A Constraint Satisfaction Model of Causal Learning and Reasoning

York Hagmayer (york.hagmayer@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen
Gosslerstr. 14, 37073 Göttingen, Germany

Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

Following up on previous work by Thagard (1989, 2000) we have developed a connectionist constraint satisfaction model which aims at capturing a wide variety of tasks involving causal cognitions, including causal reasoning, learning, hypothesis testing, and prediction. We will show that this model predicts a number of recent findings, including asymmetries of blocking, and asymmetries of sensitivity to structural implications of causal models in explicit versus implicit tasks.

Introduction

Causal reasoning has been widely investigated during the last decade, which has led to a number of interesting novel findings (see Shanks, Holyoak, & Medin, 1996; Hagmayer & Waldmann, 2001, for overviews). For example, it has been shown that participants' causal judgments are sensitive to the contingency between the cause and the effect, and that people's judgments reflect the causal models underlying the observed learning events (see Hagmayer & Waldmann, 2001; Waldmann, 1996). Moreover, causal reasoning has been studied in the context of a number of different tasks, such as learning, reasoning, categorization, or hypothesis testing.

Most psychological theories and computational models of causal learning and reasoning are rooted in two traditions. They are either based on associationistic or on probabilistic or Bayesian models (see Shanks et al., 1996; Thagard, 2000). Both kinds of models have been criticized. Associationistic learning networks have proven unable to capture the fundamental semantics of causal models because they are insensitive to the differences between learning events that represent causes versus effects (see Waldmann, 1996). By contrast, Bayesian networks are perfectly capable of representing causal models with links directed from causes to effects (see Pearl, 2000). However, although the goal of these networks is to reduce the complexity of purely probabilistic reasoning, realistic Bayesian models still require fairly complex computations, and they presuppose competencies in reasoning with numerical probabilities which seem unrealistic for untutored people (see Thagard, 2000, for a detailed critique of these models).

The aim of this paper is to introduce a more qualitatively oriented, connectionist constraint satisfaction model of causal reasoning and learning. Our model is inspired by Thagard's (2000) suggestion that constraint satisfaction models may qualitatively capture many insights underlying normative Bayesian network models in spite of the fact that constraint satisfaction model use computationally far simpler, and therefore psychologically more realistic processes. The model differs from standard associationist learning models (e.g., Rescorla & Wagner, 1972) in that it is capable of expressing basic differences between causal models. Our model embodies a uniform mechanism of learning and reasoning, which assesses the fit between data and causal models. This architecture allows us to model a wide range of different tasks within a unified model, which in the literature have so far been treated as separate, such as learning and hypothesis testing.

Constraint Satisfaction Models

Constraint satisfaction models (Thagard, 1989, 2000) aim at capturing qualitative aspects of reasoning. Their basic assumption is that people hold a set of interconnected beliefs. The beliefs pose constraints on each other, they either support each other, contradict each other, or are unrelated. Coherence between the beliefs can be achieved by processes which attempt to honor these constraints.

Within a constraint satisfaction model beliefs are represented as nodes which represent propositions (e.g., "A causes B"). The nodes are connected by symmetric relations. The numerical activation of the nodes indicates the strength of the belief in the proposition. A belief that is highly activated is held strongly, a belief that is negatively activated is rejected. The activation of a node depends on the activation of all other nodes with which it is connected. More precisely, the net input to a single node *j* from all other nodes *i* is defined as the weighted sum of the activation *a* of all related nodes (following Thagard, 1989, p.466, eq.5):

$$\mathrm{Net}_j = \sum_i w_{ij} a_i(t) \qquad (1)$$

The weights *w* represent the strength of the connection of the beliefs. In our simulations, they are generally pre-set to default values which are either positive or negative and remain constant throughout the simulation. At the beginning of the simulations, the activation of the nodes representing hypotheses are set to a low default value. However, nodes representing empirical evidence are connected to a special activation node whose activation remains constant at 1.0. This architecture allows us to capture the intuition that more faith is put into empirical evidence than into theoretical hypotheses (see Thagard, 1989). To update the activation in each

cycle of the simulation, first the net input net_j to each node is computed using Equation 1. Second the activation of all nodes is updated using the following equation (Thagard, 1989, p.446, eq.4):

$$a_j(t+1) = a_j(t)(1-\theta)+net_j(max-a_j(t)) \text{ if } net_j>0$$
$$= a_j(t)(1-\theta)+net_j(a_j(t)-min) \text{ otherwise.} \quad (2)$$

