Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society.* (pp. 426-432). Hillsdale, NJ: Erlbaum.

EXPLANATORY COHERENCE AND BELIEF REVISION IN NAIVE PHYSICS

Students of reasoning have long tried to understand how people revise systems of beliefs. We maintain that people often change their beliefs in ways driven by considerations of explanatory coherence. In this paper, we describe a computational model of how experimental subjects revise their naive beliefs about physical motion. First, we present instances in which subjects changed their beliefs while learning elementary physics. Each of these cases involved an individual's attempt to explain a surprising observation. Next, we show how their belief revisions can be modeled using ECHO, a connectionist computer program that uses constraint-satisfaction techniques to implement a theory of explanatory coherence. The resulting simulations even captured temporal characteristics of the observed changes in belief. Finally, we discuss the model's representational sensitivity and procedural robustness, and conclude by showing how ECHO can be used to generate empirical predictions about subjects' current beliefs.

EXPLANATORY COHERENCE AND BELIEF REVISION IN NAIVE PHYSICS

Michael Ranney and *Paul Thagard*

Cognitive Science Laboratory
221 Nassau Street
Princeton University
Princeton, NJ 08542

Students of reasoning have long tried to understand how people revise systems of beliefs (see Wertheimer, 1945, for example). We will describe a computational model of how experimental subjects revise their naive beliefs about physical motion. We maintain that people often change their beliefs in ways driven by considerations of explanatory coherence. After describing instances in which subjects change their beliefs while learning elementary physics, we show how their belief revisions can be modeled using ECHO, a connectionist computer program that uses constraint-satisfaction techniques to implement a theory of explanatory coherence.

THE PHENOMENA: CHANGES IN SYSTEMS OF BELIEFS

Ranney (1987a) investigated belief change in naive subjects learning elementary physics by using feedback provided on a computer display. Subjects were asked to predict the motion of several projectiles and then explain these predictions. The physical contexts were quite simple, involving objects that were either thrown or released in various ways. Analyses of verbal protocol data indicate that subjects sometimes underwent dramatic belief revisions while offering predictions or receiving empirical feedback. We will describe two kinds of revisions.

Pat's Changes

Consider "Pat," a woman who was asked to offer predictions about events including (a) the motion of a heavy object dropped by a briskly walking man and (b) the motion of a heavy object thrown *obliquely* upward. Using episodic memories and mental imagery, Pat initially predicted that the object dropped by the man would fall straight down (relative to the ground). This belief is a common finding in the naive physics literature (McCloskey, Washburn, & Felch, 1983). Although she entertained the correct prediction, that the dropped object might curve forward owing to the object's forward "force" (velocity), she preferred to stay with the straight-down belief.

Several tasks later, when faced with the "upward-throw" situation, Pat noted a similarity between it and the "walking-drop" task -- one that eventually spawned a belief revision. While she offered the correct parabolic trajectory as a prediction for the upward-throw, she noted that, at the parabola's zenith, the upwardly thrown object is comparable to that just released by the walking man. That is, at the apex of the thrown object's trajectory, it has an exactly-horizontal motion, as does the just-dropped object. Pat then mentioned that this observation was not "consistent" with what she said before and, if she *were* to be consistent, the thrown object would "stop" its horizontal motion and "then just fall straight down" from the zenith of the parabola. This "curving-up-then-straight-down" trajectory was not consistent with her past experience of falling objects.

Pat then realized that her memory-driven description of the ball dropping straight down from the walking man involved beliefs that were incoherent with her beliefs about the parabolic motion of thrown bodies. After a period of ignoring the incoherence, Pat stated that she had "constructed a consistent theory of how these things move." Remarkably, she went on to reject her straight-down prediction for the walking-drop task and accept the belief that the path would have a "slight forward" arc combining the "forward force" and gravity. Eventually, Pat generalized this notion, discriminating among the breadths of the arcs of several laterally released projectiles. (A "laterally" released projectile has an initial velocity that includes a horizontal component.)

