To appear in the working notes of the IJCAI-95 workshop *Connectionist-Symbolic Integration: From Unified to Hybrid Approaches*, Montreal, Canada, August 1995.

A Hybrid Learning Model of Abductive Reasoning[1]

Todd R. Johnson
The Ohio State University
Division of Medical Informatics and
Center for Cognitive Science
2015 Neil Ave. Room 395
Columbus, Ohio 43210
tj@medinfo.ohio-state.edu

Jiajie Zhang
The Ohio State University
Department of Psychology and
Center for Cognitive Science
1827 Neil Ave. Room 200C
Columbus, Ohio 43210
zhang@canyon.psy.ohio-state.edu

Abstract

Multicausal abductive tasks appear to have deliberate and implicit components: people generate and modify explanations using a series of recognizable steps, but these steps appear to be guided by an implicit hypothesis evaluation process. This paper proposes a hybrid symbolic-connectionist learning architecture for multicausal abduction. The architecture tightly integrates a symbolic Soar model for generating and modifying hypotheses with Echo, a connectionist model for evaluating hypotheses. The symbolic component uses knowledge compilation to quickly acquire general rules for generating and modifying hypotheses, and for making decisions based on the current best explanation. The connectionist component learns to provide better hypothesis evaluation by implicitly acquiring explanatory strengths based on the frequencies of events during problem solving.

1. Introduction

Abduction is the process of generating a best explanation for a set of observations. Symbolic models of abductive reasoning tend to be far too search-intensive (e.g., Peng & Reggia, 1990), whereas connectionist models (e.g., Thagard, 1989) have difficulty explaining higher level abductive reasoning, such as the generation and revision of explanatory hypotheses. This paper proposes a hybrid learning model for abduction that tightly integrates a symbolic Soar model (Laird, *et al.*, 1986; Newell, 1990) for forming and revising hypotheses with Echo, a connectionist model for evaluating explanations (Thagard, 1989). In this model, Soar's symbolic knowledge compilation mechanism, chunking, acquires rules for forming and revising hypotheses and for taking actions based on the evaluations of these hypotheses. Thus, chunking models the problem solver's shift from deliberate to automatic reasoning. To complement this, Echo learns to provide better hypothesis evaluations by acquiring explanatory strengths based on the frequencies of events from past experience. Since Echo does not have a learning mechanism, we propose to add a connectionist learning method that uses both the Hebbian and Rescorla-Wagner (1972) rules. This hybrid model is motivated and supported by experimental results from the literature.

2. Abduction

Abductive reasoning is the process of generating a best explanation for a set of observations (e.g., Pople, 1973; Peng & Reggia, 1990; Josephson & Josephson, 1994). An explanation is a relation between one or more hypotheses and the datum or data they account for. Consider the radar task, a tactical decision making task in which a person uses radar information to determine whether an approaching craft is friend or foe. In the radar task, the observations are the current set of information available about a craft, such as the type of craft (air, sea or land) , its speed and course, whether it has responded to a warning and so on. The explanations relate hypotheses about the nature and intent of the craft to the observations. For example, if an aircraft is heading toward your ship and is not on a commercial air route, you might consider two explanations: 1) the craft is hostile (a single hypothesis); or 2) the craft is a commercial plane that is off course (two hypotheses, commercial and off course). Your decision about what to do next depends, in part, on your confidence in these explanations. If your ship is in immediate danger and you feel that the hostile aircraft explanation is more likely, you might decide to take defensive action. On the other hand, if your confidence in the commercial aircraft explanation is much higher than the hostile aircraft explanation, then, despite

[1] This research is supported by Grant No. N00014-95-1-0241 from the Office of Naval Research, Cognitive Science Program.

the threat of immediate danger, you might decide to collect additional data or issue a warning.

3. Motivation for a Hybrid Model

The hybrid model proposed in this paper is motivated by the following observations and empirical results concerning the relationship between symbolic and connectionist processes and human abductive reasoning.

