, and Taylor 1986 and Brush 1996). The four main theories were the following:

- Moon-capture hypothesis: a fully formed moon was caught by the earth.
- Coaccretion hypothesis: the moon and earth formed concurrently from a cloud of gas and dust.
- Fission hypothesis: the moon formed by fission from a rapidly spinning earth.
- Giant-impact hypothesis: a Mars-sized body hit the earth.

The relevant evidence concerned comparisons of the composition of the earth and moon, as well as the high angular momentum of the earth-moon system.

CCC has so far been run on the moon example with groups of simulated scientists involving 4, 20, 40, and 60 members. Each simulation begins with one quarter of the scientists holding each of the four theoretical positions. Computational experiments found, as expected, that the amount of time (number of meetings between pairs of scientists) required before consensus is reached increases with the number of scientists, and decreases with higher probability of information exchange. Moreover, starting the simulation with four "lectures" in which proponents of the four theories broadcast to the whole group speeds up consensus formation. The results of these simulations are similar to those described for the ulcer example. They show that CCC is capable of interesting behavior even though it is very simple, compared with the complexities of consensus formation in real scientific communities, and they illustrate the benefits of enhancing the communication process by allowing more rapid lecturelike transmissions of information from one individual to many.

5 BENEFITS OF CONSENSUS CONFERENCES

In an earlier book (Thagard 1999, chap. 12) I assessed medical consensus conferences with respect to seven epistemic standards: reliability, power, fecundity, speed, efficiency, explanatory efficacy, and pragmatic efficacy (the first five of these derive from the work of Alvin Goldman 1992). I will not repeat that analysis here, but I will try to deepen it from the perspective of the CCC model of consensus formation.

The point of medical and scientific consensus conferences should be to help scientific communities reach common conclusions that are reliable (have a good ratio of truths to falsehoods), explanatorily powerful (make sense of the evidence), and practically efficacious (bring about nonepistemic benefits to people). In addition, they should help provide many answers to important questions (power) and make these answers available to many people (fecundity). Speed and efficiency are also relevant epistemic standards, since we want an epistemic practice to produce

answers quickly and at low cost. In accord with the CCC simulations of both the moon and ulcer cases, consensus conferences in which scientists begin with lectures and proceed with intense discussions serve to communicate evidence and explanations, and thereby produce speedy, efficient, and explanatorily efficacious decisions. Consensus conferences also increase the speed of interaction of scientists and medical practitioners, by bringing them all together in the same place. This dramatically increases the rate of pairwise and larger interactions between scientists.

Of course, there are many aspects of consensus conferences that are not captured by the CCC model as it currently stands. I have not yet attempted to model the role of the jury panel in meeting together to reach a consensus that is then communicated to a larger group and presumably has a substantial impact on the larger group's consensus. Moreover, the simulations so far have dealt only with issues of explanatory coherence, but there are legitimate and illegitimate ways in which medical decisions are also based on deliberative coherence, which evaluates the extent to which various actions affect goals. The legitimate contribution of deliberative coherence to medical decisions includes calculation of the extent to which different courses of action accomplish medical goals, such as curing as many people as possible, and social goals, such as keeping the cost of medicine down to a level that people and the government can sustain. Such common social goals are favored elements that can be communicated from one decision maker to another in the same way that favored elements representing evidence are communicated between scientists.

On the illegitimate side of theory evaluation, individual judgments about the causes and treatment of disease are sometimes affected by the individual goals of decision makers. Scientists and physicians are not saints, and their inferences may well be affected by their personal goals, such as their finances and their stature in the profession. One of the pioneers of the bacterial theory of ulcers once suggested to me that some gastroenterologists were reluctant to accept the idea of a quick antibiotic cure for ulcers because they would then lose their lucrative gastroscopy businesses and be reduced to conducting colonoscopies, making them no better than proctologists! Because CCC can incorporate any kind of coherence-based inference, it would be easy to incorporate such individual goals and deliberative coherence calculations into its simulations. Like ordinary people, the scientists in a CCC simulation of consensus building would then be liable to what Kunda (1990) calls motivated inference, in which personal goals affect the evaluation of evidence and hence the overall judgment of the reasoner. Moreover, different scientists will have different emotional valences attached to particular hypotheses, which will yield different judgments of emotional coherence.

It is crucial to note, however, that medical consensus conferences, like scientific communications in general, are structured so as to discourage the dissemination of such individual concerns. Medical practitioners cannot stand up and say, "We shouldn't adopt this treatment because it will reduce our income," even if that is what they are thinking. Decisions at consensus conferences are expected to be evidence-based, taking into account all the available data from the most carefully conducted clinical trials. (When I first heard the term "evidence-based medicine," I thought it was redundant, but in fact many medical treatments have yet to be assessed using the randomized and blinded clinical trials necessary to evaluate causal efficacy.) Public talks and comments (although not necessarily informal asides) must conform to the social norm of evaluating disease

explanations and potential treatments on the basis of dispassionately presented evidence. Hence what gets transferred between individuals at a consensus conference is not their quirky individual goals, but evidence, explanations, and socially acceptable general goals. Thus consensus conferences can serve to ensure not only that *some* decision be made, but also that the decision made does in fact maximize explanatory and deliberative coherence.

6 CONSENSUS IN VALUE JUDGMENTS

Controversies are common in science and medicine, but so is consensus reached as the result of the collective assessment of evidence. In ethics, politics, and aesthetics, however, it seems that the balance is tipped more toward controversy than consensus. My discussion of coherence and emotion in previous chapters points to several reasons why consensus is more problematic in issues concerning values. Whereas scientific controversies can be settled largely by evaluating the explanatory coherence of hypotheses with respect to the evidence, ethical and other value controversies require integration of the constraints of deliberative coherence. And whereas scientists are required to take seriously the evidence presented by other scientists, decision makers may not share the goals of other decision makers, and there is no immediate normative reason why they should. If your primary goal is world domination and human enslavement, there is no reason why I should give it any priority in my own assessments of deliberative coherence. The emotional valences that you attach to different hypotheses and possible actions need not correspond to my emotional valences.

A collective assessment of deliberative coherence has to be reached on the basis of agreed-upon high-level goals, such as flourishing, freedom, and fairness, as I discussed in chapter 5. But the relative weights of these constraints is an open question, and there is no obvious way that communication can help to overcome weighting differences. How can consensus rationally be reached between proponents of libertarianism, who put top priority on freedom, and proponents of socialism, who put top priority on fairness? The main hope for consensus comes from the possibility of constraint adjustment achieved through explanatory and analogical coherence, as political debaters consider the widest possible range of evidence concerning historical cases, social functioning, and cognitive-emotional processes. Not just cold constraints but also the hot valences that contribute to emotional coherence must be changed.

Broadening consensus formation about what to do so that it includes other kinds of coherence besides deliberative coherence does not always make consensus easier to achieve. For most people, ethical issues are closely tied in with metaphysical ones, because ethical education is commonly part of religion. Consider a person who strongly believes the following propositions:

- God exists.
- God determines what is right and wrong.
- The Bible is God's word.
- The bible says that abortion is wrong.

From these beliefs, it follows deductively that abortion is wrong, so the ethical judgment is strongly constrained by metaphysical beliefs. Achieving a consensus between this person and a proabortion atheist requires dramatic revisions in judgments based on explanatory and other kinds of coherence, as well as deliberative coherence.

Despite these impediments, the prospects for ethical and political consensus are not entirely bleak. By the end of the twentieth century, most educated people have come to agree on many ethical judgments, for example that slavery is wrong. I hope that further increase in understanding of how people think and feel and of how societies work will lead to further consensus.

7 SUMMARY

Consensus in a group can be reached as the result of communication that allows its members to exchange elements and constraints. Thus consensus arises by means of a combination of interpersonal communication and individual coherence assessments. In science and medicine, conferences are one of the means by which communication is increased and convergence on common coherence judgements is encouraged. Consensus in ethics and politics is often more problematic because exchange of goals and constraints in deliberative coherence is much harder to accomplish than exchange of hypotheses and explanations in explanatory coherence.

8 *Probability*

The model of consensus in the last chapter assumed that scientists evaluate competing hypotheses on the basis of their explanatory coherence. But this is not the only position in current philosophy of science and epistemology, where theory choice and belief revision are often discussed using probability theory. On the probabilistic view, one theory should be preferred to another if it has higher probability given the evidence. This chapter explores the relationship between probabilistic and coherentist approaches to inference.

1 TWO TRADITIONS IN CAUSAL REASONING

When surprising events occur, people naturally try to generate explanations of them. Such explanations usually involve hypothesizing causes that have the events as effects. Reasoning from effects to prior causes is found in many domains, including the following:

- Social reasoning: when friends are acting strange, we conjecture about what might be bothering them.
- Legal reasoning: when a crime has been committed, jurors must decide whether the prosecution's case gives a convincing explanation of the evidence.

- Medical diagnosis: from a set of symptoms, a physician tries to decide what disease or diseases produced them.
- Fault diagnosis in manufacturing: when a piece of equipment breaks down, a troubleshooter must try to determine the cause of the breakdown.
- Scientific theory evaluation: scientists seek an acceptable theory to explain experimental evidence.

What is the nature of such reasoning? The many discussions of causal reasoning over the centuries can be seen as falling under two general traditions, which I will call explanationism and probabilism. Explanationists understand causal reasoning qualitatively, while probabilists exploit the resources of the probability calculus to understand causal reasoning quantitatively. Explanationism goes back at least to Aristotle (1984, vol. 1, p. 128), who considered the inference that the planets are near as providing an explanation of why they do not twinkle. Some Renaissance astronomers such as Copernicus and Rheticus evaluated theories according to their explanatory capabilities (Blake 1960). The leading explanationists in the nineteenth century were the British scientist, philosopher, and historian William Whewell (1967), and the American polymath C. S. Peirce (1958). The most enthusiastic explanationists in this century have been epistemologists such as Gilbert Harman (1973, 1986) and William Lycan (1988). In the field of artificial intelligence, computational models of inference to the best explanation have been developed (Josephson et al. 1994, Shrager and Langley 1990, Thagard 1992b).

The probabilist tradition is less ancient than explanationism, for the mathematical theory of probability arose only in the seventeenth century through the work of Pascal, Bernoulli, and others (Hacking 1975). Laplace and Jevons were the major proponents of probabilistic approaches to

induction in the eighteenth and nineteenth century, respectively (Laudan 1981, chap. 12). Many twentieth-century philosophers have advocated probabilistic approaches to epistemology, including Keynes (1921), Carnap (1950), Jeffrey (1983), Levi (1980), Kyburg (1983), and Kaplan (1996).

Probabilistic approaches have recently become influential in artificial intelligence as a way of dealing with the uncertainty encountered in expert systems (D'Ambrosio 1999; Frey 1998; Jordan 1998; Neapolitain 1990; Pearl 1988, 1996; Peng and Reggia 1990). Probabilistic approaches are also being applied to natural-language understanding (Charniak 1993). The explanationist versus probabilist issue surfaces in a variety of subareas. Some legal scholars concerned with evidential reasoning have been probabilist (Lempert 1986, Cohen 1977), while some are explanationist and see probabilist reasoning as neglecting important aspects of how jurors reach decisions (Allen 1994, Pennington and Hastie 1986). In the philosophy of science, there is an unresolved tension between probabilist accounts of scientific inference (Achinstein 1991, Hesse 1974, Horwich 1982, Howson and Urbach 1989, Maher 1993) and explanationist accounts (Eliasmith and Thagard 1997; Lipton 1991; Thagard 1988, 1992b, 1999). Neither the probabilist nor the explanationist tradition is monolithic: there are competing interpretations of probability and inference within the former, and differing views of explanatory inference within the latter.

In recent years it has become possible to examine the differences between explanationist and probabilist approaches at a much finer level, because algorithms have been developed for implementing them computationally. As chapter 3 described, my theory of explanatory coherence incorporates the kinds of reasoning advocated by explanationists and is implemented in a connectionist program called ECHO, which shows how explanatory coherence can be computed in networks of propositions. Pearl (1988) and others have shown how probabilistic reasoning can be computationally implemented using networks. The question naturally arises of the relation between ECHO networks and probabilistic networks. This chapter shows how ECHO's qualitative input can be used to produce a probabilistic network to which Pearl's algorithms are applicable. At one level, this result can be interpreted as showing that ECHO is a special case of a probabilistic network.