Hal's Changes

A second kind of systematic belief revision occurred in subjects who offered predictions, received feedback, and provided explanations for a set of tasks in which pendulum-bobs were released from their supportive strings during various points in a swing. This set of tasks was adapted from stimuli used by Caramazza, McCloskey, and Green (1981). Because of the similarity among several of the subjects, we will amalgamate them into a composite subject "Hal."

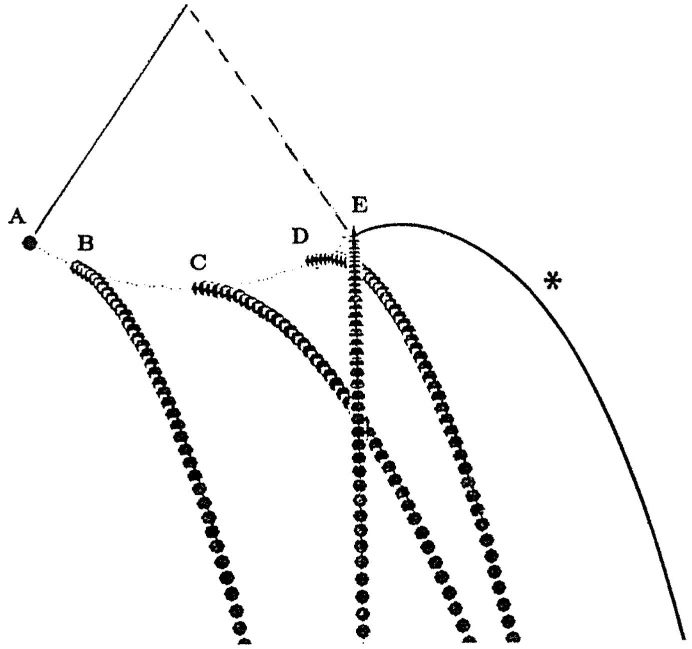

Figure 1. Hal's prediction (*) and four feedback paths

As shown in Figure 1, Hal predicted that at the extreme endpoint of a swing, a released bob will travel laterally and (eventually) downward. To some extent, this prediction was driven by images of children flying off playground swings. Via feedback, Hal learned that a bob released in this manner actually falls straight down (from position E in

Figure 1). Most of the subjects observed by Ranney (1987a) were surprised by this piece of feedback, as almost 90% of the predicted trajectories were nonvertical. Virtually all these subjects revised some beliefs, offering explanations similar to the following prototype:

> Unlike the bobs with the other release positions, this bob went directly straight down, not to the side at all. Since it had no lateral motion as it fell, this means that the object had no speed when it was released. Therefore, the pendulum must have been temporarily stopped when the bob dropped. This makes sense, since the pendulum was probably slowing down -- and it had to stop in order to change directions!

In contrast to Pat's belief change, in which two incoherent predictions caused her to reject one of the them, Hal's belief system underwent a more dramatic revision. He came to accept both the straight-down feedback and the notion of an instantaneous zero velocity, while rejecting both his earlier (lateral) prediction and an impetus-driven belief regarding pendular motion. (See Halloun & Hestenes, 1985, and Ranney, 1987b, for descriptions of different sorts of impetus beliefs.)

EXPLANATORY COHERENCE AS A MECHANISM FOR SYSTEMATIC BELIEF REVISION

How can we account for these systematic changes in beliefs? Both cases involve a subject's attempt to adjust beliefs in order to explain a surprising observation. An adequate model of these phenomena must provide a mechanism by which a coherent, revised, set of beliefs can arise from the need for explanation.

ECHO

Thagard (1988a) has proposed a theory of explanatory coherence that builds on previous ideas about the evaluation of explanatory hypotheses (Harman, 1986; Thagard, 1988b). The theory has been implemented in a connectionist computer program, ECHO, that uses parallel constraint satisfaction to accept and reject hypotheses on the basis of their explanatory coherence. ECHO has been used to analyze a variety of scientific arguments, past and present: Lavoisier's case for his oxygen theory against the phlogiston theory, Darwin's argument for evolution by natural selection, controversies about continental drift (Thagard & Nowak, 1988), and debates about why the dinosaurs became extinct. Application of ECHO to the belief revisions in Pat and Hal is novel in two respects. First, we are modeling subject protocols produced during experiments rather than finished arguments. Second, these models are dynamic, in that ECHO changes its coherence judgments in response to new evidence.