Symbolic knowledge compilation can learn explicit rules based on a single problem solving episode, but it cannot easily learn explanatory strengths from previous experience. Connectionist learning techniques can easily acquire explanatory strengths, but not explicit rules. For example, based on a single problem solving episode of identifying an aircraft as hostile, Soar's chunking can learn an explicit rule that says that under certain general conditions, it should consider a craft to be hostile, but even after many episodes there is no direct way for chunking to learn the probability (or some measure of certainty) that a craft is hostile, given the conditions in the rule. Papageorgiou & Carley (1993) noted this problem with their Soar model of a simplified radar detection task. In contrast, connectionist learning techniques can easily acquire and use such probabilistic knowledge because they are sensitive to the past frequency of events (e.g., Dennis, 1994; Gluck & Bower, 1988), but generally speaking connectionist techniques cannot easily learn explicit, complex rules. These differences suggest that a hybrid Soar/connectionist architecture in which Soar learns complex rules while the connectionist component acquires explanatory strengths might lead to significantly improved models of human learning for multicausal abductive tasks.

Modeling how people determine confidence in an abductive explanation is a difficult problem, because it involves a variety of factors that interact in ways that we do not yet completely understand. These factors include: explanatory breadth (the amount of the data explained), parsimony (some measure of simplicity), strength of explanatory relations, base rates, the reliability of each datum, knowledge of and confidence in alternative hypotheses and how thoroughly the explanation space has been searched for alternatives. Researchers in a number of fields have studied the use of these factors and their role in the overall evaluation process, but so far no clear picture has emerged and many contradictory results remain. For example, Chinn & Brewer (1993) and Shustack and Sternberg (1981) found that awareness of alternatives affects a person's belief in a hypothesis; however, Downing, Sternberg & Ross (1985) found that the strength of alternatives has little effect on a person's belief.

Symbolic models of hypothesis evaluation do not sufficiently account for human reasoning. Symbolic models of hypothesis evaluation have two major problems:

1. **Symbolic models have difficulty combining the factors needed to evaluate hypotheses.** For example, many symbolic models, including our own (Johnson & Smith, 1991), have been based on an evaluation function that considers the number of hypotheses in an explanation and the number of observations that are not explained by those hypotheses. This works for some situations, but for many others it is important to consider other factors, such as those listed above. Downing, Sternberg & Ross' (1985) research illustrates the difficulty of using simple linear combinations of factors to model hypothesis evaluation.
2. **Applying a symbolic evaluation function can lead to a combinatorial problem, because the set of hypotheses making up an explanation must be specified and evaluated as a group.** Symbolic abduction methods usually apply an evaluation function to potentially relevant explanations and then select the explanation with the highest value (e.g., Johnson & Smith, 1991; Peng & Reggia, 1990). If a person has generated *n* hypotheses to account for various portions of the observations, then there can be up to 2^n ways to combine these hypotheses into explanations. The evaluation function must be applied to each of these explanations. If we add even more data, the number of hypotheses can grow, increasing the number of important explanations that we need to consider. A symbolic model of evaluation would spend a great deal of time enumerating possibilities and then evaluating them. If a person were to employ such a method, their working memory would quickly become overloaded by the number of relevant combinations.

Echo's connectionist model of hypothesis evaluation addresses the problems with symbolic models. Echo is based on a theory of explanatory coherence (TEC) that proposes that people prefer explanations that best cohere with their beliefs. TEC defines coherence (and incoherence) in terms of principles that relate hypotheses and observations. For example, a hypothesis coheres with the observations that it explains and also with analogous explanations, but incoheres with hypotheses that provide alternative explanations. Echo implements these principles by representing propositions (hypotheses and observations) as units in a connectionist net, while explanatory relations of coherence and incoherence are represented by excitatory and inhibitory links, respectively. Weights on the links represent the strengths of explanatory relations, and additional data support units represent the reliability of data. Unit activations represent how well a unit coheres with other propositions in the net, a kind of confidence in a proposition. By synchronously updating activation values, Echo applies the principles of explanatory coherence in parallel. When the net settles, the activations at each unit represent the coherence (evaluation) for that proposition. This eliminates the problems with symbolic models as follows:

