nonBayesian (Cohen 1986), agree that a jury, asked to decide whether the state has proven a criminal defendant guilty beyond a reasonable doubt, is required to consider not only (1) whether the prosecution has negated reasonable doubts that might otherwise arise from the evidence, but also (2) whether the prosecution has introduced enough evidence to dissipate reasonable doubts. Although one might argue that the preponderance of the evidence standard (which asks whether a proposition has been shown to be more probable than not) requires the jury to decide only whether one party's account of the evidence is more coherent (Lempert 1986), the question is still open. In any event, the preponderance standard allows jurors to discount the likelihood of a party's arguments because that party has not produced the evidence likely to be available if that party's argument is accurate (Allen 1986).

TEC and ECHO fail to address issues of completeness or adequacy of evidence. The portion of TEC that comes closest is Principle 6(b), which holds that the "acceptability of a proposition P . . . is reduced" if the proposition only explains "a few" of the relevant observations. The relevant portion of ECHO is a decay parameter that ECHO increases "in proportion to the ratio of unexplained evidence to explained evidence." Neither principle nor program deal with the absence of evidence that would normally be expected if a hypothesis (or contention) at issue were true. Accordingly, if ECHO is to reflect fact-finding adequately, it should include a procedure for measuring the adequacy of the evidence and each party's offer of evidence. (It would also have to reflect the possibility that jurors might formulate hypotheses of their own – but Thagard explicitly disavows any attempt to model theory generation.)

O'Rorke (1989) in his commentary on Thagard's target article, mentioned the possibility that one could decide to get more information before deciding to accept a given hypothesis. This comes close to my point, but does not quite reach it. First, fact-finders are in no position to gather further evidence. At most, they can find against the party with the burden of proof and send a message to future litigants that such evidence is simply insufficient. Second, and more telling for TEC and ECHO, thinness of evidentiary support can cause a decision not to accept any theory, even for the purpose of further investigation, simply because all theories seemed too speculative.

To the extent that Thagard recognizes problems of the completeness of evidence, his response to Sintonen's (1989) commentary seems to treat them as pertaining to decision-making rather than to choice of the "best available" theory. Whether issues of completeness of evidence properly pertain to theory choice or the allocation of one's limited resources for action is a fine point. Yet it would be unrealistic to suppose that one would spend cognitive resources, let alone material resources, on judgments about the acceptability of theories where the evidence seems lacking unless one were simply choosing a theory to test. Otherwise, the cognitive effort would be a waste. So to make Thagard's theory psychologically real, even as a description of choice in ostensibly "pure" theory, it needs some measure of the completeness of the evidence.

Any attempt to cause ECHO to assess the completeness of its own data base might, it is true, cause the same logical difficulties that Brilmayer (1986) pointed out in Bayesian attempts to model fact-finding – that interpreting Bayes' theorem to take account of the completeness of the evidence at hand causes a logical paradox akin to the liar paradox. On the other hand, Thagard could simply enrich the measure of explanatory coherence so that it included one's beliefs about the sort of evidence that should be present if a hypothesis is true. With appropriate measures of excitation and inhibition from the nodes representing those beliefs, ECHO would be a more psychologically adequate account of the acceptance of theories, in any context, and would mitigate, if not avoid, the philosophical objection.

Network and direct methods of maximising harmony

Nick Chater

Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, Scotland
Electronic mail: *nicholas@cogsci.ed.ac.uk*

Paul Thagard (1989t) presents an interesting and important theoretical account of choosing between hypotheses in scientific and everyday domains, describing a network model, ECHO, which aims to embody these ideas. Thagard's integration of philosophical, psychological, and computational considerations is impressive in its breadth, theoretical rigor, and conceptual simplicity. It is a compelling example of the value of interdisciplinary research in the understanding of scientific and everyday reasoning. The general problem of choosing between alternative hypotheses, explanations, or theories is so profound that it might be thought that modeling is wholly infeasible. Thagard is able to finesse nicely many thorny problems, however (by presupposing a formalisation of the domain of application into discrete propositions representing hypotheses and evidence, and taking the degree of coherence between these propositions as given), leaving a constrained yet important subproblem to be solved.

Central to Thagard's thesis is the claim that theory choice can be based on the "explanatory coherence" of sets of propositions, where "the global explanatory coherence of a system S of propositions is a function of the pairwise local coherence of those propositions" (Thagard 1989t, p. 437). In theory choice, the set of propositions with the maximal explanatory coherence is preferred.

Formally, Thagard measures explanatory coherence as a simple function H:

$$H = \sum_i \sum_j a_i \cdot a_j \cdot w_{ij} \qquad (1)$$

where a_i stands for the degree to which a proposition is accepted and w_{ij} stands for the degree to which propositions a_i and a_j cohere. Finding the most explanatorily coherent set of propositions amounts to assigning the values a_i such that H is maximised.