Space constraints permit only a sketch of the theory of explanatory coherence and its implementation (see Thagard, 1988a, for greater detail). The theory is stated using seven principles of explanatory coherence that can be summarized as follows. Principle 1, *Symmetry*: Coherence and incoherence are symmetric relations. Principle 2, *Explanation*: Hypotheses that together explain a piece of evidence cohere with the evidence and with each other, and the degree of coherence decreases with the number of hypotheses used in the explanation. Principle 3, *Analogy*: Similar hypotheses that explain similar pieces of evidence cohere. Principle 4, *Data Priority*: Pieces of evidence have a degree of coherence in themselves, even though evidence can be rejected for theoretical reasons. Principle 5, *Contradiction*: Contradictory propositions are incoherent. Principles 6 and 7, *General* and *System Coherence*: The explanatory coherence of a proposition or set of propositions is determined by the pairwise relations established by principles 1-5.

ECHO is a Common LISP program whose input consists of statements about the explanatory and contradictory relations among propositions. It creates units representing propositions and sets up links between pairs of propositions in accord with these seven principles of explanatory coherence. If two propositions cohere because they are both arguments of a particular explanation, then ECHO sets up an excitatory link between them. If two propositions are incoherent because they contradict each other, then ECHO sets up an inhibitory link between them. In accord with the principle of data priority, propositions representing evidence receive a link from a special evidence unit. For modeling the physics students, we treat as evidence propositions based on either (a) the presence or absence of direct observations, (b) memories of such observations, or (c) facts that are well established for the subject, such as "gravity pulls objects downward."

The mathematics underlying ECHO are straightforward. Following typical connectionist practice (Rumelhart & McClelland, 1986), each unit has an activation that is updated by considering the units that are linked to it. A unit's excitatory link with another unit whose activation is greater than 0 tends to increase the first unit's activation, whereas an inhibitory link with the other unit tends to decrease activation. More generally, for each unit j, the activation a_j is a continuous function of the activation of all the units linked to it, with each unit's contribution depending on the *weight* w_{ij} from unit i to unit j. The activation of a unit j can be updated from time t to time t+1 using the following equation.

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

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

$$net_j = \sum_i w_{ij} a_i(t) \qquad (2)$$

Repeated updating cycles result in some beliefs gaining acceptance (activation > 0) while others are rejected (activation < 0). ECHO networks eventually settle into stable states in which the units have asymptotic activations that represent their coherence with other units.

Applying ECHO To Pat's Belief Revision

We have used ECHO to analyze the kinds of belief revision exhibited in the subjects described. In each case, a contradiction among the subject's beliefs appeared to serve as the motivation for the observed changes. ECHO deals with contradictions gracefully, treating them as pressures to change beliefs, but otherwise tolerating them. Pat's case involved a critical incoherence between two mutually exclusive predictions: a piece of evidence that was supposedly observed, and a hypothesis that was not observed, yet consistent with other observations and hypotheses.

The following is a list of Pat's initial set of propositions, as garnered from her verbal protocol of the problem-solving session. They represent her active beliefs just after she provided her straight-down prediction for the walking-drop task.

Evidence:

E1. Carried objects fall straight-down upon release.
E2. Carried objects don't fall diagonally upon release.

Negative Evidence (proposed observations that do not obtain):

NE1. Carried objects fall diagonally upon release.

Common Fact:

CF1. Gravity moves released objects downward.

Newtonian Hypotheses:

NH1. Laterally moving objects begin to curve downward (immediately) upon release.
NH2. Released objects move forward via a forward velocity.