In Equation 2, θ is a decay parameter that decrements the activity of each node in every cycle, min represents the minimum activation (-1) and max the maximum activation (+1). The activations of all nodes are updated until a stable equilibrium is reached, which means that the activation of all nodes do no longer substantially change. To derive quantitative predictions it would be necessary to specify rules that map the final activations to different types of responses. This is an important goal which should be addressed in future research. In the present article we only derive ordinal, qualitative predictions from the model.

The Model

Following causal-model theory (Waldmann, 1996) we assume that people typically enter causal tasks with initial assumptions about the causal structure they are going to observe. Even though specific knowledge about causal relations may not always be available, people often bring to bear knowledge about abstract features of the models, such as the distinction between events that refer to potential causes and events that refer to potential effects. In virtually all psychological studies this information can be gleaned from the initial instructions and the materials (see Waldmann, 1996).

Figure 1 displays an example of how the model represents a causal model. The nodes represent either causal hypotheses or observable events. The causal hypothesis node at the top represents a structural causal hypothesis (H1), in this case the hypothesis that the three events e_1, e_2, x form a common-effect structure with e_1 and e_2 as the two alternative causes and x as the common effect. The two nodes on the middle level refer to the two causal relations H2 and H3 that are part of the common-effect model with two causes and a single effect. The nodes on the lowest level refer to all patterns of events that can be observed with three events (a dot represents "and"). On the left side, the nodes represent patterns of three events, in the middle pairs, and on the right side single events. Not only the present but also the corresponding absent events are represented within this model (for example ~x). The links connecting the nodes represent belief relations. Thus, they do not represent probabilities or causal relations as in Bayesian models. There are two different kinds of connections between the nodes. Solid lines indicate excitatory links, dashed lines inhibitory links. How are the connections defined? A connection is positive if the propositions support each other. For example, if all three events are present, the observation is in accordance with both hypotheses H2 and H3. This pattern might be observed if both e1 and e2 cause x. Therefore the evidence node e1.e2.x is positively connected to H2 and H3. In general, a hypothesis is positively connected to an evidence node if the events mentioned in the hypothesis are either all present or all absent. If this is not the case, that is if one of the relevant events specified in the hypothesis is absent, the link is assigned the negative default value. Exploratory studies have shown, that participants share a common intuition whether a certain pattern of events supports or contradicts a hypothesis (Hagmayer & Waldmann, 2001). The assigned weights mirror these general intuitions. The weights of the links remain the same throughout the simulations. Figure 1 does not display the special activation node. This node was pre-set to 1.0 and attached to event nodes describing present events in the respective experiment.

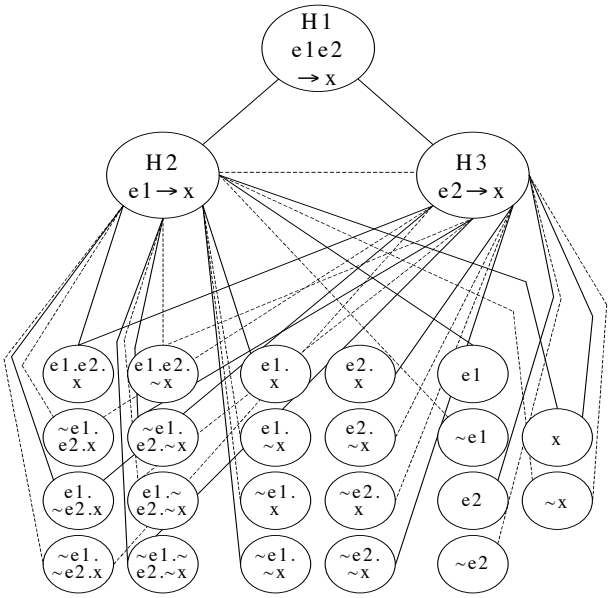

Figure 1: Constraint satisfaction model of causal learning and reasoning. See text for further explanations.