The production of a probabilistic version of ECHO highlights several computational problems with probabilistic networks. The probabilistic version of ECHO requires the provision of many conditional probabilities of dubious availability, and the computational techniques needed to translate ECHO into probabilistic networks are potentially combinatorially explosive. ECHO can therefore be viewed as an intuitively appealing and computationally efficient approximation to probabilistic reasoning. We will also see that ECHO puts important constraints on the conditional probabilities used in probabilistic networks.

The comparison between ECHO and probabilistic networks does not in itself settle the relation between the explanationist and probabilist traditions, since there are other ways of being an explanationist besides ECHO, and there are other ways of being a probabilist besides Pearl's networks. But from a computational perspective, ECHO and Pearl networks are much more fully specified than previous explanationist and probabilist proposals, so a head-to-head comparison is potentially illuminating. After briefly reviewing Pearl's approach to probabilistic networks, I shall sketch the probabilistic interpretation of explanatory coherence and discuss the computational

problems that arise. Then a demonstration of how ECHO naturally handles Pearl's central examples will support the conclusion that the theory of explanatory coherence is not obviated by the probabilistic approach.

The point of this chapter, however, is not simply a comparison of two computational models of causal reasoning. Causal reasoning is an essential part of human thinking, so the nature of such reasoning is an important question for cognitive science. Do people use coherence-based or probabilistic inference when they evaluate competing causal accounts in social, legal, medical, engineering, and scientific contexts? Many researchers in AI and philosophy assume that probabilistic approaches are the only ones appropriate for understanding such reasoning, but there is much experimental evidence that human thinking is often not in accord with the prescriptions of probability theory (see, e.g., Kahneman, Slovic, and Tversky 1982; Tversky and Koehler 1994). On the other hand, there is some psychological evidence that explanatory coherence theory captures aspects of human thinking (Read and Marcus-Newhall 1993; Schank and Ranney 1991, 1992; Thagard and Kunda 1998). Moreover, as earlier chapters showed, coherence-based reasoning is pervasive in human thinking, in areas as diverse as perception, decision making, ethical judgments, and emotion. Clarification of the relation between explanatory coherence and probabilistic accounts is thus part of the general psychological project of understanding how human causal reasoning works.

The probabilistic view assumes that the degrees of belief that people have in various propositions can be described by quantities that comply with the principles of the mathematical theory of probability. In contrast, the explanationist approach sees no reason to use probability theory to model degrees of belief. Probability theory is an immensely valuable tool for making statistical inferences about patterns of frequencies in the world, but it is not the appropriate mathematics for understanding human inference in general.

2 PROBABILISTIC NETWORKS

The theory of explanatory coherence employs vague concepts such as explanation and acceptability, and ECHO requires input specifying explanatory relations. It is reasonable to desire a more precise way of understanding causal reasoning, for example, in terms of the mathematical theory of probability. That theory can be stated in straightforward axioms that establish probabilities as quantities between 0 and 1. From the axioms it is trivial to derive Bayes's theorem, which can be written thus:

$$P(H/E) = \frac{P(H) \times P(E/H)}{P(E)}$$

This equation says that the probability of a hypothesis given the evidence is the prior probability of the hypothesis times the probability of the evidence given the hypothesis, divided by the probability of the evidence. Bayes's theorem is very suggestive for causal reasoning, since we can hope to decide what caused an effect by considering which cause has the greatest probability, given the effect. Hence probabilists are often called Bayesians.

In practice, however, application of the probability calculus becomes complicated. Harman (1986, 25) has pointed out that in general probabilistic updating is combinatorially explosive, since we need to know the probabilities of a set of conjunctions whose size grow exponentially with the number of propositions. For example, full probabilistic information about three propositions, A, B, and C, would require knowing a total of 8 different values: $P(A \& B \& C)$, $P(A \& B \& \text{not } C)$, $P(A \& \text{not } B \& \text{not } C)$, etc. Only 30 propositions would require more than a billion probabilities. As Thagard and Verbeurgt (1998) have shown, coherence maximization is also potentially intractable computationally, but the algorithms described in chapter 2 provide efficient ways of computing coherence, and the semidefinite programming algorithm is guaranteed to accomplish at least 0.878 of the optimal constraint satisfaction.

Probabilistic networks enormously prune the required number of probabilities and probability calculations, since they restrict calculations to a limited set of dependencies. Suppose you know that B depends only on A, and C depends only on B. You then have the simple network $A \to B \to C$. This means that A can affect the probability of C only through B, so that the calculation of the probability of C can take into account the probability of B while ignoring that of A.

Probabilistic networks have gone under many different names: causal networks, belief networks, Bayesian networks, influence diagrams, and independence networks. For the sake of precision, I want to concentrate on a particular kind of probabilistic network that uses the elegant and powerful methods of Pearl (1988). Though methods for dealing with probabilistic networks other than his are undoubtedly possible, I will not try to compare ECHO generally with probabilistic networks, but will make the comparison specifically with Pearl networks.

In Pearl networks, each node represents a multivalued variable such as a patient's temperature, which might take three values: high, medium, low. In the simplest cases, the variable can be propositional, with two values, true and false. Already we see a difference between Pearl networks and ECHO networks, since ECHO requires separate nodes

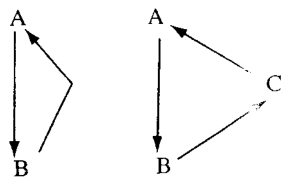

Figure 8.1
Examples of cyclic graphs.

for a proposition and its negation. But translations between Pearl nodes and ECHO nodes are clearly possible and will be discussed below.

More problematic are the edges in the two kinds of networks. Pearl networks are directed, acyclic graphs. Edges are directed, pointing from causes to effects, so that $A \rightarrow B$ indicates that A causes B and not vice versa. In contrast, ECHO's links are all symmetric, befitting the character of coherence and incoherence (principle E1 of chapter 3, section 1), but symmetries are not allowed in Pearl networks. The specification that the graphs be acyclic rules out relations such as those shown in figure 8.1. Since the nodes are variables, a more accurate interpretation of the edge $A \rightarrow B$ would be that the values of B are causally dependent on the values of A.

The structure of Pearl networks is used to localize probability calculations and surmount the combinatorial explosion that can result from considering the probabilities of everything, given everything else. Figure 8.2 shows a fragment of a Pearl network in which the variable D is identified as being dependent on A, B, and C, while E and F are dependent on D. The probabilities that D will take on its various values can then be calculated by looking only at A, B, C, E, and F and ignoring other variables in the network that D is assumed to be conditionally independent of, given the five variables on which it is directly dependent. The

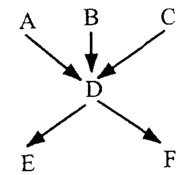

Figure 8.2
Sample Pearl network, in which the variable D is dependent on A, B, and C, while E and F are dependent on D. Lines with arrows indicate dependencies.

probabilities of the values of D can be expressed as a vector corresponding to the set of values. For example, if D is temperature and has values (high, medium, low), the vector (0.5 0.3 0.2) assigned to D means that the probability of high temperature is 0.5, of medium temperature is 0.3, and of low temperature is 0.2. In accord with the axioms of probability theory, the numbers in the vector must sum to 1, since they are the probabilities of all the exclusive values of the variable.

The desired result of computing with a Pearl network is that each node should have a stable vector representing the probabilities of its values, given all the other information in the network. If a measurement determines that the temperature is high, then the vector for D would be (1 0 0). If the temperature is not known, it must be inferred using information gathered from both the variables on which D depends and the ones that depend on D. In terms of Bayes's theorem, we can think of A, B, and C as providing prior probabilities for the values of D, while E and F provide observed evidence for them. The explanatory-coherence interpretation of figure 8.2 is that A, B, and C explain D, while D explains E and F. For each variable X, Pearl uses BEL(x) to indicate the computed degrees of belief (probabilities) that X takes on for each of its values x. BEL(x) is thus a vector with as many entries as X has values, and is calculated using the following equation:

$$\text{BEL}(x) = \alpha \times \lambda(x) \times \pi(x)$$

Here α is a normalizing constant used to ensure that the entries in the vector sum to 1; $\lambda(x)$ is a vector representing the amount of support for particular values of X coming up from below, that is, from variables that depend on X; and $\pi(x)$ is a vector representing the amount of support for particular values of X coming down from above, that is, from variables on which X depends. For a variable V at the very top of the network, the value passed down by V will be a vector of the prior probabilities that V takes on its various values. Ultimately, BEL(x) should be a function of these prior probabilities and the fixed probabilities at nodes where the value of the variable is known, which produce BEL vectors such as (1 0 0).

Calculating BEL values is nontrivial, because it requires repeatedly updating BEL and other values until the prior probabilities and the known values based on the evidence have propagated throughout the network. It has been shown that the general problem of probabilistic inference in networks is NP-hard (Cooper 1990), so we should not expect there to be a universal efficient algorithm for updating BEL. Pearl presents algorithms for computing BEL in the special case where networks are singly connected, that is, where no more than one path exists between two nodes (Pearl 1988, chap. 4; see also Neapolitain 1990). A loop here is a sequence of edges independent of direction. If there is more than one path between nodes, then the network contains a loop that can interfere with achievement of stable values of BEL, λ, and π. Hence methods have been developed for converting multiply connected networks into singly connected ones by clustering nodes into new nodes with many values. For example, consider the

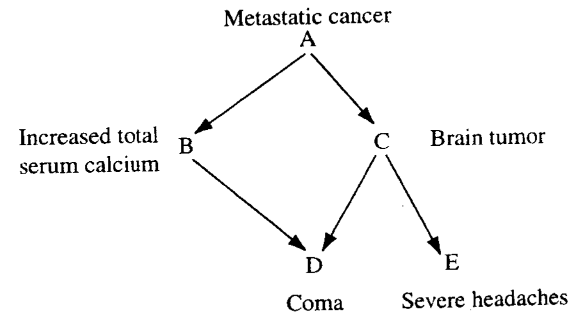

Figure 8.3
Pearl's representation of a multiply connected network that must be manipulated before probability calculations can be performed.

network shown in figure 8.3 (from Pearl 1988, 196). In this example, metastatic cancer is a cause of both increased total serum calcium and brain tumors, either of which can cause a coma. This is problematic for Pearl because there are two paths between A and D. Clustering involves collapsing nodes B and C into a new node Z representing a variable with values that are all the possible combinations of the values of B and C: increased calcium and tumor, increased calcium and no tumor, no increased calcium and tumor, and no increased calcium and no tumor. ECHO deals with cases such as these very differently, using the principle of competition, as we will see below.

There are other ways of dealing with loops in probabilistic networks besides clustering. Pearl discusses two approximating alternatives. Lauritzen and Spiegelharter (1988) have offered a powerful general method for converting any directed acyclic graph into a tree of cliques of that graph (see also Neapolitain 1990, chap. 7). Hrycej (1990) shows how approximation by stochastic simulation can be understood as sampling from the Gibbs distribution in a random Markov field. Frey (1998) uses graph-based inference techniques to develop new algorithms for Bayesian networks.

What probabilities must actually be known to compute BEL(x) even in singly connected networks? Consider again figure 8.2, where the BEL values for node D are to be computed from values for A, B, C, E, and F. To simplify, consider only values a, b, c, d, and e of the respective variables. Pearl's algorithms do not simply require knowledge of the conditional probabilities $P(d/a)$, $P(d/b)$, and $P(d/c)$. (Here $P(d/a)$ is shorthand for the probability that D has value d given that A has value a.) Rather, the calculation considers the probabilities of d, given all the possible combinations of values of the variables on which D depends. Consider the simple propositional case where the possible values of D are that it is true (d) or false (not d). Pearl's algorithm requires knowing $P(d/a$ & b & $c)$, $P(d/a$ & b & not $c)$, $P(d/a$ & not b & not $c)$ and five other conditional probabilities. The analogous conditional probabilities for not d can be computed from the ones just given.