In ECHO each proposition corresponds to a unit of the neural network, and the values of the weight between pairs of units reflect the degree to which the propositions cohere or fail to cohere. The value of each unit represents the acceptability of the associated proposition. The measure H of the explanatory coherence of a system of propositions corresponds to the standard harmony measure for symmetric networks, and the network is designed to relax iteratively into a harmony maximum, hence picking the most coherent set of propositions (i.e., the most coherent theory).

In this commentary, I point out: first, ECHO does not maximize the harmony function H. H can be maximized using, for example, a Hopfield network scheme, however. And second, there is an alternative "direct" method of comparing theories – simply pick the theory that explains the most evidence – that invariably picks the theory that has the largest value of H. Which theory is favoured is, perhaps disturbingly, independent of the internal structure of the two theories, and interconnections between them.

1. ECHO does not maximise harmony. Three aspects of ECHO are incompatible with the maximisation of H: (i) the update rule can increase H; (ii) updates are performed synchronously rather than asynchronously; and (iii) the network is started from one rather than from a range of activation states, and so, in general, will find a local rather than a global maximum.

1. The update rule. In such symmetric networks as the Hopfield net (Hopfield 1982), Smolensky's harmonium (Smolensky 1986), and the Boltzmann machine (Hinton & Sejnowski

1986), each time a unit is updated, the "harmony" of the network increases (the exact details, such as the sign of the function and hence whether the network is taken to perform maximisation or minimisation, whether or not the units have biases, etc., vary among models). In a symmetrical network, the rate at which H changes as the activation of a unit a_j is varied is given by the partial derivative of the H function with respect to the activation of a single proposition/unit:

$$\frac{\partial H}{\partial a_j} = 2 \cdot \sum_{i \neq j} a_i \cdot w_{ji} \tag{2}$$

This quantity turns out to be equal to the input, net_j, to the unit a_j. This means that if the input to a is positive then the partial derivative is positive, and hence that an increase in a_j will increase harmony and a decrease in a will decrease harmony. Similarly, if the input is negative, the partial derivative is negative, and a decrease in a_j will increase harmony. Hence, if an updating scheme is to lead to a monotonic increase in H, then the activation of a unit must increase only when the input to that unit is positive, and decrease only when the activation of that unit is negative. That is,

$$\text{change in } a_j = a_j(t+1) - a_j(t) = (\text{positive term}) \cdot \text{net}_j \tag{3}$$

Thagard's update rule is:

$$a_j(t+1) = a_j(t)(1-\theta) + \begin{cases} \text{net}_j \cdot [\max - a_j(t)] & \text{if net}_j > 0 \\ \text{net}_j \cdot [a_j(t) - \min] & \text{otherwise} \end{cases} \tag{4}$$

which may be recast as follows:

$$\text{change in } a_j = a_j(t+1) - a_j(t) = -\theta \cdot a_j(t) + (\text{positive term}) \cdot \text{net}_j \tag{5}$$

(As it stands, if the input to a unit has a large absolute magnitude, its activation may depart from the $[-1, 1]$ range of activation values specified. Presumably a learning rate, γ, is intended in the second formula. I neglect this complication here, however, which amounts to setting the learning rate equal to 1.) The additional "$-\theta \cdot a_j(t)$" term means that the updating of the network will not in general maximize H. For example, consider the update of the unit A of the following two-node

$A_i(t) = 0.5$ 1 $B_i(t) = 0.5$

Figure 1 (Chater). Initial state of a very simple ECHO network. With $\theta = 1$, the current harmony of 0.25 will decrease to 0.125.

network, with θ set to (the unrealistically high value of) 1 (Figure 1). Notice that with θ equal to 1, the first term of (4) becomes 0, with the consequence that the level of activation at time $t + 1$ is independent of the activation at time t. According to Thagard's rule, we have

$$A_i(t+1) = 0 + [1 \cdot (0.5)] \cdot [1 - (0.5)] = 0.25 \tag{6}$$

which reduces H from the initial $(0.5) \cdot (0.5) \cdot 1 = 0.25$ to $(0.5) \cdot (0.25) \cdot 1 = 0.125$.

(2) Synchronous versus asynchronous update. The proofs that symmetrical networks maximise harmony, to which we adverted above, do not apply when all the units are updated at once, as in ECHO, but only when units are updated one by one. For example, suppose that both units in our simple network are updated together (Figure 2) (where the weight between the two units is denoted by w (>0)).

$$A_i(t+1) = 0.5 + (-0.5) \cdot (w \cdot [0.5 - (-1)] = 0.5 - 0.75w \tag{7}$$

$$B_i(t+1) = -0.5 + (0.5) \cdot (w \cdot [1 - (-0.5)] = -(0.5 - 0.75w) \tag{8}$$

$A_i(t) = 0.5$ w $B_i(t) = -0.5$

Figure 2 (Chater). Initial state of a very simple ECHO network. With synchronous update, if w is large the H will increase significantly.