In Figure 1, the dashed line between the hypotheses H1 and H2, which signifies an inhibitory link, is of special interest. The network represents a common-effect structure. This means that there are two causes e1 and e2 which compete in explaining the occurrence of effect x. Therefore the two hypotheses referring to the individual causal relations have to be connected by a inhibitory link (see also Thagard, 2000). However, both hypotheses H2 and H3 are positively connected to the structural hypothesis H1. By contrast, a common-cause structure is represented slightly differently. In such a structure, event x would be the common cause of the two effects e1 and e2 (i.e., H1: x→e1.e2). A model of this structure looks almost identical to the one for the common-effect structure in Figure 1. There is only one very important difference. Because there is no competition between the effects of a common cause, a common-cause model has no inhibitory link between H2 and H3. All other nodes and links in the two models are identical.

Both the common-effect and the common-cause model were implemented using Microsoft Excel. Default values were adopted from the literature if not indicated otherwise (Thagard, 1989). Initial activations were set to 0.01, inhibitory links between nodes to –0.05, and excitatory links to +0.05. The inhibitory link between H1 and H2 within the common-effect model was pre-set to a value of –0.20. The

special activation node was attached to all evidence nodes. The additional activation was divided among the evidence nodes according to the relative frequency of the evidence in the learning input. This principle captures the intuition that more faith is put into evidence that is observed more frequently.

Evaluation

In order to evaluate the proposed constraint satisfaction model different tasks and paradigms from the literature on causal learning and reasoning were modeled. One of our main goals was to show that the same architecture can be used to simulate different types of tasks. However, different tasks required different sections of the model depicted in Figure 1. We used two principles for the construction of task specific networks. The first principle is that we only included the event nodes that corresponded to the event patterns observed in the learning phase or that corresponded to events that have to be evaluated or predicted in the test phase. For example, to model a task in which only event triples were shown, only the event nodes on the left side of the event layer in Figure 1 would be incorporated in the model. However, if the task following the learning phase required the prediction of single events, the corresponding nodes for single events would have to be added to the event layer. The second principle is that only the hypothesis nodes were included that represent hypotheses that are given or suggested to participants. These two principles ensure that for each paradigm a minimally sufficient sub-model of the complete model is instantiated.

Test 1: Asymmetries of Blocking

Blocking belongs to the central phenomena observed in associative learning which, among other findings, have motivated learning rules that embody cue competition (e.g., Rescorla & Wagner, 1972). A typical blocking experiment consists of two learning phases. In Phase 1 participants learn that two events e1 and x are either both present or absent. In Phase 2 a third event e2 is introduced. Now all three events are either present or absent. In both phases, events e1 and e2 represent cues and x the outcome to be predicted. Associative theories generally predict a blocking effect which means that participants should be reluctant about the causal status of the redundant event e2 that has been constantly paired with the predictive event e1 from Phase 1. This prediction has come under attack by recent findings that have shown that the blocking effect depends on the causal model learners bring to bear on the task (see Waldmann, 1996, 2000). If participants assume that e1 and e2 are the causes of x (common-effect structure) a blocking effect can be seen. In contrast, if participants assume that e1 and e2 are the collateral effects of the common cause x (common-cause structure), no blocking of e2 is observed. In this condition, learners tend to view both e1 and e2 as equally valid diagnostic cues of x.