More generally, if D depends on n variables with k values each, k^n conditional probabilities will be required for computation. This raises two problems that Pearl discusses. First, if n is large, the calculations become computationally intractable, so that approximation methods must be used. Probabilistic networks nevertheless are computationally much more attractive than the general problem of computing probabilities, since the threat of combinatorial explosion is localized to nodes that we can hope to be dependent on a relatively small number of other nodes. Second, even if n is not so large, there is the problem of obtaining sensible conditional probabilities to plug into the calculations. Pearl acknowledges that it is unreasonable to expect a human or other system to store all this information about conditional probabilities, and he shows

Table 8.1
Comparison of ECHO and Pearl networks

	ECHO	Pearl
Nodes represent	propositions	variables
Edges represent	coherence	dependencies
Directedness	symmetric	directed
Loops	many	must be eliminated
Node quantity updated	activation from −1 to 1	BEL: vector with values having probabilities from 0 to 1
Additional updating	none	λ, π
Additional information used	explanations, data	conditional probabilities, prior probabilities

how it is sometimes possible to use simplified models of particular kinds of causal interactions to avoid having to do many of the calculations that the algorithms would normally require. Table 8.1 summarizes the differences between ECHO and Pearl networks. We now have enough information to begin considering the relation between ECHO and Pearl networks.

3 TRANSLATING ECHO INTO PROBABILISTIC NETWORKS

To see why it is reasonable to consider translating ECHO networks into Pearl networks, let us review some simple examples that illustrate ECHO's capabilities. Given a choice between competing hypotheses, ECHO prefers ones that explain more. A Pearl network can be expected to show a similar preference, since the λ function will send more support up to a value v of a variable from variables whose values v_i are known and where $P(v/v_i)$ is high. ECHO prefers hypotheses that are explained to ones that are not, and a Pearl network will similarly send support down to a value of a variable using the π function. To determine whether Pearl networks can duplicate other aspects of ECHO networks, we will have to consider a possible translation in more detail.

A translation algorithm from ECHO networks to Pearl networks could take either of two forms. The most direct would be an immediate network-to-network translation. Every proposition node in the ECHO network would become a variable in a Pearl network, and every ECHO link would become a Pearl conditional probability between values of variables. This direct translation clearly fails, since ECHO's symmetric links would translate into two-way conditional probabilities, which are not allowed in Pearl networks. Moreover, the translation would produce many cycles, which Pearl networks exclude by definition. (Clarification: In Pearl's terminology, all cycles are loops, but not all loops are cycles. $A \to B \to C \to A$ is a cycle, because the direction of the path is maintained. However, $A \to B \to C \leftarrow A$ is a loop, since for loops direction is ignored, but it is not a cycle.) Alternatively, we can bypass the creation of an ECHO network and simply use the input to ECHO to generate a Pearl network directly. We can then try to produce a program that will take ECHO's input and produce an Pearl network suitable for running Pearl's algorithms.

Let us call this program PECHO. First we must worry about creating the appropriate nodes. ECHO creates nodes when it is given input describing a proposition P. Analogously, PECHO would create a variable node with two values, TRUE and FALSE. At this point, PECHO would have to consult ECHO's input concerning contradictions and check whether there is some proposition not P that

contradicts P. If so, there is no need to construct a new variable node, since not P would simply be represented by the FALSE value of the variable node where P represents the value TRUE. It becomes more complicated if there are several propositions in ECHO that all contradict each other, but these could all be amalgamated into one variable node with multiple values.

Since the vector representing the BEL values for a variable is required to sum to 1, PECHO will be able to duplicate ECHO's effect that the acceptability of a proposition counts against the acceptability of any proposition that contradicts it. Because we are not directly translating from ECHO networks to Pearl networks, we do not have to worry that ECHO activation values range from −1 to 1, rather than from 0 to 1 like probabilities, but in any case it is possible to normalize ECHO's activations into the range of probabilities. To normalize ECHO in the simplest case, just add 1 to a unit's activation and divide the result by 2. If two units represent contradictory propositions, normalize the resulting values by multiplying each value by 1 divided by the sum of the values.

When a proposition contradicts more than one other proposition, the normalization becomes problematic unless the propositions in question are all mutually contradictory, as when they all represent different values of the same variable. In ECHO networks this is not always the case. Simulation of Copernicus's case against Ptolemy (Nowak and Thagard 1992a) includes the following propositions:

P6 The Earth is always at the center of the heavenly sphere.

P12 The sun moves eastward along a circle about the earth in one year.

C12 The sun is immobile at the center of the universe.

Here Ptolemy's propositions P6 and P12 each contradict Copernicus's C12, but they do not contradict each other.

ECHO uses the range [−1, 1] for both conceptual and computational reasons. Conceptually, ECHO is interpreted in terms of degrees of acceptance (activation > 0) and degrees of rejection (activation < 0), sharing with some expert systems the intuition that attitudes toward hypotheses are better characterized in terms of acceptance versus rejection rather than as degrees of belief, as probabilists hold (see Buchanan and Shortliffe 1984, chap. 10). The computational reason is that activation updating in ECHO has the consequence that if a hypothesis coheres with one that is deactivated, it too will tend to be deactivated. Thus sets of hypotheses (theories) tend to be accepted and rejected as wholes, as is usually the case in the history of science (Thagard 1992b).

Now we get to the crucial question of links. When ECHO reads the input

(explain (P1 P2 P3 P4) Q)

it creates excitatory links between each of the explaining propositions and Q. PECHO correspondingly would note that the variable node whose TRUE value represents Q is causally dependent on the variable node whose TRUE value represents P1, and so on. More problematically, PECHO would have to contrive 4 conditional probabilities: P(Q/P1), P(Q/not P1), P(not Q/P1), and P(not Q/not P1). The first of these could perhaps be derived approximately from the weight that ECHO puts on the link between the nodes representing P1 and Q, and we could derive the third from the first, since P(Q/P1) and P(not Q/P1) sum to 1, but ECHO provides no guidance about the other two probabilities.

In fact, the situation is much worse, since Pearl's algorithms actually require 32 different conditional probabilities, for example, P(Q/P1 & not P2 & P3 & not P4). In the most complex ECHO network to date, modeling the case of Copernicus against Ptolemy (Nowak and Thagard 1992a), there are 143 propositions. A search through the units created by ECHO that counts the number of explainers shows that the number of conditional probabilities that PECHO would need is 45,348. This is a big improvement over the 2^{143} (more than 10^{43}) probabilities that a full distribution would require, but it is still daunting. Of the 45,348 conditional probabilities, only 469 could be directly derived from the weight on the ECHO link. Does it matter what these probabilities are? Perhaps PECHO could give them all a simple default value and still perform as well as ECHO. We will shortly see that in fact explanatory coherence requires some constraints on the conditional probabilities if PECHO is to duplicate ECHO's judgements.

A derivation of P(Q/P1) based on the ECHO link between Q and P1 would effectively implement the simplicity principle, E2 (c) from chapter 3, since it would mean that in updating the Pearl network, P1 would get less support from Q than it would if P1 explained Q without the help of other hypotheses. This is because ECHO makes the strength of such links inversely proportional to the number of hypotheses. PECHO is able to get by with a unidirectional link between the node for P and the node for Q, since the contribution of the λ and π functions to the BEL functions of the two nodes effectively spreads support in both directions.

The construction just described would enable PECHO to implement principles E2 (a) and E2 (c) of the theory of explanatory coherence, but what about E2 (b), according to which P1, P2, P3, and P4 all cohere with each other? Here PECHO encounters serious difficulties. Explanatory-coherence theory assumes that cohypotheses (hypotheses that participate together in an explanation) support each other, but this is impossible in Pearl networks, which have to be acyclic. There is thus a deep difference in the fundamental assumptions of explanatory coherence and probabilistic networks, which gain their relative efficiency by making strong assumptions of independence. In contrast, explanatory coherence assumes that every proposition in a belief system can be affected by every other one, although the effects may be indirect and small. To put it graph-theoretically, ECHO networks are strongly connected, by their symmetric links, but probabilistic networks with directed edges are emphatically not. At first glance, ECHO networks might seem to be similar to the noisy OR gates for which Pearl (1988, 188ff.) provides an efficient method. However, his method requires assumptions not appropriate for ECHO, such as that an event is presumed false if all conditions listed as its causes are false.

Pearl networks can, however, get the effects of excitatory links between cohypotheses by means of the clustering methods used to eliminate loops. We saw in the last section that Pearl considers collapsing competing nodes into single nodes with multiple values. If H1 and H2 together explain E, then instead of creating separate variable nodes for H1 and H2, PECHO would create a variable node ⟨H1-H2⟩ with values representing H1 & H2, H1 & not H2, not H1 & H2, and not H1 & not H2. To implement explanatory-coherence principle E2 (b), which establishes coherence between H1 and H2, PECHO would have to ensure that it has conditional probabilities such that P(E/H1 & H2) is greater than either P(E/H1 & not H2) or P(E/not H1 & H2). (I am simplifying the representation here: in the Pearl network, by P(E/H1 & H2) I mean the probability that variable E takes the value TRUE, given that variable ⟨H1-H2⟩ takes the value ⟨H1 & H2⟩.)

A similar method should enable PECHO to deal with the inhibitory links required by ECHO to implement explanatory coherence principle E6, Competition. We saw that PECHO can handle contradictions by constructing complex variables, but it has no direct way of expressing the negative impact of one hypothesis on another when they are competing to explain a piece of evidence. The clustering technique mentioned in the last section shows how this can be done. If an ECHO network has H1 and H2 independently explaining E, PECHO will have to replace variable nodes for H1 and H2 with a combined node with values for H1 & H2, H1 & not H2, not H1 & H2, and not H1 & not H2. PECHO can enforce competition between H1 and H2 by requiring that P(E/H1 & H2) be less than either P(E/H1 & not H2) or P(E/not H1 & H2). In very simple situations, it is possible to enforce competition in probabilistic networks without using clustering. Pearl describes how the effect of one cause *explaining away* another can be modeled in singly connected networks. Often in the examples to which ECHO has been applied, however, two hypotheses compete to explain more than one piece of evidence. Units representing those hypotheses are therefore connected by two different paths, and clustering or some other technique will be necessary to translate the network into one not multiply connected.

The clustering situation gets much more complicated, since H1 may compete with other hypotheses besides H2. In the Copernicus versus Ptolemy simulation, ECHO finds 214 pairs of competing hypotheses, and there is an important Copernican hypothesis that competes with more than 20 Ptolemaic hypotheses. In general, if a proposition P has n hypotheses participating in explaining it, either in cooperation or in competition with each other, then a clustered variable node with 2^n values will have to be created.

For example, in the Copernicus versus Ptolemy simulation, there are pieces of evidence explained by 5 Ptolemaic hypotheses working together and by 5 Copernican hypotheses. To handle both support among cohypotheses and competition between conflicting hypotheses, PECHO would need to have a single node with 1,024 values, in place of the 10 nodes corresponding to ECHO units. In fact, the node would have to be still more complicated because these 10 hypotheses compete and cohere with many others, since they participate in additional explanations. Typically, the set of hypotheses formed by collecting all those that either compete with or coexplain with some member of the set will be virtually all the hypotheses that there are. In the Copernicus simulation, there are 80 explaining hypotheses, including Ptolemaic ones, and a search shows that each is connected to every other by chains of coherence or competition. We would thus require a single hypothesis node with 2^{80} values, more than the number of milliseconds in a billion years.

Thus, dealing with competition and cohypotheses in Pearl networks can be combinatorially disastrous, although there may be more efficient methods of clustering. Pearl (1988, 201) considers an alternative method, akin to that of Lauritzen and Spiegelharter (1988), in which the nodes represent cliques in an ECHO network, such as the cluster of H1, H2, and E. (Cliques are subgraphs whose nodes are all adjacent to one another.) In the Copernicus simulation, there are more than 4,000 such cliques, so the reconstituted Pearl network would be much larger than the original. More important, it is not at all clear how to assign conditional probabilities in ways that yield the desired results concerning cohypotheses and competitors. Thus in principle ECHO input can be used to drive a Pearl network, but in practice the computational obstacles are formidable.

