Causal mechanisms as temporal bridges in a connectionist model of causal attribution

Michael E. Young
Dept. of Psychology
University of Minnesota
317 Elliott Hall
Minneapolis, MN 55455
young@turtle.psych.umn.edu

Brian DeBauche
Dept. of Philosophy
University of Minnesota
205 Elliott Hall
Minneapolis, MN 55455
debauche@text2.psych.umn.edu

Abstract

We use a connectionist model which relies on the encoding of temporal relationships among events to investigate the role of causal mechanisms in causal attribution. Mechanisms are encoded as intervening events with temporal extent that occur between the offset of a causal event and the onset of an effect. In one set of simulations, the presence of intervening events facilitated acquisition of a relationship between cause and effect via the mechanism. In a second set of simulations, prior experience with mechanisms enhanced development of a cause-effect relationship during later training absent the mechanism. The results provide evidence that causal mechanisms can facilitate causal attribution via Humean cues-to-causality.

Introduction

Our intuitive notion of causality leads us to believe that causes and effects are linked via some underlying physical mechanism. However, it becomes readily apparent that we do not *perceive* this mechanism in operation. When we see a billiard ball strike another causing it to move, we do not *see* energy passing from one to the other. Psychological theories which rely on our intuitive notions of causality (e.g. White, 1988) focus on adult introspection as a valid indicator of our cognitive processes. In this paper, we wish to explore the *foundations* of our causal attributions, relying upon the most objective of perceptual information. This motivates a focus on learning and on basic object and event attributes. This contrasts with a focus on the precise scientific basis for causal relationships.

Singularists believe single-observation causal attributions to be paradigmatic. For the singularist the subject learns *causal principles or rules* which are applied to current perceptions. The rule of *generative transmission* or *causal mechanism* is argued to be primary in making causal judgments (e.g. Bullock, Gelman & Baillargeon, 1982; Shultz, 1982; Shultz & Kestenbaum, 1985). This rule presupposes the presence of an objective, knowable causal mechanism (Harré & Madden, 1975) to mediate between the proposed cause and effect. Candidate mechanisms are identified as a result of prior experience with similar instances or by applying culturally transmitted knowledge.

Neo-Humean theorists require observation of various *cues-to-causality* to build knowledge of causation (Hume, 1739/1978). While it may often seem otherwise, both classes of theorists are relying on the subject *learning* the relations among events. For the neo-Humean the causal judger gathers empirical evidence, principally relying upon the cues-to-causality. These cues include temporal priority (causes must precede their effects), temporal contiguity (causes should occur near in time to their effects), spatial contiguity, and contingency or covariation (causes consistently precede an effect and do not occur in its absence) (see Einhorn & Hogarth, 1986). When two events possess all of these cues, a causal association from the earlier to the later is made. When any of the cues are absent this detracts from the certainty of a causal attribution (e.g. Koslowski & Okagaki, 1986; Siegler & Liebert, 1974; for a review, see Shultz & Kestenbaum, 1985). The cues are differentially weighted in their importance to the causal judgment (e.g. Einhorn & Hogarth, 1986). Singularist arguments against the Humean position involve causal judgments made after only one observation, demonstrating the non-necessity of covariation. Yet Hume (1739/1978) offered an explanation: "...this difficulty will vanish, if we consider, that tho' we are here suppos'd to have had only one experiment of a particular effect, yet we have many millions to convince us of this principle; that like objects, plac'd in like circumstances, will always produce like effects...." (p. 105). When the events are sufficiently similar to known causally related events, we will readily (i.e. after a single observation) infer causality.

This research supported in part by the Center for Research in Learning, Perception and Cognition and the National Institute of Child Health and Human Development (HD-07151).

The Role of Causal Mechanisms

For neo-Humeans it is necessary to explain causal ascriptions in the absence of cues. Lack of contingency can be overcome by a subject positing unknown or unobserved factors mediating the cause/effect relationship. Lack of temporal or spatial contiguity can be explained by hypothesizing the presence of an intervening event (with temporal duration) or object/force (which has spatial extent) to bridge the gap. This intervening event or force then becomes the causal mechanism. For a singularist the mechanism provides a medium for transmission of physical energy while a neo-Humean argues that, due to the mechanism, a series of events is perceived each of which is spatially and temporally contiguous with the subsequent event in a causal sequence.