To model blocking, we used a network that was extended after Phase 1. In Phase 1, the net consisted of a hypothesis node (H2) and the nodes for patterns of two events (e1, x). After Phase 1, the final activation of the hypothesis node was transferred to Phase 2. In Phase 2, the network consisted of two nodes for the two causal hypotheses (H2 and H3), and nodes that represented patterns of three events, the patterns participants observed within the learning phase. Furthermore, the node H1 was included, which, depending on the condition, either coded a common-cause or a common-effect hypothesis. The nodes for the event pairs from Phase 1 were removed.

Figure 2 shows the activation of the two hypotheses referring to the causal relations in Phase 1 and 2. Figure 2A depicts the activation for the common-cause model and Figure 2B for the common-effect model.

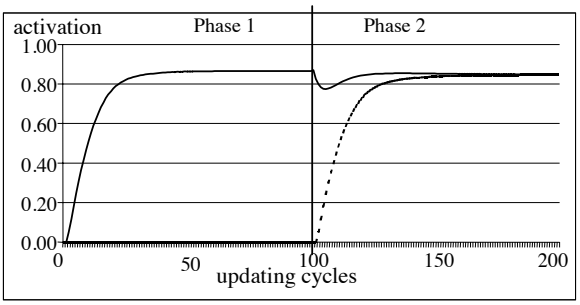

Figure 2A: Simulation of a blocking paradigm (Test 1). Activation of hypothesis nodes for a common-cause model. The solid line represents the activation of H2:x→e1, the dotted line of H3:x→e2. Phase 2 started at the 101st. cycle.

The model shows no blocking for event e2 in the context of the common-cause model. It quickly acquires the belief that there is a causal connection between x and e2.

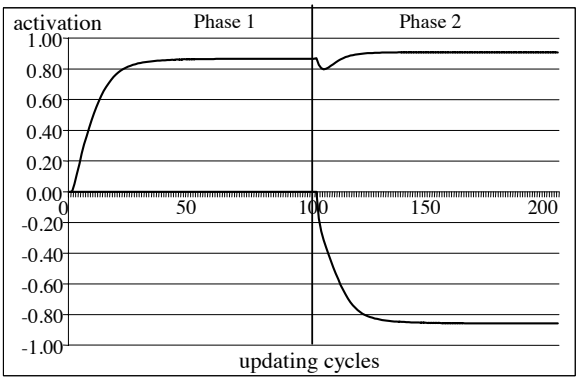

Figure 2B: Simulation of a blocking paradigm (Test 1). Activation of hypothesis nodes for a common-effect structure. The upper line represents the activation of H2:e1→x, the lower line of H3:e2→x. Phase 2 started at the 101st cycle.

For the common-effect model the simulation shows blocking of the second cause, that is the second hypothesis is believed to be wrong. Thus, the simulations closely correspond to the empirical finding that blocking interacts with the structure of the causal model used to interpret the learning data.

Test 2: Testing Complex Causal Hypotheses

The first test of the model used a phenomenon from the literature on causal learning. We now want to turn to a completely different paradigm, hypothesis testing. In experiments on causal learning participants are typically instructed about a causal structure, and the task is to learn about the causal relations within the structure. They are not asked whether they believe that the structure is supported by the learning data or not. In recent experiments (Hagmayer, 2001; Hagmayer & Waldmann, 2001) we gave participants the task to test a complex causal model hypothesis. For example, we asked them whether three observed events support a common-cause hypothesis or not. Normatively this task should be solved by testing the implications of the given structural hypothesis. For example, a common-cause model implies a (spurious) correlation of the effects of the single common cause. In contrast, a common-effect structure does not imply a correlation of the different causes of the joint effect. Unless there is an additional hidden event that causes a correlation among the causes, they should be uncorrelated. In the experiment, participants were given data which either displayed a correlation between all three events (data set 1) or correlations between e1-x and e2-x only, that is e1 and e2 were marginally independent in this data (data set 2). Data set 1 was consistent with a common-cause hypothesis which implies correlations between all three events. In contrast, data set 2 favors the common-effect hypothesis with x as the effect and e1 and e2 as independent causes. However, in a series of experiments we found that participants were not aware of these differential structural implications when testing the two hypotheses. Instead they checked whether the individual causal relations within the complex structures held (e.g., e1-x). Thus, participants dismissed a hypothesis if one of the assumed causal links was missing. However, they proved unable to distinguish between the common-cause and the common-effect structure when both structures specified causal connections between the same events (regardless of the direction).