The input to ECHO also includes information about data and analogies. PECHO can implement something like explanatory-coherence principle E4, Data Priority, by special treatment of variable nodes corresponding to evidence propositions in ECHO. It would be a mistake to instantiate an evidence variable node with a value (1 0), since that would not allow the possibility that the evidence is mistaken. (ECHO can reject evidence if it does not fit with accepted hypotheses.) Pearl (1988, 170) provides a method of dummy variables, which allows a node to represent virtual evidence, an effective solution if updating can lead to low BEL values for such nodes. As for analogy, there is no direct way in a Bayesian network in which a hypothesis H1 can support an analogous one H2, but once again dummy nodes might be constructed to favor the hypothesis in question. Analogy is normally used to support a contested hypothesis by analogy with an established one, so little is lost if there is no symmetric link and the contested one is simply viewed as being slightly dependent on the established one. Analogy is thus viewed as contributing to the prior probability of the contested hypothesis.

In sum, the theory of explanatory coherence that is implemented in ECHO by connectionist networks (and by the other coherence algorithms in chapter 2) can also be approximately implemented by probabilistic networks. There is, however, a high computational cost associated with the alternative implementation. A massively greater amount of information in the form of conditional probabilities is needed to run the algorithms for updating probabilistic networks, and the problem of creating a probabilistic network is nontrivial: reconstruction is required to avoid loops, and care must be taken to retain information about cohypotheses and competitors. Combinatorial explosions must also be avoided.

Hence while ECHO's connectionist networks can be abstractly viewed in probabilistic terms, there are potentially great practical gains to be had by not abandoning the explanationist approach for the apparently more general probabilist one. Practically, the probabilist approach must use the explanationist one for guidance in assessing probabilities. We saw that consideration of cohypotheses and competitors puts constraints on the conditional probabilities allowable in probabilistic networks, and explanatory coherence theory can also contribute to setting prior probabilities. One can also think of the principles of analogy and data priority as giving advice on how to set prior probabilities. How far can we go with ECHO alone?

4 TACKLING PROBABILISTIC PROBLEMS WITH ECHO

If ECHO is to qualify as an alternative to probabilistic networks, it must be able to handle cases viewed as prototypical from the probabilist perspective. Consider Pearl's (1988, 49) example of Mr. Holmes at work trying to decide whether to rush home because his neighbor Mr. Watson, a practical joker, has called to say that his alarm at home has sounded. If the alarm has sounded, it may be because of a burglary or because of an earthquake. If he hears a radio report of an earthquake, his degree of confidence that there was a burglary will diminish. Appropriate input to ECHO would be the following:

(explain (BURGLARY) ALARM)

(explain (EARTHQUAKE) ALARM)

(explain (EARTHQUAKE) RADIO-REPORT)

(explain (ALARM) WATSON-CALLED)

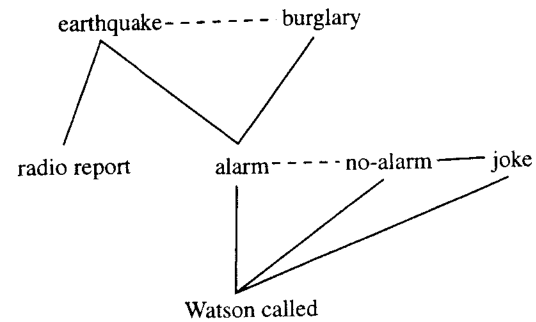

Figure 8.4
The ECHO network created for Pearl's burglary example for input given in the text. Solid lines indicate positive constraints, while dotted lines indicate negative ones.

(explain (JOKE NO-ALARM) WATSON-CALLED)

(contradict ALARM NO-ALARM)

(data (WATSON-CALLED RADIO-REPORT))

The network created by ECHO using this input is shown in figure 8.4. In implementing the competition principle E6, ECHO automatically places inhibitory links between BURGLARY and EARTHQUAKE and between ALARM and JOKE. From the above input, ECHO reaches the conclusion that there was an earthquake rather than a burglary.

This simple qualitative information may give misleading results in cases where statistical information is available. Suppose that Holmes knows that burglaries almost always set off his alarm, but earthquakes do so only rarely. ECHO need not assume that every explanation is equally good; it allows the input to include an indicator of the strength of the explanation. We could, for example, alter the above input to include these statements:

(explain (BURGLARY) ALARM 0.8)

(explain (EARTHQUAKE) ALARM 0.1)

This has the effect of making the excitatory link between BURGLARY and ALARM eight times stronger than the link between EARTHQUAKE and ALARM, so, other things being equal, ECHO will prefer the burglary hypothesis to the earthquake hypothesis.

Statistical information that provides prior probabilities can be used in similar ways. Suppose that an alarm is as likely if there is a burglary as if there is an earthquake, but Mr. Holmes knows that in his neighborhood burglaries are far more common than earthquakes. Without the radio report of the earthquake, Holmes should prefer the burglary hypothesis to the earthquake hypothesis. In Bayesian terms, the burglary base rate is higher. ECHO can implement consideration of such prior probabilities by assuming that the base rates provide a statistical explanation of the occurrences (see Harman 1986, 70). The base rates can be viewed as hypotheses that themselves explain statistical information that has been collected. We could thus have this additional input to ECHO:

(explain (BURGLARY-RATE) BURGLARY 0.1)
(explain (BURGLARY-RATE) BURGLARY-STATISTICS)
(explain (EARTHQUAKE-RATE) EARTHQUAKE-STATISTICS)
(explain (EARTHQUAKE-RATE) EARTHQUAKE 0.01)
(data (BURGLARY-STATISTICS EARTHQUAKE-STATISTICS
WATSON-CALLED RADIO-REPORT))

The network constructed by ECHO is shown in figure 8.5.

Similar cases in which prior probabilities need to be taken into account often arise in medical diagnosis. Medical students are cautioned to prefer routine diagnoses to exotic ones with the adage, "When you hear hoof beats, think horses, not zebras."

ECHO is also capable of handling the cancer example (figure 8.3) with the prior and conditional probabilities

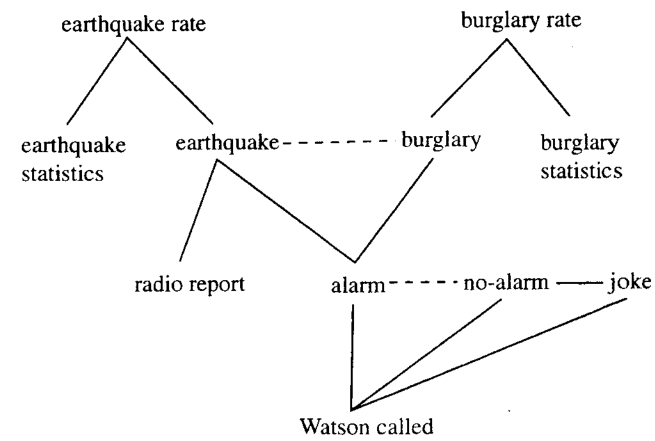

Figure 8.5
An enhanced ECHO network for the burglary example with statistical explanations.

provided by Pearl (1988, 197). Of course, ECHO's final activations are not exactly equivalent to the final probabilities that Pearl calculates, but without recourse to clustering methods, ECHO gets results that are qualitatively very similar. ECHO strongly accepts just the propositions to which Pearl's calculation gives high probability, and strongly rejects just the propositions to which he gives low probability.

ECHO is thus capable of using probabilistic information when it is available, but does not require it. There may well be cases in which a full probability distribution is known and ECHO can be shown to give a defective answer because activation adjustment does not exactly mirror the calculation of posterior probabilities. In such cases where there are few variables and the relevant probabilities are known, it is unnecessary to muddy the clear probabilistic waters with explanatory-coherence considerations. But in most real cases found in science, law, medicine, and ordinary life, the explanationist will not be open to the charge of being probabilistically incoherent, since the probabilities are sparsely available and calculating them is computationally unfeasible. What matters, then, are the qualitative considerations that explanatory coherence theory takes into account, and probabilities are at best epiphenomenal. See Thagard (1999) for further argument that explanatory coherence is crucial to causal reasoning in medicine.

It might be argued that probabilistic approaches are preferable because they provide a clear semantics for numerical assessment of hypotheses. While probability theory undoubtedly has a clear syntax, the meaning or meanings of probability is an unsolved problem. All the available interpretations, in terms of frequencies, propensities, degrees of belief, and possible worlds, can be challenged (for a review, see Cohen 1989). For scientific purposes, statistical inference based on frequencies in observed populations suffices, and we can dispense with the logically problematic and psychologically implausible notion of probabilities as degrees of belief. Frequency views of probability are difficult to apply to individual events such as "Fred has a brain tumor" and to causal hypotheses such as "Fred's headaches are caused by a brain tumor." Whereas probability theory is only a few hundred years old and requires expert calculations, people have been offering and evaluating explanations at least since the pre-Socratic philosophers. Moreover, explanatory reasoning is part of everyday life when people try to understand the behavior of the physical world and other people. Hence, instead of trying to contrive probabilistic accounts of reasoning where frequencies are not available, we should adopt the psychologically plausible and computationally efficient explanationist approach.

5 CONCLUSION

At the most general level, this chapter can be understood as offering a reconciliation of explanationism and probabilism. ECHO, the most detailed and comprehensive explanationist model to date, has a probabilistic interpretation. This interpretation should make the theory of explanatory coherence more respectable to probabilists, who should also appreciate how explanatory-coherence issues such as data priority, analogy, cohypotheses, and competition place constraints on probability values.

At a more local level, however, it is an open question whether explanationist or probabilist accounts are superior. Local disputes can be epistemological, psychological, or technological. If one accepts the view of Goldman (1986) that power and speed, as well as reliability, are epistemological goals, then explanationist models can be viewed as desirable ways of proceeding apace with causal inferences, while probabilistic models are still lost in computation. Similarly, the computational cost associated with the probabilistic interpretation of explanatory coherence suggests that such models may be inappropriate as models of human psychology. ECHO and probabilistic networks can be compared as models of human performance, with probabilistic networks apparently predicting that people should be much slower at inference tasks that require the most work to translate into probabilistic terms. We saw such cases arise when there are cohypotheses and competitors. ECHO takes such complications in stride, whereas Pearl networks require computations to realign networks and many more conditional probabilities to handle such cases. It should therefore be possible to present people with examples of increasing complexity and determine whether their reasoning ability declines rapidly, as the complexity of probabilistic computations suggest.

Similarly for technological applications in expert systems, ECHO may perform better than probabilistic networks. If rich probabilistic information is generally not available and if the domains are complex enough with cohypotheses and competitors, then ECHO may be more effective than probabilistic cases. The issue must be decided on a local basis, application by application, just as the psychological issue depends on experiments that have not yet been done. My conjecture is that the psychological and technological applicability of explanationist and probabilist techniques will vary from domain to domain with the following approximate ordering from most appropriate for explanationist to most appropriate for probabilist approaches: social reasoning, scientific reasoning, legal reasoning, medical diagnosis, fault diagnosis, games of chance. The psychological and technological answers need not be the same: diagnosis may well be an enterprise where a nonpsychological probabilistic approach can bring technological gains.

Much remains to be done in the comparative evaluation of the computational and psychological merits of the probabilistic and coherentist approaches to causal reasoning. Following Pearl's seminal 1988 book, there have been improvements in the computational implementation of Bayesian networks, but solutions have not been found for such fundamental problems as the need for many unavailable conditional probabilities and the lack of a frequency interpretation for individual events and causal hypotheses. Since the theory of explanatory coherence has a probabilistic approximation, albeit a computationally expensive one, the analysis in this chapter suggests that probabilism might reign supreme in the epistemology of Eternal Beings.

But explanationism survives in epistemology for the rest of us.