1. **Echo uses parallel constraint satisfaction to combine a wide range of factors in parallel.** These factors include: parsimony, explanatory strength, explanatory breadth, reliability of data, strength of alternatives, analogies, contradictory evidence and explanations that are themselves explained. No other theory of hypothesis evaluation combines as many factors.
2. **The set of hypotheses that form the best explanation emerges from Echo's connectionist net; potential combinations of hypotheses do not need to be deliberately enumerated.** Echo computes explanatory coherence for each hypothesis. The set of hypotheses that make up the best (most coherent) explanation emerges as a result of Echo's parallel constraint satisfaction process. For example, if the most coherent explanation is a commercial craft with a malfunctioning radio that is off-course, then the hypotheses "commercial craft", "malfunctioning radio", and "off-course" will all have high explanatory coherence and competing hypotheses will have low coherence.

Research in a number of domains support Echo as a model of human hypothesis evaluation. Thagard used Echo to model conceptual change in scientific discovery (Thagard, 1992a), jury decisions in murder cases (Thagard, 1989) and adversarial problem solving (Thagard, 1992b). Ranney and Thagard (1988) used Echo to model how beginning students solve physics problems and Miller and Read (1991) used Echo to study how people perceive social relationships. Read and Marcus-Newhall (1993) conducted one of the most detailed psychological investigations of Echo's predictions using a series of simple social situations in which people were asked to judge their belief in several explanatory hypotheses. Their research shows that Echo can predict both the hypothesis preferred by the subjects as well as the subjects' belief levels for each hypothesis.

Research on implicit acquisition and use of event frequencies supports the hybrid Soar/Echo architecture. When conditional probability and base rates of occurrence are presented explicitly in terms of numeric values, they are very difficult to learn and utilize (see Kahneman, Slovic & Tversky, 1982). However, when they are presented in terms of real events and occurrences, they can often be learned implicitly and used correctly (e.g., Christensen-Szalanski, & Bushyhead, 1981; Medin & Edelson, 1988). A number of studies indicate that the learning of frequency of occurrence is usually implicit (unconscious) and automatic. The Soar/Echo hybrid architecture is consistent with these results, because Echo appears to Soar as an opaque mechanism that automatically and constantly provides confidence values for hypotheses. Soar sees only the results of Echo's evaluations. It knows nothing about the acquisition or evaluation processes within Echo nor does it ever deliberately call upon Echo. Thus, to Soar, Echo is an implicit mechanism that supplies appropriate information based on the current situation and previous experience.

4. A Hybrid Learning Model of Abduction

The major architectural components of the hybrid Soar/Echo architecture are shown in Figure 1. It consists of 4 major components: 1) Soar for modeling deliberative problem solving (forming and revising hypotheses), interaction with the environment (through the perceptual and motor system), and symbolic rule learning; 2) Echo for evaluating hypotheses, updating beliefs and acquiring event frequencies; 3) a visual system that mediates between the external environment and central cognition; and 4) a motor system that translates motor intentions from central cognition into motor commands. We first give an overview of how these components interact and then describe each component in detail.

Let us consider how these components would interact during a radar detection task scenario. Information presented on the radar and other parts of the display flows from the environment, through the visual system and into Soar's working memory. This information is limited by the visual system's focus of attention. Objects near the center of attention show up with the most details, whereas objects far from the center might only register as a visual event. The symbolic problem solving knowledge encoded in Soar, uses this information to begin to build, in working memory, a situation model of the environment. This model contains the observations as well as hypotheses about the nature of the situation. This information is organized into a network in which nodes represent observations and hypotheses and links represent explanatory relations. This network is shared between Soar and Echo. As Soar constructs the network, the connectionist Echo component augments the network with additional links and nodes that are needed to support the connectionist belief updating process. Echo also provides link weights and activation values based on its long term memory of event frequencies. Once the network is fully specified, Echo updates the node activations, which become values in Soar's working memory that Soar interprets as belief values (or certainty factors). Soar then uses these belief values to decide what to do next. It might decide to accept a hypothesis, reject a hypothesis, discount one or more observations, collect additional data, or take some other action appropriate to the task. These actions can lead to changes in working memory and in the environment. As the environment changes, these changes flow back into Soar's working memory, where Soar uses them to update its situation model. Changes to the situation model are detected by Echo and result in updated activation values.