To model this task we used the model without the nodes for event pairs and individual events. The special activation node was connected to the patterns of three events. As before the activation of the individual event patterns was proportional to the frequency of the respective pattern in the data. To test the model, we used three sets of data. Either all three events were correlated (data set 1), e1 and x, and e2 and x were correlated and e1 and e2 were marginally independent (data set 2), or e1 and x, and e1 and e2 were correlated, and e2 and x were uncorrelated (data set 3). As competing hypotheses we either used a common-cause model with x as the common cause, or a common-effect model with x as the common effect. Figure 3 shows the activation of the node H1 which represents the hypothesis that the respective causal model underlies the observed data.

Figure 3A shows the results for the common-cause hypothesis, Figure 3B for the common-effect hypothesis. The results clearly mirror the judgments of our participants. Whenever the two assumed causal relations within either causal model were represented in the data, the structural hypothesis was accepted (solid lines), if one link was missing the hypothesis was rejected (dotted line).

One slight deviation from our empirical findings was observed. In early cycles there seems to be an effect favoring the common-effect hypothesis with data consistent with this hypothesis. However, the difference between the hypotheses is relatively small and further decreases after 100 updating cycles. Thus, the results are consistent with participants' insensitivity to structural implications of causal models in hypothesis testing tasks.

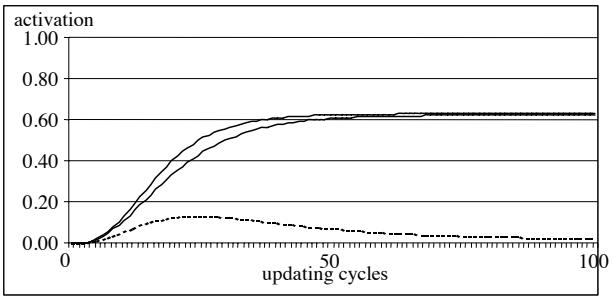

Figure 3A: Activation of hypothesis node H1 for a common-cause model (Test 2). The solid lines represent the activations for data set 1 and 2, the dotted line the activations for data set 3.

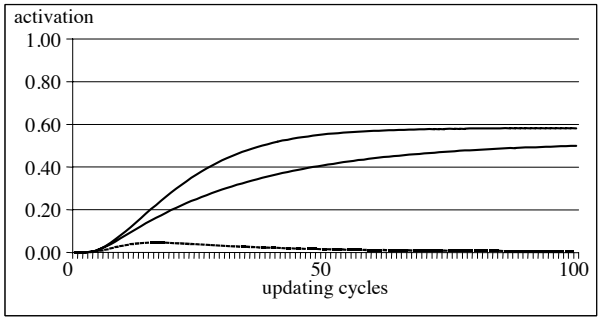

Figure 3B: Activation of hypothesis node H1 for a common-effect model (Test 2). The solid lines represent the activations for data set 1 and 2, the dashed line at the bottom the activations for data set 3

Why does the model not differentiate between the two causal structures? The reason is that it is assumed that complex structural hypotheses are not directly linked to empirical evidence. In our model empirical evidence is connected to the hypotheses that represent individual causal links which in turn are linked to more complex model-related hypotheses. This architecture allows it to model learning and hypothesis testing within the same model. It also seems to capture the empirical finding that participants can easily decide whether a certain pattern of events supports a simple causal hypothesis, but have a hard time to relate event patterns to complex causal hypotheses.