6 SUMMARY

Causal reasoning can be understood qualitatively in terms of explanatory coherence or quantitatively in terms of probability theory. Comparison of these approaches can be done most informatively by looking at computational models, using ECHO's coherence networks and Pearl's probabilistic ones. ECHO can be given a probabilistic interpretation, but there are many conceptual and computational problems that make it difficult to replace coherence networks with probabilistic ones. On the other hand, ECHO provides a psychologically plausible and computationally efficient model of some kinds of probabilistic causal reasoning. Hence coherence theory need not give way to probability theory as the basis for epistemology and decision making.

9 *The Future of Coherence*

One of the most attractive reasons for putting probability theory at the center of epistemology is that it ties belief closely with decision: combining probabilities with utilities allows us to calculate the expected utilities of different actions and choose the best. This book has presented an alternative approach in which inference concerning what to believe and what to do are both based on coherence. The mathematically exact, computationally feasible, and psychologically plausible account of coherence presented in chapter 2 provided the basis for understanding the development of ordinary, scientific, and metaphysical knowledge (chapters 3 and 4). Adding deliberative coherence into the picture provided a basis for understanding how people make decisions, including judgments about what is right and wrong (chapter 5). Human inference is a matter of emotion as well as cold cognition, and chapter 6 showed how a theory of emotional coherence can be constructed as an extension of the theory of coherence as constraint satisfaction, with applications to understanding diverse judgments ranging from trust to aesthetics. The development of knowledge is a social as well as a cognitive process, and chapter 7 described a theory of consensus based on coherence and communication. In chapter 8, I argued that causal reasoning in many domains is more naturally construed in terms of explanatory coherence than in terms of probability theory.

The results of these inquiries illustrate, I hope, the fecundity of cognitive naturalism, the approach to philosophy in which psychological theories and computational models are combined with philosophical reflection to produce theories of knowledge, reality, ethics, politics, and aesthetics. Cognitive naturalism does not abandon the traditional philosophical concern with epistemological and ethical justification, nor does it try to derive the normative from the descriptive. The aim, rather, is to interweave normative philosophical theories with empirical scientific ones so that they form a coherent whole. Connecting philosophy with empirical and computational investigations does not signal its demise, but rather opens up new possibilities for pursuing answers to its ancient and inescapable questions.

I do not pretend to have answered all these questions in this essay. Although the treatment of coherence in chapter 2 and later is far more comprehensive than previous discussions by philosophers and cognitive scientists, my application of coherence notions to problems in epistemology, metaphysics, ethics, political philosophy, and aesthetics has sometimes devoted only a few pages to important issues that deserve volumes. I have aimed for demonstration of the breadth of the idea of coherence as cognitive and emotional constraint satisfaction, at the expense of depth in many of the suggested applications. It is not circular reasoning to note that one of the great advantages of my version of coherentism is that is highly coherent, applying the same conception of coherence as constraint satisfaction to many diverse kinds of thinking.

Much remains to be done to work out the philosophical and psychological consequences of the hypothesis that a great deal of human thought consists of coherence

judgments that maximize constraint satisfaction. The remainder of this chapter suggests a series of research projects that would help to fill in the substantial gaps in the coherentist approach to cognition and philosophy that this book has merely sketched.

In ethics and epistemology, many philosophers have advocated the usefulness of Rawls's notion of reflective equilibrium. According to Elgin (1996, ix), "A system of thought is in reflective equilibrium when its components are reasonable in light of one another, and the account they comprise is reasonable in light of our antecedent convictions about the subject at hand." Elgin sees reflective equilibrium as an alternative to coherence, claiming that a system is coherent if its components mesh but that reflective equilibrium requires in addition reasonableness with respect to antecedent commitments. Obviously, the kind of coherence she rejects is very different from the discriminating and broad coherence that I advocated in chapter 4. In fact, you legitimately reach reflective equilibrium only if your system is maximally coherent, that is, if it maximizes satisfaction of multiple constraints, including ones involving evidence based on observation and experiment. Reflective equilibrium is an attractive metaphor for describing the end state of inquiry, but it depends on well-developed theories of coherence-based inference to provide an explanation of how equilibrium can and should be reached. Coherence as computational constraint satisfaction provides the overall framework for understanding reflective equilibrium in both epistemology and ethics, with specific theories of explanatory, deductive, conceptual, analogical, perceptual, and deliberative coherence providing the details concerning the elements and constraints involved. I agree with Stich (1988) that reflective equilibrium is an insufficient basis for a theory of epistemological and ethical justification.

Still, it would be desirable to have a fuller account of how coherence-based inference dynamically produces reflective equilibrium. The examples discussed in chapter 3 and 4 presume that an individual is presented all at once with an array of elements and coherence relations, with maximization of constraint satisfaction proceeding in a single step. More realistically, people's beliefs develop incrementally, with equilibrium being achieved in smaller steps than one global coherence calculation (Hoadley, Ranney, and Schank 1994). Studying this process psychologically and computationally should provide a better understanding of how people can reach reflective equilibria that are optimal in that they maximize the coherence of all available information, and also a better understanding of how people sometimes reach equilibria that are suboptimal.

A psychologically realistic theory of coherence-based inference should also have practical applications to help people reason better. I often teach an undergraduate class on critical thinking, and do so within the cognitive naturalist framework presented in this book. Most critical-thinking textbooks assume, in line with philosophical orthodoxy, that human inference is and should be based on arguments, with deduction providing the gold standard of what an argument should look like. Although arguments are important for indicating the elements and constraints relevant to making an inference, they give a misleadingly linear picture of how inferences are actually made. If inference is coherence-based, with emotional as well as cognitive constraints contributing, it becomes much easier to see why people are so frequently prone to inferential errors that have nothing to do with deduction. The standard philosophical list of fallacies does not begin to capture the array of common reasoning errors that psychologists have identified (e.g., Gilovich 1991). Cognitive

naturalism can draw on research concerning the psychological processes that can lead people to think poorly, while at the same time urging reasoning strategies that encourage assembly of all the information that people need to maximize explanatory and other kinds of coherence. Ranney and Schank (1998) describe an educational program, Convince Me, that uses explanatory-coherence computations to help students develop and revise arguments, but working out how the coherentist understanding of reason and emotion can be used systematically to produce a new approach to critical thinking is a task that remains to be done. It is also possible to derive insights into how people can improve their decision making by drawing lessons from the theories of deliberative and emotional coherence (Thagard, forthcoming). In chapter 5, I rejected the common philosophical view that intuitions contribute to ethical justification, but intuitions can be a valuable part of individual decision making when they provide an emotional summary of tacit judgments about what is most important to a person.

The metaphysical applications of coherence theory also need to be much further developed. I hope, for example, that someone with an interest in theology will work out in much greater detail the explanatory and analogical structure of the case for and against the existence of God. However, I suspect that further analysis along these lines would only account for the attitudes of small numbers of religious believers, with many more asserting that their beliefs rest on faith rather than reason. I would like to see the development of a theory of faith as a kind of emotional coherence, in which belief in God is adopted because of its contribution to satisfaction of personal and social goals that are important to many people. This theory would not provide any further justification of theistic beliefs, but would be valuable for solving the psychological puzzle of why so many people believe in God despite the paucity of good evidence.

For more philosophical purposes, it would also he highly desirable to say more about the connection between coherence and truth. Millgram (2000) raises doubts about whether the constraint-satisfaction characterization of coherence is fully adequate for philosophical purposes. His main objection is that it is not appropriate for epistemology, because it provides no guarantee that the most coherent available theory will be true. In a forthcoming reply, I argue that the constraint-satisfaction account of coherence is not at all flawed in the ways that Millgram describes and in fact satisfies the philosophical, computational, and psychological prerequisites for the development of epistemological and ethical theories (see http://cogsci.uwaterloo.ca/Articles/Pages/coh.price.html). Nevertheless, I would like to see a much fuller account of the conditions under which progressively coherent theories can be said to approximate the truth.

Chapter 5 barely begins the discussion of the applicability of coherentist ideas to ethics and politics. The topics discussed in that chapter—capital punishment, abortion, and the justification of the state—need to receive a much fuller treatment to bring out many more of the elements and constraints that are relevant to reaching conclusions by coherence maximization. Moreover, there are many other ethical and political issues that deserve a full discussion from the perspective of coherence as constraint satisfaction. At the methodological level, more thorough critical analysis is needed of the appropriate contribution of ethical thought experiments to analogical and general coherence. If I am right that political decisions primarily are and should be based on maximizing the three constraints of freedom, flourishing, and fairness, then much more needs to be said about how we can assess the relative importance and appropriate tradeoffs of these constraints.

The theory of emotional coherence developed in chapter 6 is limited by its emphasis on positive and negative valences, and needs to be expanded to take into account the full range of human emotions. Our understanding of the cognitive neuroscience of emotions is increasing rapidly, and I hope that the theory of emotional coherence will expand to take these developments into account. Like other artificial-neural-network models used in cognitive science, my computational models of coherence are enormously simplified in comparison with the complexity of the brain and its neurons. In recent years, dramatic progress has been made in understanding the brain structures and mechanisms involved in emotions (e.g., Panksepp 1998). I plan to make my computational models more neurologically realistic by introducing distributed representations and more complex structures corresponding to brain anatomy.

The current version of HOTCO uses localist representations in which each unit (neuronlike node) represents a whole concept or proposition. Obviously, the brain does not have a single node for representations such as *Clinton*, but somehow distributes the information across numerous neurons. In current work on artificial neural networks, there are two main ways of distributing complex information across multiple nodes: vector coding and neural synchrony. Vector coding represents a complex piece of information such as a proposition by a vector of k real numbers, corresponding to the firing rates of k neurons. Encoding and decoding schemes have been devised to perform variable binding and thus distinguish the proposition *Clinton loves Hillary* from *Hillary loves Clinton*, a distinction that was not possible in early artificial neural networks with simple nodes (Smolensky 1990). Eliasmith and Thagard (forthcoming) employ vector coding to produce distributed representation of complex relational propositions used in analogical mapping.

Within vector-coding schemes, the natural way to attach emotional valences to representations is to treat them as vectors that are algebraically blended with the vector that represents the proposition. Just as the vector representing *Clinton loves Hillary* is built by combining vectors for *Clinton*, *loves*, and *Hillary*, an enhanced vector could combine the proposition vector with an emotion vector representing the emotional attitude toward the proposition. In contrast to HOTCO, which can only associate positive and negative valences with nodes, using vectors to encode emotions would make possible the association of many different emotions with a proposition or other representation. The positive or negative emotions associated with *Clinton*, for example, could include liking, disliking, admiration, disgust, and so on.

The other main method for producing complex distributed representations in artificial neural networks is neural synchrony, which uses time as an additional means of binding information together (e.g., Hummel and Holyoak 1997). Representations such as *Clinton*, *Hillary*, *loves*, *agent*, and *recipient* are each represented by groups of artificial neurons with their own firing patterns, and relations between the representations are modeled by synchronies among those firing patterns, with neurons for related representations all firing or all not firing. Within this system, an emotion could be represented by a group of neurons that fire in synchrony with the neurons corresponding to the object of the emotion. It would be desirable to produce both neural-synchrony and vector-coding models of emotional inference in order to determine which is a more psychologically and neurologically plausible way of combining emotions with distributed representations.

Another way in which artificial-neural-network models such as HOTCO are not neurologically realistic is that they have few neuronal units and lack the high degree of anatomical organization found in the brain. As chapter 6 reported, Damasio and his colleagues have identified regions of the human brain whose damage consistently compromises processes involving connections between reasoning and emotion, which leads to defective reasoning in the personal and social domains (Damasio 1994; Damasio, Damasio, and Christen 1996). The crucial regions include the ventromedial prefrontal cortices, the amygdala, and the somatasensory cortices in the right hemisphere. Contrary to the popular view that emotions interfere with rational thought, Damasio and his colleagues have found that in patients with damage to these regions, the inability to integrate emotional considerations with cognitive planning actually produces inferior decisions. I hope to model the importance of these regions by organizing the units in my artificial neural networks in much more modular fashion.

