Mangan, B. B., & Palmer, S. E. (1989). New science for old. A peer commentary on P. Thagard. Explanatory coherence. *Behavioral and Brain Science, 12*(3), 480–482.

New science for old

Bruce Mangan and Stephen Palmer

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

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

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

On the other hand, the actual structure of Thagard's simulation looks much closer to Kant, with a tincture of Bacon. There is nothing wrong with Bacon or Kant. As philosophers of science they are a bit out of style, but that does not make them less important or less potentially valuable for current thinking; much current thinking is built on them. However, if one were given the exercise of putting some of the more recent ideas about the nature of science and scientific explanation into connectionist terms, an architecture rather different from Thagard's would probably emerge, one which would take advantage of more of the resources of connectionism. For purposes of comparison, we will later sketch an example of this sort. But for the most part, we will consider some of the less "progressive" components of Thagard's model and see how they contrast with the theoretical ideas Thagard seems to believe he incorporated in ECHO.

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

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

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

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

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

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

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

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

In contrast, consider a more "holistic" and "deductive" way of choosing between hypotheses, but one that is still roughly within the ECHO format: Activation enters the system, not through evidence units, but via a given *hypothesis* unit that is "clamped" on. This hypothesis unit then evokes its own best-fitting configuration of activation in the network. The hypothesis is then deactivated, the next hypothesis unit is clamped on, and the process is repeated for each remaining hypothesis. The winning hypothesis is the one that creates the most coherent network. Note that in this case we are choosing the best hypothesis by observing its direct effect on the network as a whole and not by using any indirect measure such as the relative activation of the hypothesis units compared in isolation. Further changes in ECHO's architecture would probably be necessary to implement this idea, but the general point should be clear: The present proposal attempts simultaneously (1) to bring ECHO closer to the modern view of hypotheses as central organizing devices and (2) to use system coherence directly to evaluate explanations in Thagard's sense.

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

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

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

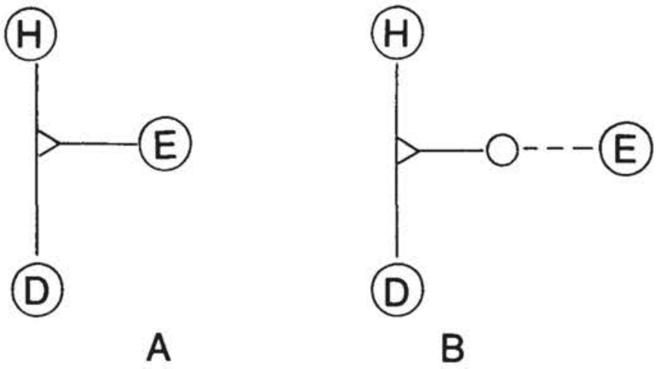

Figure 1 (Mangan and Palmer). Modeling explanatory principles as "gating" units in a connectionist network. Triangular symbols represent special connections in which the product of each pair of units gets transmitted to the third. Figure 1A shows how an *excitatory* connection between a datum unit (D)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

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

References

Bacon, F. (1859) Novum organum (first published 1620). In: *The works of Francis Bacon,* vol. 1, ed. J. Spedding, R. Ellis & D. N. Heath. Longmans.

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

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

Himmelfarb, G. (1962) *Darwin and the Darwinian revolution.* Doubleday.

Kant, I. (1787/1963) *Critique of pure reason.* N. Kemp Smith, translator, 2nd edition. Macmillan.

Kant. I. (1790/1951) *Critique of judgement.* J. H. Bertrand, translator. Collier-Macmillan. [BM]

Kuhn, T. S. (1970) *Structure of scientific revolutions* (2nd ed., first published 1962). University of Chicago Press.

Hinton, G. E. (1981) A parallel computation that assigns canonical object-based frames of reference. *Proceedings of the International Joint Conference on Artificial Intelligence,* Vancouver, Canada.

Leibniz, G. W. (1765/1981) *New essays on human understanding,* ed. P. Remnant & J. Bennett. Cambridge University Press.

MacKinnon, E. (1978) The development of Kant's conception of scientific explanation. Proceedings of the 1978 Biennial Meeting of the Philosophy of Science Association**.**

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

Quine, W. V. (1960) *Word and object.* MIT Press. [WCL]

Quine, W. V. (1961) From *a logical point of view* (2nd ed.). Harper Torch-books.