Test 3: Causal Inferences

In the previous section we have mentioned studies showing insensitivity to spurious relations implied by causal models. A last test for our model is a task in which participants have to predict other events under the assumption that a certain causal model holds. Interestingly we have empirically demonstrated sensitivity to structural implications of causal models in this more implicit task (Hagmayer & Waldmann, 2000). In this task participants do not have to evaluate the validity of a causal model in light of observed evidence but rather are instructed to use causal models when predicting individual events. In our experiments we presented participants with two learning phases in which they learned about two causal relations one at a time. Thus, in each phase participants only received information about the presence and absence of two events (x and e1, or x and e2). They never saw patterns of all three events during the experiment. The initial instructions described the two causal relations, which were identically presented across conditions, either as parts of a common-cause model with x as the cause or as part of a common-effect model with x as the effect. After participants had learned about the two causal relations we asked them to predict whether e1 and e2 were present given that x was present. We found that participants were more likely to predict that both e1 and e2 would co-occur when x was viewed as the common cause than when it was seen as a common effect. Thus, in this more implicit task the predictions expressed knowledge about structural implications of causal models. In particular, the patterns the participants predicted embodied a spurious correlation among the effects of a common cause, whereas the causes of a common effect tended to be marginally uncorrelated in the predicted patterns. By contrast, in a more direct task which required explicit judgments about correlations, no such sensitivity was observed, which is consistent with the results reported in the previous section.

To model this experiment we eventually used the complete network depicted in Figure 1 which was successively augmented according to our two principles. In Phase 1, the learning phase, patterns of two events were connected to the hypotheses H2 and H3. Depending on the learning condition, these two hypotheses were either linked to a common-cause or a common-effect hypothesis (H1). The activations of the hypothesis nodes at the end of Phase 1 were used as initial activation values in Phase 2. In Phase 2 the model consisted of the three hypothesis nodes, the nodes for patterns of three events and the nodes representing single events. The single event nodes were included because the task required the prediction of individual events. The special activation node was now attached to event x. The model then predicted the other two individual events and patterns of all three events.

The model quickly learned the causal relations during Phase 1 of the experiment. Figure 4 depicts the results of Phase 2. Figure 4A shows the predictions of the model for the condition in which participants assumed a common-cause model, Figure 4B shows the results for the common-effect condition. The results of the simulations are consistent with the behavior we have observed in our participants. When the model assumes a common-cause model the presence of x leads to a high positive activation of the two effects e1 and e2. This means that the model tends to prefer the prediction that the two effects of a common cause co-occur. In contrast, for the common-effect structure the model does not show such a preference. In this condition, both causes or either one of them equally qualify as possible explanations of the observed effect. This means that our model, similar to the one Thagard (2000) has proposed, tends to "explain away" the second cause when one of the competing causes is present. This is a consequence of the competition between the two causal hypothesis H2 and H3.

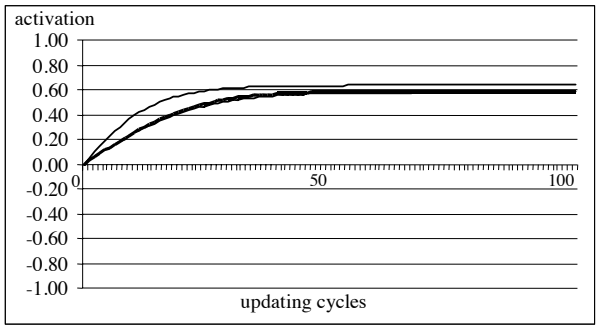

Figure 4A: Implicit causal inferences (Test 3). Activation of single event nodes for the common-cause model: Event x (top), events e1 and e2 (bottom)