Another promising area for research is the role of emotions in scientific thinking. Scientists are supposed to be dispassionate, but scientific cognition is often highly emotional. Here is a passage from James Watson's *Double Helix*, describing work leading up to the discovery of the structure of DNA; I have highlighted in boldface the positive emotion words and in italics the negative emotion words.

As the clock went past midnight I was becoming more and more **pleased**. There had been far too many days when Frances and I *worried* that DNA structure might turn out to be superficially very *dull*, suggesting nothing about either its replication or its function in controlling cell biochemistry. But now, to my **delight** and **amazement**, the answer was turning out to be profoundly **interesting**. For over two hours I **happily**

> lay awake with pairs of adenine residues whirling in front of my closed eyes. Only for brief moments did the *fear* shoot through me that an idea this good could be wrong. (Watson 1969, 118)

Watson's short book contains hundreds of such emotional expressions. Positive emotions involved in mental states such as interest, wonder, and excitement contribute to the pursuit of potentially important scientific ideas, while negative emotions involved in boredom, worry, and fear help to steer scientists away from unpromising pursuits. I hope to extend my theory of emotional coherence and link it to previous computational work on scientific discovery, producing a theory of the role of emotions as inputs and outputs of scientific discoveries.

In addition to helping to motivate problem solving and discovery, emotions attend the evaluation of scientific theories: highly coherent theories are viewed as elegant and beautiful, while ad hoc theories are rejected as ugly. My theory of emotional coherence can be extended to model the positive aesthetic feelings that attend the adoption of a highly coherent theory, as well as the negative feelings involved in the entertainment of unsatisfactory ones. Scientists' decisions to pursue answers to some questions rather than others seem based in part on emotional reactions such as surprise and excitement. I am more interested in the aesthetics of science than in the aesthetics of art, literature, or music, but I hope that philosophers and psychologists more inclined towards those areas will expand on my sketchy account of the role of emotional coherence in aesthetics.

A long-term objective for future work on emotional coherence would be to develop a theory of emotional change. The theory and computational model of emotional coherence are intended to explain why people make the emotional judgments that they do, but the theory

and model do not address the question of how such judgments can change over time. Emotion changes can include minor alterations in attitudes (e.g., "I used to like football, but I don't anymore") to major emotional shifts such as occur when people fall in love, turn their lives around through psychotherapy, or undergo religious or political conversions. It should be possible to build onto the theory of emotional coherence to develop a comprehensive theory of the cognitive and affective mechanisms that underlie emotional change, analogous to the theory of conceptual change that I developed to explain scientific revolutions (Thagard 1992b, 1999). We need to be able to answer questions such as the following: How are emotional constraints formed? How do elements acquire new input valences? How do changes in the valences of some elements contribute to dramatic shifts in attitudes towards persons and situations? Computational answers to these questions should help generate a model of both minor and major emotional changes. There is a substantial literature in social psychology on variables that affect attitude changes, but there is very little work on the cognitive-emotional mechanisms that produce such changes.

As chapter 7 stated, my model of consensus as communication plus coherence is highly idealized, and the development of more realistic models would shed further light on how consensus is achieved in science and other enterprises. As I indicated at the end of chapter 8, there is a need for further comparative evaluation of the computational and psychological merits of probabilistic and coherentist approaches to causal reasoning. I would like to see expanded computational experiments in which large ECHO networks are reinterpreted as Bayesian networks and simulated using one of the various programs now available for computing probabilistically (see, for example, the HUGIN system at http://www.hugin.dk/). Such experiments should be done in numerous domains, such as scientific, medical, and legal reasoning. For example, it would be desirable to construct a very large analysis of a legal trial, comparable to that performed by Wigmore (1937), and to determine the comparative feasibility of implementing the causal relations essential to legal inferences in Bayesian networks and the explanatory coherence program ECHO.

I have outlined these projects to indicate that there is much to be done on the coherentist research project in cognitive science, in synchrony with the philosophical movement of cognitive naturalism. Philosophy can go beyond analyzing concepts, conducting a priori investigations, and studying the great philosophers of the past. Borrowing ideas and methods from psychology and other sciences, it can help to develop robust theories of how people do and should think. Computational modeling provides a valuable methodology for working out and testing the feasibility of different theories of how people can increase their empirical and ethical knowledge. The coherentist approach, working within the theory of coherence as constraint satisfaction, is psychologically realistic and computationally feasible, yet it can contribute to the traditional goal of philosophy to be prescriptive as well as descriptive of human thought and action. Philosophy and cognitive science can thrive together in the twenty-first century.

References

Achinstein, P. (1991). *Particles and waves*. Oxford: Oxford University Press.

Allen, R. J. (1994). Factual ambiguity and a theory of evidence. *Northwestern University Law Review* 88: 604–660.

Anderson, N. (1974). Information integration theory: a brief survey. In D. H. Krantz, R. C. Atkinson, R. D. Luce, and P. Suppes (eds.), *Contemporary developments in mathematical psychology* (vol. 2, pp. 236–305). San Francisco: W. H. Freeman.

Anonymous. (1996). *Primary colors*. New York: Random House.

Aristotle. (1984). *The complete works of Aristotle*. Princeton: Princeton University Press.

Arrow, K. J. (1963). *Social choice and individual values*. Second ed. New York: Wiley.

Ash, M. G. (1995). *Gestalt psychology in German culture, 1890–1967*. Cambridge: Cambridge University Press.

Audi, R. (1993). Fallibilist foundationalism and holistic coherentism. In L. P. Pojman (ed.), *The theory of knowledge: classic and contemporary readings* (pp. 263–279). Belmont, Calif.: Wadsworth.

Bacchus, F., and van Beek, P. (1998). On the conversion between non-binary and binary constraint satisfaction problems. *Proceedings of the National Conference on Artificial Intelligence (AAAI-98)* (pp. 311–318). Menlo Park, Calif.: AAAI Press.

Baird, R. M., and Rosenbaum, S. E. (eds.), (1993). *The ethics of abortion*. Buffalo: Prometheus Books.

Baker, G. L., and Gollub, J. P. (1990). *Chaotic dynamics: an introduction*. Cambridge: Cambridge University Press.

Barnes, A. (1998). *Reading other minds*. Unpublished Ph.D. thesis, University of Waterloo, Waterloo, Ontario.

Barnes, A., and Thagard, P. (1997). Empathy and analogy. *Dialogue: Canadian Philosophical Review* 36: 705–720.

Batson, C. D., Sympson, S. C., Hindman, J. L., Decruz, P., Todd, R. M., Weeks, J. L., Jennings, G., and Burris, C. T. (1996). "I've been there, too": effect on empathy of prior experience with a need. *Personality and Social Psychology Bulletin* 22: 474–482.

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. New York: International Universities Press.

Beck, A. T., Rush, A. J., Shaw, B. F., and Emery, G. (1979). *Cognitive therapy of depression*. New York: Guilford.

Bender, J. W. (ed.), (1989). *The current state of the coherence theory*. Dordrecht: Kluwer.

Bianco, W. T. (1994). *Trust: representatives and constituents*. Ann Arbor: University of Michigan Press.

Blake, R. (1960). Theory of hypothesis among renaissance astronomers. In R. Blake, C. Ducasse, and E. H. Madden (eds.), *Theories of scientific method* (pp. 22–49). Seattle: University of Washington Press.

Blanchette, I., and Dunbar, K. (1997). Constraints underlying analogy use in a real-world context: politics. In M. G. Shafto, and P. Langley (eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 867). Mahwah, N.J.: Erlbaum.

Blanshard, B. (1939). *The nature of thought*. Vol. 2. London: George Allen and Unwin.

Bloor, D. (1991). *Knowledge and social imagery*. Second ed. Chicago: University of Chicago Press.

BonJour, L. (1985). *The structure of empirical knowledge*. Cambridge: Harvard University Press.

Bosanquet, B. (1920). *Implication and linear inference*. London: Macmillan.

Bower, G. H. (1981). Mood and memory. *American Psychologist* 36: 129–148.

Bower, G. H. (1991). Mood congruity of social judgments. In J. P. Forgas (ed.), *Emotion and social judgments* (pp. 31–53). Oxford: Pergamon Press.

Bradley, D. R., and Petry, H. M. (1977). Organizational determinants of subjective contour: the subjective Necker cube. *American Journal of Psychology* 90: 253–262.

Bradley, F. H. (1914). *Essays on truth and reality*. Oxford: Clarendon Press.

Brink, D. O. (1989). *Moral realism and the foundations of ethics*. Cambridge: Cambridge University Press.

Brush, S. G. (1996). *Fruitful encounters: the origin of the solar system and of the moon from Chamberlin to Apollo*. Vol. 3 of *A history of modern planetary physics*. Cambridge: Cambridge University Press.

Buchanan, B., and Shortliffe, E. (eds.), (1984). *Rule-based expert systems*. Reading, Mass.: Addison Wesley.

Byrne, M. D. (1995). The convergence of explanatory coherence and the story model: a case study in juror decision. In J. D. Moore, and J. F. Lehman (eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 539–543). Mahwah, N.J.: Erlbaum.

Caputi, M. (1996). National identity in contemporary theory. *Political psychology* 17: 683–694.

Carnap, R. (1950). *Logical foundations of probability*. Chicago: University of Chicago Press.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Clarendon Press.

Chalmers, D. J. (1996). *The conscious mind*. Oxford: Oxford University Press.

Charniak, E. (1993). *Statistical language learning*. Cambridge: MIT Press.

Churchland, P. M. (1995). *The engine of reason: the seat of the soul*. Cambridge: MIT Press.

Churchland, P. S. (1986). *Neurophilosophy*. Cambridge: MIT Press.

Cohen, L. J. (1977). *The probable and the provable*. Oxford: Clarendon Press.

Cohen, L. J. (1989). *An introduction to the philosophy of induction and probability*. Oxford: Clarendon.

Collingwood, R. G. (1997). *Outlines of a philosophy of art*. Bristol: Thoemmes Press.

Cooper, G. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence* 42: 393–405.

Cooper, J., and Fazio, R. H. (1984). A new look at dissonance theory. In L. Berkowitz (ed.), *Advances in experimental social psychology* (vol. 17). New York: Academic Press.

Cottrell, G. W. (1988). A model of lexical acces of ambiguous words. In S. L. Small, G. W. Cottrell, and M. K. Tanenhaus (eds.), *Lexical ambiguity resolution* (pp. 179–194). San Mateo: Morgan Kaufman.

Crick, F. (1994). *The astonishing hypothesis: the scientific search for the soul.* London: Simon and Schuster.

Cummins, R. (1998). Reflection on reflective equilibrium. In M. R. DePaul, and W. Ramsey (eds.), *Rethinking intuition* (pp. 113–127). Lanham: Rowman and Littlefield.

D'Ambrosio, B. (1999). Inference in Bayesian networks. *AI Magazine* 20 (no. 2, Summer): 21–36.

Damasio, A. R. (1994). *Descartes' error.* New York: G. P. Putnam's Sons.

Damasio, A. R., Damasio, H., and Christen, Y. (eds.), (1996). *Neurobiology of decision making.* Berlin: Springer-Verlag.

Daniels, N. (1979). Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy* 76: 256–282.

Daniels, N. (1996). *Justice and justification: reflective equilibrium in theory and practice.* Cambridge: Cambridge University Press.

Davidson, D. (1986). A coherence theory of truth and knowledge. In E. Lepore (ed.), *Truth and interpretation.* Oxford: Basil Blackwell.

Davies, P., and Brown, J. (1988). *Superstrings.* Cambridge: Cambridge University Press.

De Sousa, R. (1988). *The rationality of emotion.* Cambridge: MIT Press.

DeGeorge, R. (1990). Ethics and coherence. *Proceedings and Addresses of the American Philosophical Association* 64 (no. 3): 39–52.

DeMarco, J. P. (1994). *A coherence theory in ethics.* Amsterdam: Rodopi.

Dennett, D. (1991). *Consciousness explained.* Boston: Little, Brown.