Causal mechanisms are used as explanations for a lack of spatial or temporal contiguity between a (previously attributed) cause and an effect. They are also used as explanations when there is spatial and temporal contiguity but specific cause-effect mechanisms are unknown, as when an appeal is made to a lower level of analysis (e.g. the subatomic). Despite the evidence (via the cues-to-causality) which suggests that two events are causally related, the lack of an adequate explanation may indicate that the two events are merely correlated. However, if the two events are readily connected by a sufficient explanatory construct, then less evidence (e.g. covariation data) is necessary. Note that even in the presence of sufficient mechanisms, absence of support for a causal judgment from the cues-to-causality results in a significant negative impact on the likelihood of attribution (Koslowski, et al, 1989). Thus, while the cues-to-causality *may* be insufficient for causal attribution by adults, they are necessary.

One of the problems with the causal mechanism rule as defined by the singularist is the lack of specification regarding its source. There are many occasions under which knowledge of causal mechanism is not necessary, e.g. operation of a light switch (mechanism is electricity), operation of a remote control (mechanism is some sort of electric beam - even we don't know the real mechanism), and starting a car (how *does* that work anyway?). Under circumstances like these causal attributions are readily made and mechanisms are frequently manufactured as *post facto* explanations by individuals (who are frequently wrong, if my personal experience is any indication). Three year old children show no such concern for identifying a causal mechanism (Bullock, 1984; Shultz & Mendelson, 1975). When 4-5 year olds in Bullock's study discovered that the obvious causal mechanism was absent, they did not change their causal attribution in the absence of the mechanism but instead 77% of them hypothesized alternative mechanisms, from magnets (to bridge a spatial gap) to invisible strings and magic.

In addition to their explanatory role, causal mechanisms may also serve a facilitatory role during the early learning process. This role is suggested by the human and animal conditioning literature. Shanks & Dickinson (1987) and Wasserman (1987) suggest that causal judgments are built on the basic processes of learning as exemplified by classical and instrumental conditioning. Similar factors affect both conditioning and causal attribution (most notably the cues-to-causality). A temporal gap between cause (conditioned stimulus) and effect (unconditioned stimulus) makes learning the cause-effect relationship more difficult, but filling the temporal gap with intervening events, thus establishing an uninterrupted chain of contiguous events, facilitates learning (e.g. Kehoe, 1979; Reed, 1992). These intervening events may serve the same role that causal mechanisms do in causality judgment. That this facilitatory effect is observed in animals suggests that this role of "mechanism" may not be based on prior knowledge of how mechanisms operate but rather may be rooted in the conditioning processes underlying learning.

A Connectionist Model of Causality

Connectionist models manifest some of the principles espoused by the suggested conditioning-causal attribution homology. We present evidence that intervening events for a connectionist model demonstrate the explanatory and facilitatory properties ascribed to causal mechanisms. This is done without any hypothesis of energy transmission. The present model relies heavily on the temporal cues-to-causality: temporal contiguity and temporal priority (covariation can also be processed by the system but is not varied in the present simulations). The system learns the temporal relationships among events and predicts consequents from presentations of antecedents. An event "causes" another if its presence predicts the occurrence of the second. There is a strict reliance on temporal priority as a causal cue. Temporal contiguity is important for the model in that the longer the time interval between onset and offset of a cause's occurrence and the onset of a effect, the longer it will take the system to learn the relationship, if it learns it at all.