If w is, say, 4, then the activities change from 0.5, -0.5 to -2.5, 2.5, with a concomitant change in H from $(0.5) \cdot (-0.5) \cdot 4 = -1$ to $(-2.5) \cdot (2.5) \cdot 4 = -25$, which amounts to a drastic decrease in H. More telling, if $w = 1.33 \ldots$, then the activation of each unit will oscillate between 0.5 and -0.5, rather than settling into a stable state. A basic result of dynamical systems theory is that such "limit cycles" cannot occur if a system is governed by a potential function – hence there can be no alternative harmony measure G according to which ECHO's behaviour is appropriate. It is interesting to note that the possibility of oscillation is not merely a theoretical curiosity, but was actually encountered – "ECHO undergoes activation oscillations only when the excitation parameter is high relative to inhibition" (Thagard 1989t, p. 457).

Given that harmony minimisation is central to Thagard's account, a natural way to implement the ideas underlying ECHO is simply to use a trivial generalisation of the Hopfield net. As before, weights would be continuously valued and represent coherence relations between propositions (in the standard Hopfield net, weights take only the values $\{-1, 0, 1\}$), and units update asynchronously. Instead of (4) we have a much simpler activation function – the activation of a unit can be either 1 or -1, depending purely on the sign of the input. The Hopfield net is known to maximize H (and the proof applies equally to networks with continuously valued weights). On this revised formulation, a proposition is either accepted or not, rather than being associated with a graded activation value. This may seem to lead to a loss of subtlety, relative to Thagard's original approach, in which propositions can be accepted to varying degrees. The amount of input that a unit receives, however, (whether it is strongly or weakly positive or negative) may be used as an alternative graded measure, if required.

(3) Local and global maximisation. In the kind of symmetrical networks discussed here, there are typically a number of stable states (or local maxima) into which the network may settle. Indeed, in illustrating the principles on which his model is based, Thagard discusses the well-known Necker cube example, which is bistable, having stable states associated with the two ways in which the figure may be given a three dimensional interpretation (Feldman & Ballard 1982). The starting activation values of the network determine into which stable state the network settles. Thagard starts all simulations with the activations of all units close to zero. It is entirely possible (indeed this will be the typical case) that, given some other starting configuration, the network will settle into an alternative stable state, perhaps favouring the theory rejected by the "zero" start. Such an alternative stable state may have a greater H value, and hence represent a more coherent interpretation. Because the concern is to find a global rather than a local maximum of H, it is important to assess the stable states obtained from a large range of varying activation starts and finding the best of these, rather than implicitly assuming that the "zero" start corresponds to the most coherent solution.

2. A direct method of assessing which of two theories is most coherent. As McDermott (1989) points out in his commentary, although Thagard's central claim is that theory choice can be effected by attempting to maximise the H function, the particular way in which H is maximised is of relatively minor concern. Specifically, the use of neural network hardware is something of a distraction (and particularly so in view of the looseness of the

mapping between the maximisation and the network implementation). I should like to close this commentary by suggesting a much simpler way Thagard's maximisation criterion can be implemented without using any hardware at all. Hobbs (1989) wonders if the theory choice judgements produced by ECHO could be rivalled by simply counting the difference between the number of pieces of evidence explained and the number of hypotheses in each theory. It turns out that an even simpler rule – counting the number of pieces of evidence explained by each theory – will serve to pick the theory that has the highest value of H. I shall call this the direct method of assessing relative values of H. It is derived as follows:

Suppose that we have two theories containing sets of propositions A and B, which comprise the set of hypotheses, and a set of propositions, E, concerning evidence. We wish to assess which of A and B has the highest H value. To assess the H value for A (the case in which A is true and B is false), simply assume that propositions in A and E have value 1, and that propositions in B have value -1. Similarly, to assess the H value for B (that B is true and A is false), set B and E to 1, and A to -1. Let us consider the contributions to H provided by terms that involve pairs of hypotheses, pairs of pieces of evidence, and pairs containing one hypothesis and one piece of evidence. In switching from considering A true to considering B true, all the hypotheses have changed sign, and the pieces of evidence are still assigned the value 1. Hence only the third class of product, containing one hypothesis and one piece of evidence, will change sign from considering A true to considering B true. (This means that all contradictory links among hypotheses of different theories, all complex explanatory relationships among hypotheses within a theory, and contradictions among pieces of evidence can be ignored without effecting the results of H maximization.) Specifically, if this third term is positive for A, it will be negative for B, and hence A will have a higher H value. It turns out that it is extremely easy to assess the value of this key term, given the way coherence is assigned in ECHO. Let us consider the contribution to H of a particular piece of evidence $E1$. The harmony associated with $E1$, if it is explained by a single proposition $H1$ in A, will be