Derbyshire, J. D., and Derbyshire, I. (1996). *Political systems of the world.* New York: St. Martin's Press.

Deutsch, M. (1973). *The resolution of conflict.* New Haven: Yale University Press.

Dunn, J. (1993). Trust. In R. E. Goodin, and P. Pettit (eds.), *A companion to contemporary political philosophy* (pp. 638–644). Oxford: Blackwell.

Elgin, C. Z. (1996). *Considered judgment.* Princeton: Princeton University Press.

Eliasmith, C., and Thagard, P. (1997). Waves, particles, and explanatory coherence. *British Journal for the Philosophy of Science* 48: 1–19.

Ellis, A. (1962). *Reason and emotion in psychotherapy.* New York: Lyle Stuart.

Ellis, A. (1971). *Growth through reason.* Palo Alto: Science and Behavior Books.

Ellis, R. E. (1992). *Coherence and verification in ethics.* Lanham, Md.: University Press of America.

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

Feldman, J. A. (1981). A connectionist model of visual memory. In G. E. Hinton, and J. A. Anderson (eds.), *Parallel models of associative memory* (pp. 49–81). Hillsdale, N.J.: Erlbaum.

Fenno, R. F. (1978). *Home style: house members in their districts.* Boston: Little, Brown.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford: Stanford University Press.

Fishbein, M., and Ajzen, I. (1975). *Belief, attitude, intention, and behavior.* Reading, Mass.: Addison-Wesley.

Fiske, S., and Pavelchak, M. (1986). Category-based vs. piecemeal-based affective responses: developments in schema-triggered affect. In R. Sorrentino, and E. Higgins (eds.), *Handbook of motivation and cognition* (vol. 1, pp. 167–203). New York: Guilford.

Flanagan, O. (1996). Ethics naturalized: ethics as human ecology. In L. May, M. Friedman, and A. Clark (eds.), *Mind and morals: essays on ethics and cognitive science* (pp. 19–44). Cambridge: MIT Press.

Frank, R. H. (1988). *Passions within reason.* New York: Norton.

Frege, G. (1964). *The basic laws of arithmetic.* Trans. by M. Furth. Berkeley: University of California Press.

Frey, B. J. (1998). *Graphical models for machine learning and digital communication.* Cambridge: MIT Press.

Frijda, N. H. (1993). Moods, emotion episodes, and emotions. In M. Lewis, and J. M. Haviland (eds.), *Handbook of emotions* (pp. 381–403). New York: Guilford.

Frith, U. (1989). *Autism: explaining the enigma.* Oxford: Basil Blackwell.

Frith, U., and Snowling, M. (1983). Reading for meaning and reading for sound in autistic and dyslexic children. *British Journal of Developmental Psychology* 1: 329–342.

Fukuyama, F. (1995). *Trust: social virtues and the creation of prosperity.* New York: Free Press.

Gambetta, D. (ed.), (1988). *Trust: making and breaking cooperative relations.* Oxford: Basil Blackwell.

Gardner, H. (1985). *The mind's new science.* New York: Basic Books.

Garey, M., and Johnson, D. (1979). *Computers and intractability.* New York: Freeman.

Gibbard, A. (1990). *Wise choices, apt feelings.* Cambridge: Harvard University Press.

Giere, R. (1988). *Explaining science: a cognitive approach.* Chicago: University of Chicago Press.

Giere, R. N. (1999). *Science without laws.* Chicago: University of Chicago Press.

Gilovich, T. (1991). *How we know what isn't so.* New York: Free Press.

Glynn, P. (1997). *God: the evidence.* Rocklin, Calif.: Prima Publishing.

Goemans, M. X., and Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the Association for Computing Machinery* 42: 1115–1145.

Goldman, A. I. (1986). *Epistemology and cognition.* Cambridge: Harvard University Press.

Goldman, A. I. (1992). *Liaisons: philosophy meets the cognitive and social sciences.* Cambridge: MIT Press.

Goodman, N. (1965). *Fact, fiction, and forecast.* Second ed. Indianapolis: Bobbs-Merrill.

Group for the Advancement of Psychiatry (GAP) (1987). *Us and them: the psychology of ethnonationalism.* New York: Brunner/Mazel.

Gwartney, J., and Lawson, R. (1997). *Economic freedom in the world, 1997.* Vancouver: Fraser Institute.

Gwartney, J., and Lawson, R. (1998). *Economic freedom in the world: 1998/1999 interim report.* Vancouver: Fraser Institute.

Haack, S. (1993). *Evidence and inquiry: towards reconstruction in epistemology.* Oxford: Blackwell.

Hacking, I. (1975). *The emergence of probability.* Cambridge: Cambridge University Press.

Hardwig, J. (1991). The role of trust in knowledge. *Journal of Philosophy* 88: 693–708.

Hardy, G. H. (1967). *A mathematician's apology.* Cambridge: Cambridge University Press.

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

Harman, G. (1986). *Change in view: principles of reasoning.* Cambridge: MIT Press.

Hartmann, W. K., Phillips, R. J., and Taylor, G. J. (eds.), (1986). *Origin of the moon.* Houston: Lunar and Planetary Institute.

Hegel, G. (1967). *The phenomenology of mind.* Trans. by J. Baillie. New York: Harper and Row. Originally published in 1807.

Heider, U. (1994). *Anarchism: left, right, and green.* San Francisco: City Light Books.

Hesse, M. (1974). *The structure of scientific inference.* Berkeley: University of California Press.

Hoadley, C. M., Ranney, M., and Schank, P. (1994). Wander ECHO: a connectionist simulation of limited coherence. In A. Ram, and K. Eiselt (eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 421–426). Hillsdale, N.J.: Erlbaum.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., and Thagard, P. R. (1986). *Induction: processes of inference, learning, and discovery.* Cambridge: MIT Press.

Holmes, J. G. (1991). Trust and the appraisal process in close relationships. In W. H. Jones, and D. Perlman (eds.), *Advances in personal relationships* (vol. 2, pp. 57–104). London: Jessica Kingsley.

Holyoak, K. J., and Spellman, B. A. (1993). Thinking. *Annual Review of Psychology* 44: 265–315.

Holyoak, K. J., and Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science* 13: 295–355.

Holyoak, K. J., and Thagard, P. (1995). *Mental leaps: analogy in creative thought*. Cambridge: MIT Press.

Holyoak, K. J., and Thagard, P. (1997). The analogical mind. *American Psychologist* 52: 35–44.

Horwich, P. (1982). *Probability and evidence*. Cambridge: Cambridge University Press.

Howson, C., and Urbach, P. (1989). *Scientific reasoning: the Bayesian tradition*. Lasalle, Ill.: Open Court.

Hrycej, T. (1990). Gibbs sampling in Bayesian networks. *Artificial Intelligence* 46: 351–363.

Hummel, J. E., and Biederman, I. (1997). Dynamic binding in a neural network for shape recognition. *Psychological Review* 104: 427–466.

Hurley, S. L. (1989). *Natural reasons: personality and polity*. New York: Oxford University Press.

Husserl, E. (1962). *Ideas: general introduction to pure phenomenology*. Trans. by W. R. B. Gibson. New York: Collier.

Hutcheson, F. (1973). *Francis Hutcheson: an inquiry concerning beauty, order, harmony, design*. The Hague: M. Nijhoff.

Ignatieff, M. (1993). *Blood and belonging: journeys into the new nationalism*. Toronto: Viking.

Jeffrey, R. (1983). *The logic of decision*. Second ed. Chicago: University of Chicago Press. First published in 1965.

Johnson, M. L. (1993). *Moral imagination: implications of cognitive science for ethics*. Chicago: University of Chicago Press.

Johnson, M. L. (1996). How moral psychology changes moral theory. In L. May, M. Friedman, and A. Clark (eds.), *Mind and morals: essays on ethics and cognitive science* (pp. 45–68). Cambridge: MIT Press.

Jordan, M. I. (ed.), (1998). *Learning in graphical models*. Dordrecht: Kluwer.

Josephson, J. R., and Josephson, S. G. (eds.), (1994). *Abductive inference: computation, philosophy, technology*. Cambridge: Cambridge University Press.

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

Kaplan, M. (1996). *Decision theory as philosophy*. Cambridge: Cambridge University Press.

Kecmanovic, D. (1996). *The mass psychology of ethnonationalism*. New York: Plenum.

Keith-Spiegel, P. (1972). Early conceptions of humor: varieties and issues. In J. H. Goldstein, and P. E. McGhee (eds.), *The psychology of humor* (pp. 3–39). New York: Academic Press.

Keynes, J. M. (1921). *A treatise on probability*. London: Macmillan.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review* 95: 163–182.

Kintsch, W. (1998). *Comprehension: a paradigm for cognition*. Cambridge: Cambridge University Press.

Kitcher, P. (1983). *The nature of mathematical knowledge*. New York: Oxford University Press.

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

Koffka, K. (1935). *Principles of gestalt psychology*. New York: Harcourt Brace.

Kosslyn, S. M. (1994). *Image and brain: the resolution of the imagery debate*. Cambridge: MIT Press.

Kramer, R. M., and Tyler, T. R. (eds.), (1996). *Trust in organizations*. Thousand Oaks, Calif.: Sage.

Kunda, Z. (1987). Motivation and inference: self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology* 53: 636–647.

Kunda, Z. (1990). The case for motivated inference. *Psychological Bulletin* 108: 480–498.

Kunda, Z., Miller, D., and Claire, T. (1990). Combining social concepts: the role of causal reasoning. *Cognitive Science* 14: 551–577.

Kunda, Z., and Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology* 68: 565–579.

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

Kusch, M. (1995). *Psychologism*. London: Routledge.

Kyburg, H. (1983). *Epistemology and inference*. Minneapolis: University of Minnesota Press.

Lakoff, G. (1996). *Moral politics: what conservatives know that liberals don't*. Chicago: University of Chicago Press.

Larson, D. W. (1997). Trust and missed opportunities in international relations. *Political Psychology* 18: 701–734.

Latouche, D. (1990). Betrayal and indignation on the Canadian trail. In P. Resnick (ed.), *Letters to a Québécois friend* (pp. 85–119). Montreal: McGill-Queen's University Press.

Latour, B., and Woolgar, S. (1986). *Laboratory life: the construction of scientific facts*. Princeton, N.J.: Princeton University Press.

Laudan, L. (1981). *Science and hypothesis*. Dordrecht: Reidel.

Lauritzen, S., and Spiegelharter, D. (1988). Local computation with probabilities in graphical structures and their applications to expert systems. *Journal of the Royal Statistical Society* B 50: 157–224.

LeDoux, J. (1996). *The emotional brain*. New York: Simon and Schuster.

Lefcourt, H. M., and Martin, R. A. (1986). *Humor and life stress: antidote to adversity*. New York: Springer-Verlag.

Lehrer, K. (1990). *Theory of knowledge*. Boulder: Westview.

Lehrer, K., and Wagner, C. (1981). *Rational consensus in science and society*. Dordrecht: Reidel.

Lempert, R. (1986). The new evidence scholarship: analyzing the process of proof. *Boston University Law Review* 66: 439–477.

Lévesque, R. (1968). *An option for Quebec*. Toronto: McClelland and Stewart.

Levi, I. (1980). *The enterprise of knowledge*. Cambridge: MIT Press.

Lewis, J. D., and Weigert, A. (1985). Trust as a social reality. *Social Forces* 63: 967–985.

Lipton, P. (1991). *Inference to the best explanation*. London: Routledge.

Lodge, M., and Stroh, P. (1993). Inside the mental voting booth: an impression-driven process model of candidate evaluation. In S. Iyengar, and W. J. McGuire (eds.), *Explorations in political psychology* (pp. 225–295). Durham: Duke University Press.

Lycan, W. (1988). *Judgement and justification*. Cambridge: Cambridge University Press.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review* 101: 676–703.

Maher, P. (1993). *Betting on theories*. Cambridge: Cambridge University Press.

Marr, D., and Poggio, T. (1976). Cooperative computation of stereo disparity. *Science* 194: 283–287.