Alternative (non-Newtonian) Hypotheses:

AH1. Horizontally moving objects fall straight-down (immediately) upon release.
AH2. Released objects move forward via a forward "force."

The following are Pat's original verbalized explanations, manifested in ECHO as excitatory links among each of the propositions involved.

Explanations:

E1 is explained by AH1;
E2 is explained by NH1;
E2 is explained by AH1;
NE1 is explained by CF1 and AH2;
NH1 is explained by CF1 and NH2;

The next set of relations are the inconsistencies that Pat originally mentioned. Recall that the contradiction that disturbed Pat was the one between E1 and NH1; she couldn't accept both (a) that laterally released objects curve downward and (b) that carried objects (also being laterally released) fall straight-down.

Contradictions:

E1 versus NH1;
E2 versus NE1;
NH1 versus AH1;

When Pat was later asked to offer a prediction for the upward-throw task, she added the following beliefs:

New Evidence:
E3. Upwardly thrown objects curve up-and-down.
E4. Upwardly thrown objects do not curve up and fall straight-down.

New Negative Evidence:
NE2. Upwardly thrown objects curve up, then fall straight-down.

Finally, Pat verbalized the following explanations and contradictions. Note that the explanation of NE2 is essentially a (higher-order) explanation of a hypothesis *by* other hypotheses:

New Explanations:
E3 is explained by NH1;
E4 is explained by NH1;
NE2 is explained by NH1 and AH1;

New Contradictions:
E3 versus AH1;
E4 versus NE2;
E4 versus AH1;

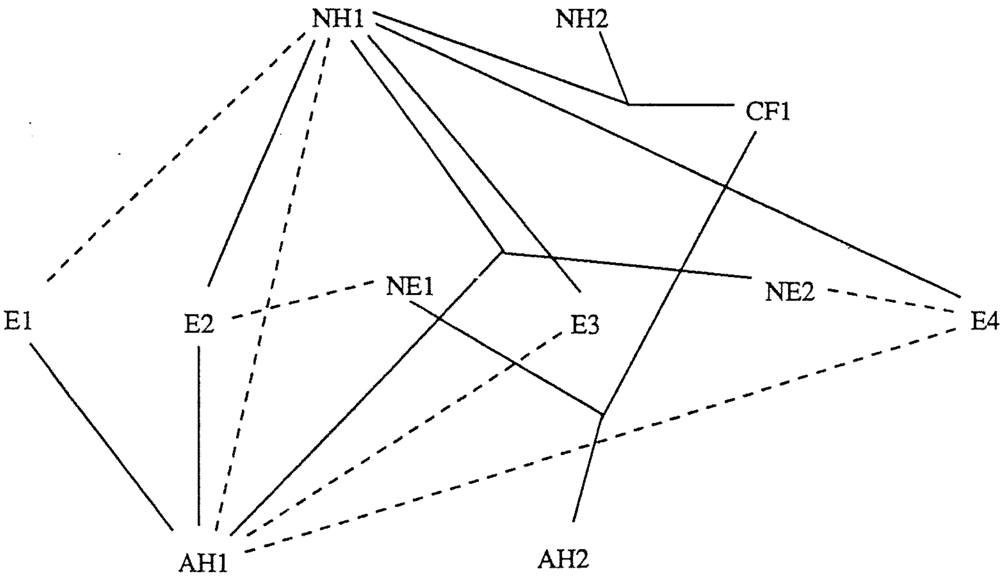

Figure 2. Pat's explanatory coherence network

Figure 2 displays the network ECHO forms from these explanations and contradictions, with solid lines representing symmetrical excitatory links and dashed lines representing symmetrical inhibitory links. We suggest that the figure displays the essential structural aspects of Pat's working memory during the belief change in question. The graph shows that prediction NH1 is well-supported by evidence E2, E3, and E4, as well as by fact CF1 and hypothesis NH2. Prediction E1, being a "remembered" observation, has a direct source of activation via principle (4) yet is supported only by the Aristotelian hypothesis AH1.