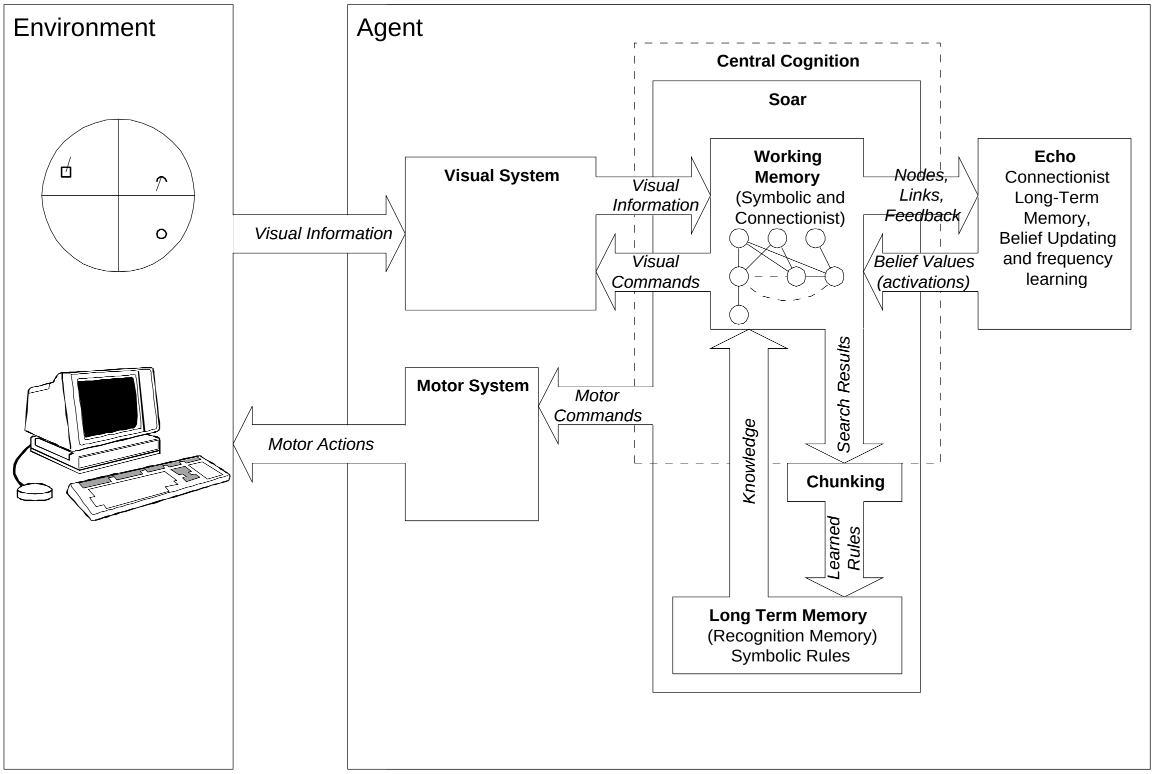

Figure 1: The major architectural components of the hybrid Soar/Echo learning architecture.

Soar and Echo interact dynamically on a very fine-grained time scale. Soar does not call Echo when it wants an evaluation, rather Echo is constantly watching working memory for changes and then updating the network based on these changes. From the perspective of the Soar theory of cognition, Echo is an implicit process for updating working memory.