May, L., Friedman, M., and Clark, A. (eds.), (1996). *Mind and morals: essays on ethics and cognitive science*. Cambridge: MIT Press.

McAllister, J. W. (1996). *Beauty and revolution in science*. Ithaca: Cornell University Press.

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception. Part 1: An account of basic findings. *Psychological Review* 88: 375–407.

McClelland, J. L., and Rumelhart, D. E. (1989). *Explorations in parallel distributed processing*. Cambridge: MIT Press.

Medin, D. L., and Ross, B. H. (1992). *Cognitive psychology*. Fort Worth: Harcourt Brace Jovanovich.

Menzel, W. (1998). Constraint satisfaction for robust parsing of spoken language. *Journal of Experimental and Theoretical Artificial Intelligence* 10: 77–89.

Millgram, E. (1991). Harman's hardness arguments. *Pacific Philosophical Quarterly* 72: 181–202.

Millgram, E. (2000). Coherence: the price of the ticket. *Journal of Philosophy* 97: 82–93.

Millgram, E., and Thagard, P. (1996). Deliberative coherence. *Synthese* 108: 63–88.

Minsky, M. (1997). A framework for representing knowledge. In J. Haugeland (ed.), *Mind design II* (pp. 111–142). Cambridge: MIT Press.

Misztal, B. A. (1996). *Trust in modern societies*. Cambridge: Polity Press.

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

Neurath, O. (1959). Protocol sentences. In A. J. Ayer (ed.), *Logical positivism* (pp. 199–208). Glencoe, Ill.: Free Press.

Nowak, G., and Thagard, P. (1992a). Copernicus, Ptolemy, and explanatory coherence. In R. Giere (ed.), *Cognitive models of science* (vol. 15, pp. 274–309). Minneapolis: University of Minnesota Press.

Nowak, G., and Thagard, P. (1992b). Newton, Descartes, and explanatory coherence. In R. Duschl, and R. Hamilton (eds.), *Philosophy of Science, Cognitive Psychology, and Educational Theory and Practice* (pp. 69–115). Albany: SUNY Press.

O'Laughlin, C., and Thagard, P. (forthcoming). Autism and coherence: a computational model. *Mind and Language*.

Oatley, K. (1992). *Best laid schemes: the psychology of emotions*. Cambridge: Cambridge University Press.

Ortony, A., Clore, G. L., and Collins, A. (1988). *The cognitive structure of emotions*. Cambridge: Cambridge University Press.

Paley, W. (1963). *Natural theology: selections*. Indianapolis: Bobbs-Merrill.

Panksepp, J. (1998). *Affective neuroscience: the foundations of human and animal emotions*. Oxford: Oxford University Press.

Paulos, J. A. (1980). *Mathematics and humor*. Chicago: University of Chicago Press.

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

Peirce, C. S. (1958). *Charles S. Peirce: selected writings*. New York: Dover.

Peng, Y., and Reggia, J. (1990). *Abductive inference models for diagnostic problem solving*. New York: Springer-Verlag.

Pennington, N., and Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology* 51: 242–258.

Pojman, L. P. (ed.), (1993). *The theory of knowledge: classic and contemporary readings*. Belmont, Calif.: Wadsworth.

Polya, G. (1957). *How to solve it*. Princeton, N.J.: Princeton University Press.

Prince, A., and Smolensky, P. (1997). Optimality: from neural networks to universal grammar. *Science* 275: 1604–1610.

Putnam, H. (1983). There is a at least one *a priori* truth. In H. Putnam (ed.), *Realism and reason*, vol. 3 of *Philosophical papers* (pp. 98–114). Cambridge: Cambridge University Press.

Quine, W. V. O. (1960). *Word and object*. Cambridge: MIT Press.

Quine, W. V. O. (1963). *From a logical point of view*. Second ed. New York: Harper Torchbooks.

Railton, P. (1986). Moral realism. *Philosophical Review* 95: 163–207.

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

Rawls, J. (1971). *A theory of justice*. Cambridge: Harvard University Press.

Rawls, J. (1996). *Political liberalism*. New York: Columbia University Press.

Raz, J. (1992). The relevance of coherence. *Boston University Law Review* 72: 273–321.

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

Reed, E. S. (1997). *From soul to mind: the emergence of psychology from Erasmus Darwin to William James*. New Haven: Yale University Press.

Richardson, H. S. (1994). *Practical reasoning about final ends*. Cambridge: Cambridge University Press.

Rock, I. (1983). *The logic of perception*. Cambridge: MIT Press.

Rosen, J. (1975). *Symmetry discovered*. Cambridge: Cambridge University Press.

Rosen, J. (1995). *Symmetry in science: an introduction to the general theory*. New York: Springer-Verlag.

Rumelhart, D., Smolensky, P., Hinton, G., and McClelland, J. (1986). Schemata and sequential thought processes in PDP models. In J. McClelland, and D. Rumelhart (eds.), *Parallel distributed processing: explorations in the microstructure of cognition* (vol. 2, pp. 7–57). Cambridge: MIT Press.

Russell, B. (1973). *Essays in analysis*. London: Allen and Unwin.

Sanders, J. T., and Narveson, J. (eds.), (1996). *For and against the state*. Lanham, Md.: Rowman and Littlefield.

Sayre-McCord, G. (1996). Coherentist epistemology and moral theory. In W. Sinnott-Armstrong, and M. Timmons (eds.), *Moral knowledge? New readings in moral epistemology* (pp. 137–189). Oxford: Oxford University Press.

Schank, P., and Ranney, M. (1991). Modeling an experimental study of explanatory coherence. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 892–897). Hillsdale, N.J.: Erlbaum.

Schank, P., and Ranney, M. (1992). Assessing explanatory coherence: a new method for integrating verbal data with models of on-line belief revision, *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 599–604). Hillsdale, N.J.: Erlbaum.

Sears, D., Huddy, L., and Schaffer, L. (1986). A schematic variant of symbolic politics theory, as applied to racial and gender equality. In R. Lau, and D. Sears (eds.), *Political cognition* (pp. 159–202). Hillsdale, N.J.: Erlbaum.

Selman, B., Levesque, H., and Mitchell, D. (1992). A new method for solving hard satisfiability problems. *Proceedings of the Tenth National Conference on Artificial Intelligence* (pp. 440–446). Menlo Park: AAAI Press.

Shelley, C., Donaldson, T., and Parsons, K. (1996). Humorous analogy: modeling *The Devil's Dictionary*. In J. Hulstijn, and A. Nijholt (eds.), *Proceedings of the Twente Workshop on Language Technology 12: Automatic Interpretation and Generation of Verbal Humor*. Twente: University of Twente.

Shrager, J., and Langley, P. (1990). *Computational models of scientific discovery and theory formation*. San Mateo: Morgan Kaufmann.

Shultz, T. R., and Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review* 103: 219–240.

Sinclair, L., and Kunda, Z. (1999). Reactions to a black professional: motivated inhibition and activation of conflicting stereotypes. *Journal of Personality and Social Psychology* 77: 885–904.

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

Spivey-Knowlton, M. J., Trueswell, J. C., and Tanenhaus, M. K. (1993). Context effects in syntactic ambiguity resolution: discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology* 47: 276–309.

St. John, M. F., and McClelland, J. L. (1992). Parallel constraint satisfaction as a comprehension mechanism. In R. G. Reilly, and N. E. Sharkey (eds.), *Connectionist approaches to natural language processing* (pp. 97–136). Hillsdale, N.J.: Erlbaum.

Stern, P. C. (1995). Why do people sacrifice for their nations? *Political Psychology* 16: 217–235.

Stich, S. (1988). Reflective equilibrium, analytic epistemology, and the problem of cognitive diversity. *Synthese* 74: 391–413.

Stocker, M., and Hegeman, E. (1996). *Valuing emotions*. Cambridge: Cambridge University Press.

Swanton, C. (1992). *Freedom: a coherence theory*. Indianapolis: Hackett.

Swinburne, R. (1990). *The existence of God*. Second ed. Oxford: Oxford University Press.

Swinburne, R. (1996). *Is there a god?* Oxford: Oxford University Press.

Taylor, G. J. (1994). The scientific legacy of Apollo. *Scientific American*, July, 40–47.

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

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

Thagard, P. (1991). The dinosaur debate: explanatory coherence and the problem of competing hypotheses. In J. Pollock, and R. Cummins (eds.), *Philosophy and AI: essays at the interface* (pp. 279–300). Cambridge: MIT Press.

Thagard, P. (1992a). Adversarial problem solving: modelling an opponent using explanatory coherence. *Cognitive Science* 16: 123–149.

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

Thagard, P. (1993). Computational tractability and conceptual coherence: why do computer scientists believe that P ≠ NP? *Canadian Journal of Philosophy* 23: 319–364.

Thagard, P. (1996). *Mind: introduction to cognitive science*. Cambridge: MIT Press.

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

Thagard, P. (forthcoming). How to make decisions: coherence, emotion, and practical inference. In E. Millgram (ed.), *Varieties of practical inference*. Cambridge: MIT Press.

Thagard, P., Eliasmith, C., Rusnock, P., and Shelley, C. P. (forthcoming). Knowledge and coherence. In R. Elio (ed.), *Common sense, reasoning, and rationality* (vol. 11). New York: Oxford University Press.

Thagard, P., Holyoak, K., Nelson, G., and Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence* 46: 259–310.

Thagard, P., and Kunda, Z. (1998). Making sense of people: coherence mechanisms. In S. J. Read, and L. C. Miller (eds.), *Connectionist models of social reasoning and social behavior* (pp. 3–26). Hillsdale, N.J.: Erlbaum.

Thagard, P., and Millgram, E. (1995). Inference to the best plan: a coherence theory of decision. In A. Ram, and D. B. Leake (eds.), *Goal-driven learning* (pp. 439–454). Cambridge: MIT Press.

Thagard, P., and Shelley, C. P. (forthcoming). Emotional analogies and analogical inference. In D. Gentner, K. H. Holyoak, and B. N. Kokinov (eds.), *The analogical mind: perspectives from cognitive science*. Cambridge: MIT Press.

Thagard, P., and Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science* 22: 1–24.

Thomson, J. J. (1971). A defense of abortion. *Philosophy and Public Affairs* 1: 47–66.

Trabasso, T., and Suh, S. (1993). Understanding text: achieving explanatory coherence through on-line inferences and mental operations in working memory. *Discourse Processes* 16: 3–34.

Tversky, A., and Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychological Review* 101: 547–567.

Van den Broek, P. (1994). Comprehension and memory of narrative texts: inferences and coherence. In M. A. Gernsbacher (ed.), *Handbook of psycholinguistics* (pp. 539–588). San Diego: Academic Press.

Watson, J. D. (1969). *The double helix*. New York: New American Library.

Westen, D. (2000). Integrative psychotherapy: integrating psychodynamic and cognitive-behavioral theory and technique. In C. R. Snyder, and R. Ingram (eds.), *Handbook of psychotherapy: the processes and practices of psychological change*. New York: Wiley.

Westen, D., and Feit, A. (forthcoming). All the president's women: affective constraint satisfaction in ambiguous social cognition. Unpublished manuscript, Department of Psyclology, Boston University.

Whewell, W. (1967). *The philosophy of the inductive sciences*. New York: Johnson Reprint Corp. Originally published in 1840.

Wigmore, J. H. (1937). *The science of judicial proof as given by logic, psychology, and general experience and illustrated in judicial trials*. Third ed. Boston: Little Brown.

Wilson, D. J. (1990). *Science, community, and the transformation of American philosophy, 1860–1930*. Chicago: University of Chicago Press.

Wood, J. A. (1986). Moon over Mauna Loa: a review of hypotheses of formation of Earth's moon. In W. K. Hartmann, R. J. Phillips, and G. J. Taylor (eds.), *Origin of the moon* (pp. 17–55). Houston: Lunar and Planetary Institute.

Zajonc, R. (1980). Feeling and thinking: preferences need no inferences. *American Psychologist* 35: 151–175.

Zemach, E. M. (1997). *Real beauty*. University Park: Pennsylvania State University Press.

Index