In order to approximate Pat's belief change, ECHO should exhibit an initial acceptance of E1, followed by its rejection in favor of NH1. As Figure 3 illustrates, these characteristics are indeed captured by ECHO. The activation (from -1 to +1 on the y-axis) of each node is plotted against time (from 0 to 200 cycles of activation updating). With each node initially set to zero activation, the system relaxes into more and more coherent states, such that E1's trajectory follows the desired nonmonotonic path -- rising sharply, then falling into the rejected region -- as NH1 advances and AH1 declines. The other propositions are similarly accepted or rejected (or held in limbo, as is AH2), depending upon their local coherence relationships within the overall constraint-satisfaction system. Note that the model also simulates the temporal aspect of Pat's reasoning, as the "new" propositions, E3, E4, and NE2, as well as their associated explanations and contradictions, are introduced after a brief lag (after 15 cycles). The final, most stable configuration of beliefs happens to be one that roughly corresponds to Newtonian motion. (Of course, if Pat had happened to recall other evidence that supported her alternative hypotheses, this need not have been the case.)

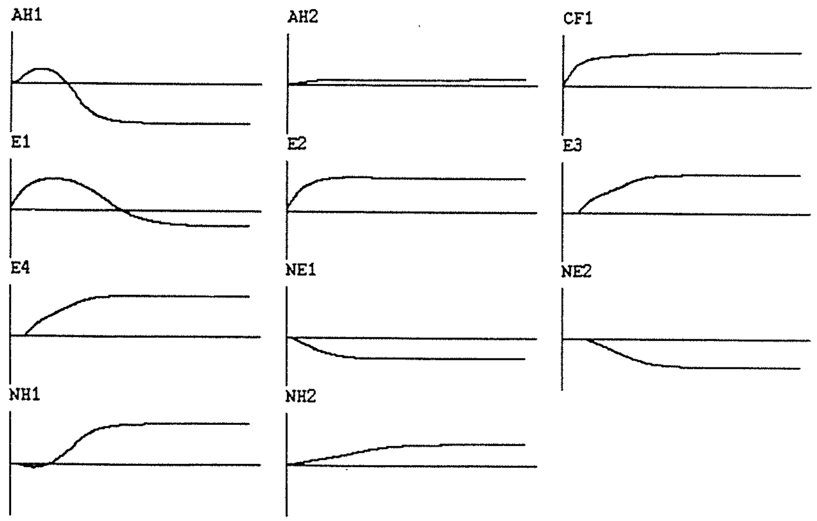

Figure 3. The activation trajectories of Pat's beliefs

A Dynamic Simulation Of Hal's Belief Revision

A simulation of Hal's belief changes involves a more intensive temporal analysis. Recall that Hal's revision was due to an empirically driven contradiction, in contrast to Pat's more memory-driven contradiction. Here are Hal's essential original beliefs (i.e., his beliefs prior to receiving *any* trajectory feedback about pendulum-bobs that are released during a swing). Keep in mind that Hal is a composite subject: these are beliefs that were characteristic of many of the subjects who underwent essentially the same belief revision.

Evidence:

E1. Kids can fly off the end of a playground swing.
E5. A pendulum reverses direction at the endpoints.

Common Facts:

CF1. Gravity pulls objects downward.
CF2. A swing is a pendulum.

Classical Physical (Newtonian) Hypotheses:

CP1. At the endpoints, a pendulum is at rest.
CP2. A laterally released object moves over and down.
CP3. The slower a pendulum-bob's speed at release, the smaller the curved trajectory.

Alternative (non-Newtonian) Hypotheses:

AH1. At the endpoints, a pendulum-bob continues its preceding lateral motion.

Predictions:

P1. At the endpoint, a released bob will move over and down.
P2. At the endpoint, a released bob will fall straight-down.

Both E1 and E5 are remembered observations; E2, E3, and E4 are intentionally left out, because these pieces of evidence will be sequentially added as feedback, as described later. The following explanations and contradictions were common to protocols reflecting Hal's belief revision. Note that the critical incoherence (which feedback eventually resolves) is between P1 and P2, two mutually exclusive predictions with different levels of support and competition.