Figure 2 illustrates how Soar and Echo are integrated. Soar operates by repeatedly executing decision cycles (DC). Echo is run just before Soar's elaboration phase (where knowledge is retrieved from long-term production memory). Within a single Soar DC, Echo will synchronously update the activations for each unit until either the net settles into a stable state or more than a specified number of synchronous cycles (not Soar DCs) have been run. Echo will then insert these activations into Soar's working memory. To Soar the activations will be indistinguishable from information produced by production rules. Echo is like a continuously running background process that is given time slices at the beginning of each decision phase. At the second DC, Echo continues to run, but it does so by taking into account any changes (deletions and/or additions) that Soar has made.

Learning in this model takes place in both Soar and Echo. As Soar reasons about what to do (both in terms of cognitive and external actions), it learns recognition knowledge that allows it to reason faster in similar future situations. This is done through Soar's learning mechanism, chunking, which compiles the results of search into new production rules for long term memory. Thus, chunking models the shift from slow, deliberative reasoning to quick, automatic reasoning. Event frequencies are acquired through Echo using a connectionist learning technique as described later. This technique depends, in part, on feedback from Soar concerning the correctness of the current hypotheses.

4.1 Echo

Echo is a connectionist implementation of Thagard's Theory of Explanatory Coherence (TEC). Thagard (1989) has proposed that people prefer the explanation that best coheres with their beliefs. TEC specifies seven principles that define coherence (and incoherence). These principles are implemented in a connectionist model, called Echo, that combines the principles in parallel to determine the coherence of hypotheses. Unlike symbolic theories of multicausal hypothesis evaluation, Echo does not provide an evaluation of complete multicausal hypotheses, nor does it require these to be enumerated and represented. Instead, Echo evaluates each individual hypothesis that has been generated to account for some of the observations. The hypotheses with the highest activations represent the most coherent hypotheses for the observations. If these hypotheses are part of the same explanation, then they represent the current best explanation; however, if they are part of competing explanations, they represent the best alternative explanations.

The primary problem with Echo as a process model as many researchers have noted (see open peer commentary in (Thagard, 1989)) is that it assumes that humans have the

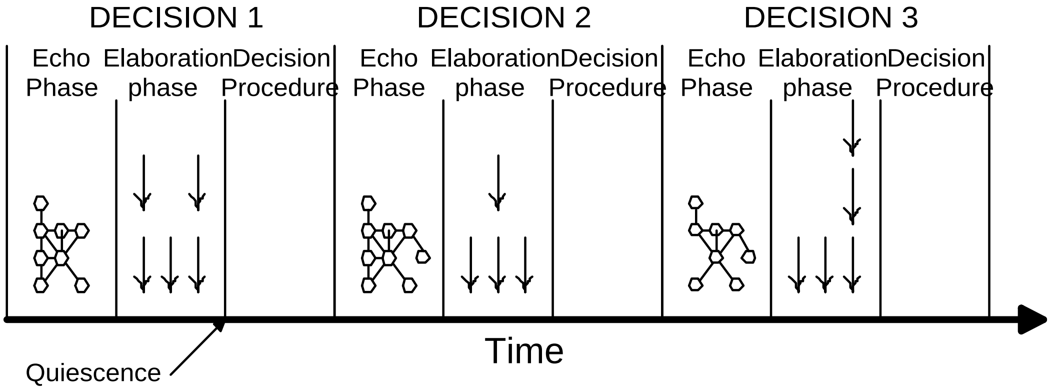

Figure 2: Echo is interleaved with Soar's decision cycle. By placing Echo just before the elaboration phase, the belief values produced by Echo appear to Soar as implicit information, automatically flowing into working memory from an impenetrable source. The diagram illustrates how the network can change from decision cycle to decision cycle.

ability to consider all of the interdependencies between an explanation and the available data. O'Rorke (1989) has also pointed out that Echo models abduction as an automatic unconscious process whereas most complex abductive tasks appear to require conscious deliberative reasoning. We believe that these problems can be avoided by integrating Echo with Soar. Most of the criticism of Echo is based on Thagard's models in which he first builds a network containing all relevant observations and hypotheses and then runs the network once. By combining Echo and Soar, our model combines deliberate abductive reasoning (generating and revising hypotheses and explanatory links) with Echo's automatic hypothesis evaluation. As we noted earlier, a number of experiments support this model.