$$\langle \text{value of } E1\rangle \cdot \langle \text{value of } H1\rangle \cdot \langle k\rangle = 1 \cdot 1 \cdot k \qquad (9)$$

where k is the default weight. Suppose that the evidence is explained by n, rather than just 1, proposition in A. In accordance with his explanatory principle 2(c) (p. 437), Thagard assigns weights with a strength inversely proportional to the number of hypotheses involved in the explanation. The harmony associated with the explanation in this case will be

$$1 \cdot 1 \cdot k/n + 1 \cdot 1 \cdot k/n + \ldots + 1 \cdot 1 \cdot k/n = 1 \cdot 1 \cdot k \quad (n \text{ terms}) \qquad (10)$$

Thus, the harmony associated with a piece of evidence explained by A is constant, and not dependent on the number of propositions of A that it is explained by. Equally, the negative harmony associated with a piece of evidence explained by the theory (say, B) whose propositions have been set to -1, will also be constant, having the same absolute magnitude, by the opposite sign (i.e., $1 \cdot -1 \cdot k$). Hence the harmony associated with all the pieces of evidence can be assessed simply by comparing the number of hypotheses explained by A with the number explained by B. The theory that explains more evidence will invariably be associated with a higher H, quite independently of the structure of that theory – how broad its explanations are, how many unmotivated assumptions are introduced and so on.

It is interesting to compare the results of the direct method to those produced by ECHO in the four examples that Thagard details. In three cases, the accounts agree; in the fourth, the judgement on the Peyer murder trial, unlike the direct method, ECHO favours a guilty verdict. Because the direct method assesses H analytically, and ECHO optimises H very imperfectly, I am inclined to conclude that this is an example of the failure of ECHO to reach what Thagard's theoretical considerations dictate to be the right decision.

The observation that H maximisation may be replaced with a simple counting rule effects a considerable computational saving in implementing Thagard's account of theory choice; the fact that the account is insensitive to the structure of explanation, however, undermines its plausibility as a model of theory choice in philosophy of science or psychology.

Empirical investigation or rational reconstruction?

Stephen M. Downes

Philosophy Department, University of Cincinnati, Cincinnati, OH 45221
Electronic mail: *downes@ucbem.bitnet*

Thagard's (1989t) reconstruction of Lavoisier's choice of the oxygen theory does not do justice to the historical facts. Although several of the first-round commentators bring up this point, it is one worth developing in more detail. Thagard offers ECHO's choice, according to criteria of explanatory coherence, between the oxygen and the phlogiston theories in chemistry as a model of Lavoisier's psychological processes. Yet he draws his evidence for setting up the model from a paper written late in the oxygen debate, the 1783 paper, "Reflexions sur le Phlogistique." Thagard says that "the input given to ECHO represents Lavoisier's argument in his 1783 polemic against phlogiston" (1989, p. 444). And yet he claims that ECHO models the actual psychological processes of Lavoisier in choosing the oxygen theory over the phlogiston theory. By 1783, Lavoisier had established the oxygen theory in some detail, and both this theory and the phlogiston theory had gone through several modifications. The historical evidence indicates that Lavoisier himself was convinced of the falsity of the phlogiston theory in 1772, the so called "crucial year" (see Guerlac 1981; cf. Perrin 1989). The point of the 1783 "polemic" was to persuade other members of the scientific community of the superiority of the new oxygen theory. And so, as Giere (1989) correctly points out, Thagard is modeling the arguments that Lavoisier presents; and yet it is not clear that he is modeling the psychological processes that led Lavoisier himself to choose the oxygen theory. Perhaps ECHO can be better characterized as a model of a later dispute between Lavoisier and one of his remaining opponents (e.g., Priestly), but this again undermines the claim that ECHO models Lavoisier's actual psychological processes at the time of his discovery.

The process of comparison between the arguments a scientist puts forward in defense of a theory and the arguments his opponents put forward is more of a social process than a psychological one. Can Thagard's model be a sociological one? Wetherick (1989) thinks that the model is sociological, but Thagard denies this, without explanation. Thagard also applies ECHO to a case that looks to be a prima facie social process, however: jury decision making.

Thagard uses ECHO to model the decisions of juries in two court cases, presenting the case for the prosecution as one theory and the case for the defense as the competing theory. The evidence is the same for both cases, although on occasion defense and prosecution will introduce contradictory evidence, and both bits of evidence are included. ECHO produces a decision as to which "theory" coheres better, and thus pronounces the defendant guilty or not guilty. The pertinent question in this case is what is ECHO modelling? Is it modelling the psychological processes of any one particular jury member, or all of the jury members? Given Thagard's concern for the psychological realism of his account, his primary aim may be to model psychological processes of individual jurors. I contend