The architecture of the model is based upon that presented by Elman (1990). At a given instance of time, it predicts future events from an internal memory of past events and current conditions. For the purposes of causal attribution, the model predicts effects from causes. The model's architecture was designed to accommodate data from the literature on conditioning and causality and was developed to test the proposal that conditioning processes may underlie causal attribution processes (Shanks & Dickinson, 1987; Wasserman, 1987). The constraints placed upon the connections from the hidden units to the predictions of the causes

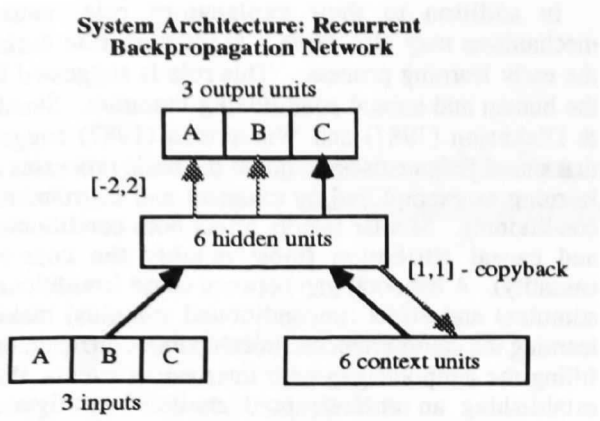

Figure 1: Architecture used for the Model of causal attribution. Weights represented by gray arrows are constrained to values within the range [minimum, maximum].

(not the final effect) are one technique for encouraging cue competition among the alternative predictors. Cue competition is a facet of conditioning, as exemplified by the well known phenomenon of blocking, and of causal attribution (e.g. Waldmann & Holyoak, 1992 on analogs to blocking and Einhorn & Hogarth, 1986 and the role of alternative causes).

Simulations of the model in a serial conditioning paradigm (Young, 1992) demonstrated that the presence of an intervening event (B) which fills a time gap between two events, A and C facilitates acquisition of the A->C relationship (i.e. expectation of C after A's occurrence develops faster following A->B->C experience than after A->gap->C experience). This facilitation has been demonstrated in human and animal conditioning experiments (e.g. Kehoe, 1979; Reed, 1992). Without the intervening event in the causal chain, the lack of temporal contiguity between the two events implies the absence of a causal relation. However, the presence of the event serves to fill the gap thus "explaining" the lack of contiguity and establishing a causal chain.

A replication of the facilitatory effect of an intervening event will be presented for a model that is a simplified version of that presented in Young (1992). The simulations involve training with much higher learning and momentum settings (.4 and .5 respectively) to increase the rate of learning without creating prediction instabilities. A second set of simulations will then be presented which demonstrate the effect that prior learning with a perceptible causal mechanism has on later experience in its absence (i.e. A->B->C trials followed by A->gap->C trials).

Simulation 1

A series of simulations were run using the recurrent architecture of Figure 1. This network differs from that presented in Young (1992) in that there are 6 hidden and context units here rather than 4, and there were 3 additional, constant inputs used in Young (1992) that played no role in those simulations and were not used in the present ones. Time is represented in an Elman (1990) network by using discrete time slices of a constant duration (one "time step"). Event occurrence is identified as an increase in activation for one of the network inputs (from 0 to 1). There are three orthogonal events used here (A, B and C) as input to the model, one input unit representing each event.

Seven architecturally equivalent networks with different random initial weights were used. These networks received two different forms of training. The *Mechanism* group was trained on repeated A->B->C event sequences. The last event in this and all subsequent simulations was one time step long. All other events were two time steps. All sequences of events were separated by a 20 time step intertrial interval during which all inputs to the network were 0. The *No-Mechanism* group received repeated A->gap->C event sequences, where the gap was 2 time steps long and consisted of presentation of 0's at all inputs.

A *trial* consisted of one presentation of the appropriate event sequence. Every 25 or 50 trials, the network's responses to a set of test events were recorded. During these test trials, no learning took place. The set of test stimuli consisted of three trial types: 1) presentation of A alone, 2) presentation of B alone, and 3) presentation of A->B (A followed by B). All test events were two time steps in duration. The network's expectancy of the occurrence of event C (the "effect") was the dependent measure of interest. *The expectancy is the value of an event's output node.* Unless otherwise noted, all figures and discussions will refer to the C expectancy at the appropriate time step, i.e. 4 time steps after A's onset or 2 time steps after B's onset.