Explanations:

E1 is explained by AH1 and CF2;
E5 is explained by CP1;
P1 is explained by AH1 and CP2;
P2 is explained by CP1 and CP3;
P2 is explained by CP1 and CF1;
CP2 is explained by CF1;

Contradictions:

E1 versus P2;
P1 versus P2;
CP1 versus AH1;

Figure 4 shows that when ECHO is loaded with this information at time t_0, the system reaches a stable state by t_1 (after 150 processing cycles). Among other dynamic relations, these graphs show that P1 (the curving-down-at-endpoint prediction) is believed, while its antagonist, P2 (the straight-down-at-endpoint prediction), is disbelieved -- as indicated by its negative activation. Thus, t_1 represents the state of Hal's belief system prior to any feedback.

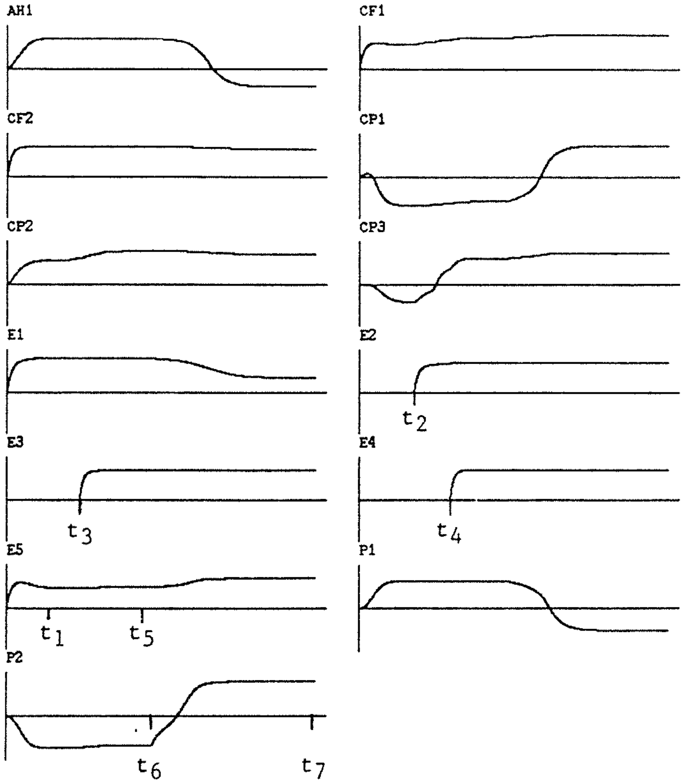

Figure 4. The activation trajectories of Hal's beliefs

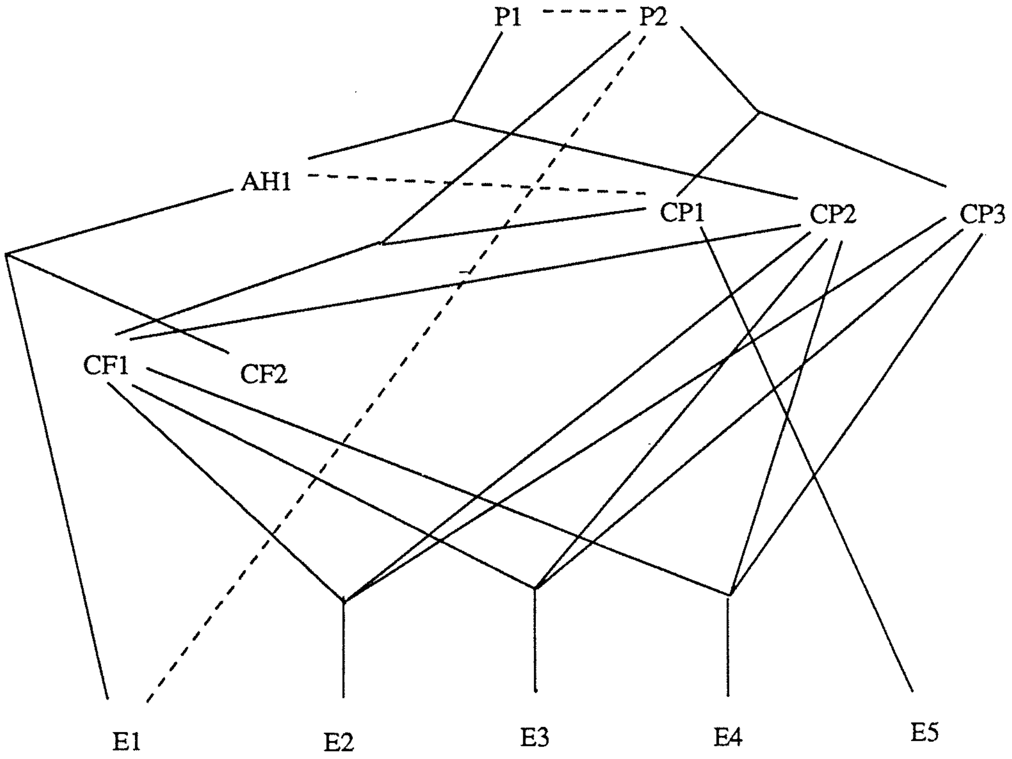

Figure 5. Hal's explanatory coherence network

At t_2, t_3, and t_4 (of Figure 4), evidence about other pendular-release positions is acquired in the form of direct observations (i.e., feedback) E2, E3, and E4. These "within-swing" paths are readily explained by (and hence support) propositions CF1, CP2, and CP3:

New Evidence:
E2. A bob released on a downswing curves down after its release.
E3. A bob released from midswing curves out (a lot) after its release.
E4. A bob released on an upswing curves up-and-down after its release.

New Explanations:
E2 is explained by CF1, CP2, and CP3;
E3 is explained by CF1, CP2, and CP3;
E4 is explained by CF1, CP2, and CP3;

The system then settles into state t_5 (after 400 total cycles). Figure 4 shows that, except for the generalization expressed in CP3 (relating release-velocity to the breadth of curves), little has changed from state t_1; P1 is still believed and P2 is not. Figure 5 shows Hal's belief system from t_5 onward, including all excitatory and inhibitory links.