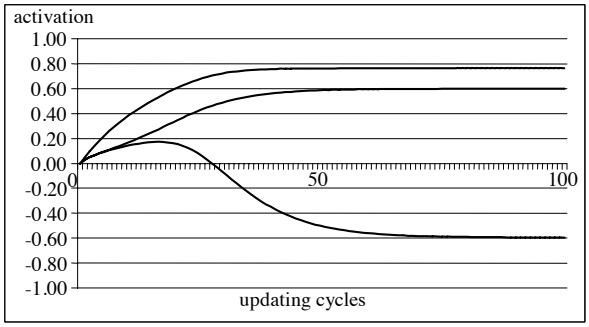

Figure 4B: Implicit causal inferences (Test 3). Activation of single event nodes for the common-effect model: Event x (top), event e1 (middle), event e2 (bottom)

Discussion

A constraint satisfaction model of causal learning and reasoning was presented in this paper that extends the architecture and scope of the model proposed by Thagard (2000). Thagard's model focuses upon causal explanations of singular events and belief updating. Our aim was to create a model that allows it to model both learning and reasoning within causal models. The model was successfully applied to three different tasks. It modeled people's sensitivity to structural implications of causal models in tasks involving learning and predictions whereas the same model also predicted that people would fail in tasks which required explicit knowledge of the statistical implications of causal models.

One question that might be raised is whether the proposed model really captures learning or just models causal judgment. In our view, the concept of learning does not necessarily imply incremental updating of associative weights Our model embodies a hypothesis testing approach to learning which assumes that learners modify the strength of belief in deterministic causal hypotheses based on probabilistic learning input. This view also underlies recent Bayesian models of causality (Pearl, 2000). In the model the activation (i.e., degree of belief) of the hypothesis nodes is modified based on the learning input. This way the model is capable of modeling trial-by-trial learning as well as learning based on summary data within the same architecture.

Thus far we have pre-set the weights connecting evidence and hypotheses. In our view, the assigned values reflect everyday qualitative intuitions about whether an event pattern supports or contradicts a hypothesized causal hypothesis. These weights remained constant throughout the simulations. Despite this restriction the model successfully predicted empirical phenomena in learning and reasoning. However, pre-setting these weights is not a necessary feature of the model. It is possible to add a learning component that acquires knowledge about the relation between event patterns and hypotheses based on feedback in a prior learning phase (see Wang et al., 1998, for a model adding associative learning to Echo).

In summary, our constraint satisfaction model seems to offer a promising new way to model causal learning and reasoning. It is capable of modeling phenomena in a wide range of different tasks, which thus far have been treated as separate in the literature. Relative to normative Bayesian models, our connectionist model allows it to simulate a large number of different tasks and different phenomena while using fairly simple computational routines. It proved capable of capturing a number of recent phenomena that have presented problems to extant models of causal cognition. More tests of the model clearly seem warranted.

References

Hagmayer, Y. (2001). *Denken mit und über Kausalmodelle*. Unpublished Doctoral Dissertation, University of Göttingen.

Hagmayer, Y., & Waldmann, M. R. (2000). Simulating causal models: The way to structural sensitivity. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 214-219). Mahwah, NJ: Erlbaum.

Hagmayer, Y., & Waldmann, M. R. (2001). Testing complex causal hypotheses. In M. May & U. Oestermeier (Eds.), *Interdisciplinary perspectives on causation* (pp. 59-80). Bern: Books on Demand.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II. Current research and theory* (pp. 64-99) New York: Appleton-Century-Crofts.

Shanks, D. R., Holyoak, K. J., & Medin, D. L. (Eds.)(1996). *The psychology of learning and motivation, Vol. 34: Causal learning*. San Diego: Academic Press.

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

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak & D. L. Medin (Eds.), *The psychology of learning and motivation, Vol. 34: Causal learning* (pp. 47-88). San Diego: Academic Press.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 53-76.

Wang, H., Johnson, T.R., & Zhang, J. (1998). UEcho: A model of uncertainty management in human abductive reasoning. In M. A. Gernsbacher & S. R. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1113-1118). Mahwah, NJ: Erlbaum.