Results and Discussion

C expectancies after presentation of A, B and A->B are plotted in Figure 2. The formation of the A->C relationship is stronger after A->gap->C training than after A->B->C training. The obvious reason for this difference is that the No-Mechanism group was trained to expect C following only A while the Mechanism group's experience suggests that the presence of B is expected as a forerunner to C. The C expectancy for the latter is significantly higher and develops much quicker after presentation of A->B, the sequence on which this group was trained. Note that the C expectancy following the series of stimuli is much greater than the sum of the expectancies following A or B individually. This demonstrates that the strong C expectancy following the A->B configuration is not carried solely by the contiguous B event but is the result of an interaction of the two (although it can be argued that the individual expectancies are sub-threshold and

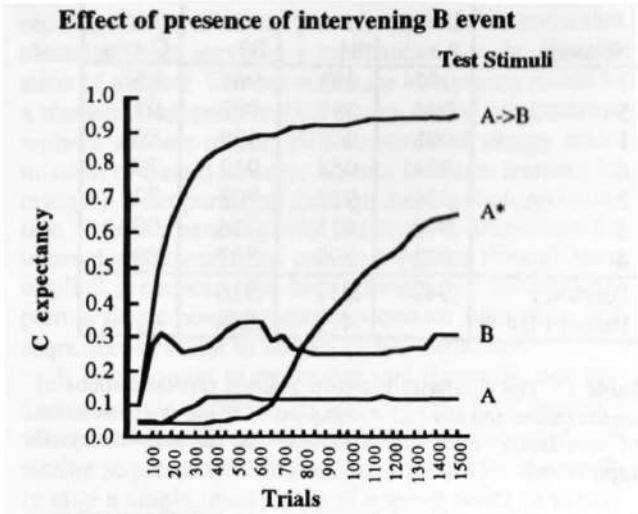

Figure 2: Growth of predictive relationship between earlier events (A, B or A->B) and the effect (C) as a function of experimental group (A* is for No-Mechanism group).

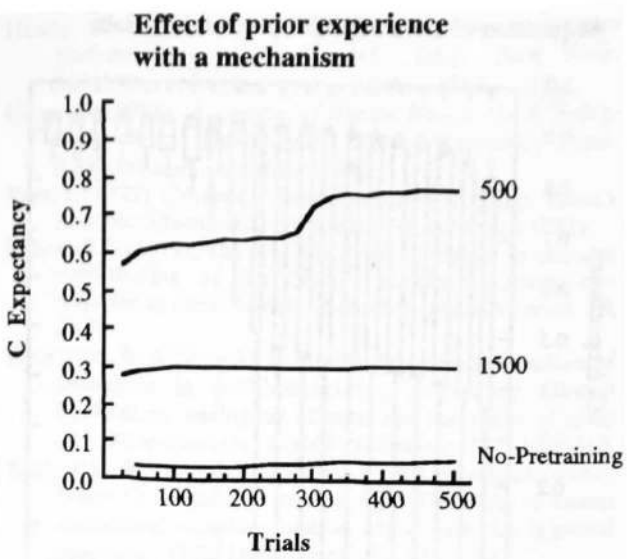

Figure 3: Comparison of effects of experience with A->B->C trials (none, 500, or 1500) on later learning of the A->C relationship during A->gap->C training.

their co-occurrence exceeds this threshold, resulting in the higher expectancy to the configuration (Kehoe, personal communication)). During early training the C expectancy following A->B is carried almost entirely by the contiguous B event, but as training progresses B facilitates the early development of the A->C relationship (compare A->gap->C). B provides an early bridge from A to C. However, this advantage disappears with time (after approximately 700 trials).

In the networks, an intervening event (our reification of a causal mechanism) can facilitate acquisition of a cause-effect relationship by bridging a temporal gap. This facilitation is limited to occasions on which the mechanism is present - the mechanism strengthens the A->C relationship via B, not in its absence. These results contrast with those found in Young (1992). In the latter, the intervening event facilitated development of A->C despite the lack of A->gap->C experience. For many of the networks the B played an early facilitatory role but contributed little later in training once the A->C relationship had been established. Informal simulations with the current model using lower learning rate and momentum values exhibited performance similar to those of Young (1992), demonstrating that the difference was due to parameter values and not the change in architecture. The second set of simulations investigated other potential benefits (or disadvantages) afforded by experience with intervening events.