As described earlier, it is at time t_6 that the dramatic belief revision begins, driven by the surprising feedback that, contrary to P1, the endpoint release yields a straight-down path (as predicted by the disbelieved P2). This feedback is simulated in ECHO by making P2 a data node, thus providing it with a direct source of activation (like E1-E5, CF1, and CF2, which also have data priority.) As Figure 4 indicates, this single change has five dramatic consequences between t_6 and Hal's ultimate state (after 850 total cycles), t_7: (a) P2 gains acceptance, flipping from a negative to a positive activation-state, while (b) the antagonistic P1 is rejected. (c) CP1, the notion of instantaneous zero velocity, achieves acceptance, while (d) its non-Newtonian antagonist, AH1, is rejected. (e) Even E1, a fallacious piece of "evidence" (i.e., that kids can fly off the end of swings) loses support. These changes essentially reflect the belief revisions verbalized by subjects like Hal.

ASSESSING THE MODEL

Although these simulations provide general correspondence with Pat's and Hal's changes in belief, there are several methodological questions to consider. We must ask how sensitive ECHO is to (a) the particular representation of an individual's beliefs and (b) the particular parameters involved in activation-passing.

How arbitrary are the representations that are put to ECHO? Pat's beliefs were garnered directly from audio-taped protocols. Nevertheless, there is no fool-proof algorithm for translating utterances into propositions, so analysis has some latitude. Similarly, although we tried to include only relations that were explicitly used in Pat's explanations, this part of the analysis also involves some subjectivity. It is particularly difficult for the coder to refrain from adding an obvious node or a link even though the particular subject didn't vocalize that obvious belief or relation. (For instance, the authors found it difficult *not* to add an inhibitory link between Pat's AH2 and NH2.) Constructing Hal's belief system allowed more latitude than Pat's, since he is a composite. Still, care was taken to create the network first -- before tinkering with the processing parameters -- so that we would be less likely to "kludge" the representation.