4.2 Learning in Echo

Echo itself does not have a learning algorithm. Its weights are pre-determined, not learned. In the hybrid model, we propose to augment Echo with both the Rescorla-Wagner (1972) rule and the Hebbian rule. The Rescorla-Wagner rule is used to change the weights between observations and hypotheses because it is good at learning the associations between observation and hypothesis units in terms of conditional probabilities. The Hebbian rule is used to change the weights between hypotheses because it is good at learning the associations between hypothesis units in terms of correlations.

We selected these learning rules based on our modeling goals and on the desired interaction between Soar and Echo. The modeling goals demand a learning model that can acquire frequencies without a large number of supervised training trials and that can exhibit human behavioral regularities. In addition, we assume that the connectionist component could pick up information about correct and incorrect explanations from Soar's WM, but nothing else. Rescorla-Wagner and Hebbian learning is consistent with all of these requirements. In addition, each learning rule plays a different and important functional role in causal reasoning: Rescorla-Wagner for learning causal strength and Hebbian for learning strength of association between concepts.

5. Conclusion

We are currently implementing an initial version of the hybrid Soar/Echo architecture. Our goal is to produce a general modeling tool that can be applied to tasks ranging from adversarial problem solving (such as the radar task) to diagnosis. We plan to evaluate and refine this architecture by comparing model and human behavior on the radar task. We also plan to explore several variations of the model, including some that are purely symbolic. With this methodology, we can begin to identify the kinds of tasks that are most appropriate for either symbolic or connectionist learning techniques and determine potential trade-offs between hybrid and pure symbolic models.

References

Chinn, C. A. & Brewer, W. F. (1993). Factors that influence how people respond to anomalous data. *In Proc. of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 318-323). Hillsdale, NJ: Lawrence Erlbaum.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7* (4), 928-935.

Dennis, S. J. (1994). *The integration of learning into models of human memory*. Ph.D. Dissertation. Dept. of Computer Science, University of Queensland.

Downing, C. J., Sternberg, R. J. & Ross, B. H. (1985). Multicausal inference: Evaluation of evidence in causally complex situations. *Journal of Experimental Psychology: General*, 114(2), 239-263.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117* (3), 227-247.

Johnson, T. R. & Smith, J. W. (1991). A framework for opportunistic abductive strategies. *In Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, (pp. 760-764). Chicago: Lawrence Erlbaum Associates.

Josephson, J. R. & Josephson, S. G. (Ed.). (1994*). Abductive Inference: Computation, Philosophy, Technology*. Cambridge University Press.

Laird, J., Rosenbloom, P. & Newell, A. (1986) *Universal Subgoaling and Chunking*. Kluwer Academic Publishers.

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

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information form experience. *Journal of Experimental Psychology: General, 117* (1), 68-85.

Newell, A. (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Papageorgiou, C. P. & Carley, K. (1993). *A cognitive model of decision making: Chunking and radar detection task* (Technical Report No. 93.45). Carnegie Mellon University.

Peng, Y. & Reggia, J. A. (1990) *Abductive Inference Models for Diagnostic Problem-Solving*. New York: Springer-Verlag.

Pople, H. (1973). On the mechanization of abductive logic. *In Proc. of the International Joint Conference on Artificial Intelligence* (pp. 147-152). IJCAI.

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

Read, S. J. & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3), 429-447.

Rescorla, R. A., & Wager, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W.

Shustack, M. W. & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. Journal of *Experimental Psychology: General*, 110(1), 101-120.

Thagard, P. (1989). Explanatory Coherence. *Behavioral and Brain Sciences*, 12, 435-502.

Thagard, P. (1992a) *Conceptual Revolutions*. Princeton, New Jersey: Princeton University Press.

Thagard, P. (1992b). Adversarial problem solving: Modeling an opponent using explanatory coherence. *Cognitive Science*, 16(1), 123-149.