Simulation 2

This set of simulations investigated the effect prior experience with a causal mechanism has on later training in its absence. This was investigated by training networks (the same networks used in the other simulations, i.e. same initial weight settings) on A->gap->C trials, but each group of networks had differing degrees of prior experience with A->B->C. Group *No-Pretraining* is identical to the A->gap->C group used in the earlier simulations; Groups *500* and *1500* had 500 and 1500 trials of A->B->C pretraining, respectively. Two groups were chosen to examine the effect that amount of experience with causal mechanism would have on later learning.

Results and Discussion

The basic results are presented in Figure 3. We first note that prior experience with a mechanism filling the gap facilitates the later acquisition of the A->C relationship during A->gap->C trials. The benefit provided by the pretraining is immediately apparent. After as few as 25 trials, the expectancy of C following A alone progressed from an initial median value of .058 (value at the end of pretraining) to a median value of .793 for Group 500. For Group 1500 results were mixed. Average improvement appears significant as evident in Figure 4, but this non-zero average was entirely the result of two networks who benefited from the 1500 trials of pretraining (C expectancy median at 25 trials was .843); the other 5 networks had C expectancies of zero. This trend continued: after 500 trials of A->gap->C the medians were .95 (for the two) and 0 (for the five). It is interesting to note that these two networks were the ones which began with the highest C expectancies to A alone at the end of 1500 trials (.683 and .089).

Figure 4 graphically depicts the special benefit afforded by prior mechanism experience for Group 500.

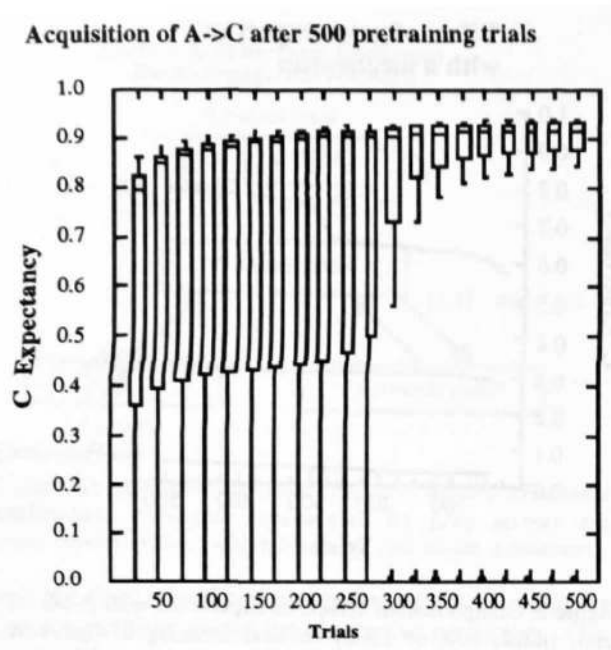

Figure 4: Strength of C expectancy following presentation of A alone. These networks were trained with A->B->C for 500 trials before receiving the graphed A->gap->C training.

Only one network failed to learn the A->C relation after 500 trials. Prior experience with an intervening event (the causal mechanism) has tremendous benefit for the group pretrained for 500 trials yet retards those networks receiving more of this apparently beneficial experience.

As a network attempts to learn an A->gap->C relationship, it is necessary for it to construct an internal, mediating representation that persists internally throughout the duration of the gap. A recurrent network of the type used here (Elman, 1990) bridges time by creating such intermediary representations. The prior experience with A->B->C provides such a mediating event explicitly. During the trials where the mechanism is now absent the network need only learn not to explicitly expect B (i.e. the output of the B node should not be appreciably above zero or an error will back-propagate through the network), while maintaining the internal representation of B as a mediating event. This unlearning requires changes in the weights between the hidden layer and the output layer only. Evidence for this use of B as an internal mediator is evident in the correlation between similarity in internal representations (hidden unit vector after pretraining and after 25 trials of A->gap->C) and the performance of the network. Table 1 shows the similarity between the internal representations (measured using a normalized dot product between hidden unit vectors) and the network's performance after 25 trials of A->gap->C (performance is captured as the strength of the A->C relationship).