There is also another kind of representational question: What does one of these networks actually represent? Generally, we conceive of the networks as models of the current contents of working memory. Note, however, that by "current" we also mean "contextual," because subjects can hold a belief in one context that they disbelieve in another. For instance, in an abstract context, most subjects explicitly held CP1, that there is no speed at a pendulum's endpoints -- even those, like Hal, who would reject it (in favor of AH1) during the context of the pendular-release tasks.

A general problem with connectionist cognitive models is that they usually have numerous numerical parameters that can be manipulated to produce desired results. Does our simulation depend on fine parameter tuning? The most important parameters in ECHO include the weight value of excitatory links, the negative weight value of inhibitory links, the weight value of the (data priority) links between evidence and the special evidence unit, and the decay of each unit at each cycle. The simulations of both Pat and Hal used the same parameter settings, and yielded the desired trajectories over similar ranges for each parameter. These common parameter ranges were: .015 to .05 for excitatory weights, -.05 to -.065 for inhibitory weights, .035 to .075 for data-priority weights, and .01 to .065 for the decay rate.

The simulations *might* have employed even more parameters. For instance, we treated units representing direct observations, memories, and facts all as evidence, with each linked to the special evidence unit by the same weight. But one can argue for varying these weights for different kinds of evidence, increasing them for current observations and decreasing them for fuzzy memories. Not all evidence has the same epistemic status. In particular, when Hal is directly presented with a phenomenon on the computer screen in front of him, this becomes a very salient piece of evidence. Accordingly, one might argue that the unit representing the surprising observation that the pendulum bob falls straight down at the end of its swing should be a multiple of the data priority of remembered evidence.

FUTURE RESEARCH

We have been modeling previously performed experiments, but ECHO can also be used to make predictions about the beliefs of subjects. Our simulation of Hal predicted that he would come to doubt the belief that kids can fly off the end of a playground swing, but very few subjects *explicitly* re-evaluated this belief. ECHO predicts that the subjects may have experienced this belief change even if they did not mention it, and this prediction can be tested in new experiments by asking subjects to state their confidence in belief E1 following the relevant feedback. Additional experimental tests of the extent to which ECHO models human performance can be done in situations where people face difficult inference problems involving judgments of explanatory coherence. We conjecture that problems that are relatively hard for people, as measured perhaps by the length of time to generate answers, will also be relatively hard for ECHO, as measured by the number of cycles it takes the system to reach a stable state. Legal reasoning, in which jurors attempt to construct a coherent account of the evidence (Pennington & Hastie, 1987), appears to be a particularly promising domain for future empirical tests of the ECHO model.

REFERENCES

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. *Cognition, 9*, 117-123.

Halloun, I., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics, 53*, 1056-1065.

Harman, G. (1986). *Change in View*. Cambridge, MA: MIT Press.

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*, 636-649.

Pennington, N., & Hastie, R. (1987). Explanation-based decision making. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, 682-690.

Ranney, M. (1987a). *Changing Naive Conceptions of Motion*. Doctoral dissertation, University of Pittsburgh, Learning Research and Development Center.

Ranney, M. (1987b, April). *Restructuring Conceptions of Motion in Physics-Naive Students*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Rumelhart D., & McClelland, J. (Eds.). (1986). *Parallel Distributed Processing*. (Vols. 1 & 2). Cambridge, MA: MIT Press.

Thagard, P. (1988a). *Explanatory coherence*. Princeton University Cognitive Science Laboratory Technical Report 16. Princeton, NJ.

Thagard, P. (1988b). *Computational Philosophy of Science*. Cambridge, MA: MIT Press/Bradford Books.

Thagard, P., & Nowak, G. (1988). *The explanatory coherence of continental drift*. Manuscript submitted for publication.

Wertheimer, M. (1945). *Productive Thinking*. New York: Harper.