The similarity in B internal representations (the two vectors represent the first and second time steps) was positively correlated with the network's success at

	Dot products			
Network	A	B1	B2	C exp.
7	.997	.993	.961	.857
5	.994	.984	.902	.837
1	.991	.908	.939	.806
6	.994	.964	.912	.793
2	.986	.912	.903	.725
3	.978	.917	.776	.002
4	.972	.899	.837	.001
Pearson r	.942	.611	.916	
Partial r B#.A		.475	.888	

Table 1: The similarity between internal representations of events before and after 25 A->gap->C trial and the correlation of similarity with performance (as measured by C expectancy).

learning the A->C relationship, i.e. prior experience with a mediating mechanism benefits later learning through the hypothesized existence of the unobserved event. This correlation held even after partialling out the effect of A's similarity. Given the advantages afforded later learning in the absence of the earlier causal mechanism, there should be greater facilitation when perceptual support for the presence of B exists (e.g. a low input activation value for the B unit rather than zero) and when the intervening event is similar to B (through generalization). The latter is especially interesting since there appears to be many occasions when the perceptual support for a causal mechanism consists of events or objects that are similar to but not identical to experienced mechanisms.

We suggest that overlearning explains why too much experience with a mechanism results in a benefit for only two of the seven networks. The networks appear to have so overlearned the A->B->C relationship that they lacked generalizability. This is not surprising in the context of causal attribution. It has been repeatedly demonstrated that spatial and temporal gaps in a launching paradigm (a rather overlearned attribution) have significant impact on subjects' reports of a causal relationship. These results suggest that training studies involving such gaps should be more effective in younger subjects than they would in older subjects. How young human subjects must be to behave like Group 500 rather than Group 1500 is an open question.

General Discussion

The results presented here suggest that encoding causal relations in the environment is reducible to representing the regularity of temporal sequences given that other cues-to-causality are present (e.g. spatial contiguity). There's a lot of debate on this notion of causality (see White, 1990). Our intuitions tell us otherwise, suggesting hypotheses similar to White's (1989) theory of causal powers. Our work is addressing the origins of

our knowledge of causal powers: how do we know that electricity can serve as a mechanism for the illumination of a light? Our networks use intervening events in a manner analogous to facilitatory causal mechanisms, without the use of singularist notions of energy transmission or causal powers. Events facilitate learning by bridging a temporal gap through their explicit presentation. After experience with mechanism sequences the internal representations can be effective through their implicit presence. An intervening event can also explain a single novel temporal sequence when the novel sequence is similar to known causal sequences.

It is important to remember that Hume did not suggest multiple experiences with a *particular* sequence were necessary for causal attribution. Experience with similar sequences is adequate for a judgment of causality after a single observation of a novel series of events. The singularist should differ with the neo-Humean primarily over the representational form prior experience takes and its function in generating current perceptions. For the singularist, prior experience is stored in symbolic rules with varying generality. No representational form is specified by Hume but prior experience impacts current perception via similarity and inference. The present neural network model encodes prior experience in a non-symbolic form, which gives the surface appearance of rule-like behavior to an outside observer.

The singularist view is premised on knowable causal relations (Harré & Madden, 1975), following Kant's view of nature as conforming to general laws (Kant, 1781/1929). The concept of causality advocated by Hume is counter-intuitive to many adults: causality is a psychological construct and has no provable physical instantiation. "It appears that in single instances of the operation of bodies we never can, by our utmost scrutiny, discover anything but one event following another, without being able to comprehend any force or power by which the cause operates...." (Hume, 1748/1955, pp. 84-85). While we do not deny the Kantian metaphysics, we suggest that people are Humean cognizers.

References

Bullock, M. (1984). Preschool children's understanding of causal connections. *British Journal of Developmental Psychology, 2*, 139-148.

Bullock, M., Gelman, R. & Baillargeon, R. (1982). The development of causal reasoning. In W. Friedman (Ed.), *The Developmental Psychology of Time*. New York: Academic Press.

Einhorn, H.J. & Hogarth, R.M. (1986). Judging probable cause. *Psychological Bulletin, 99*, 3-19.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Harré, R. & Madden, E.H. (1975). *Causal Powers: A Theory of Natural Necessity*. Oxford: Blackwell.

Hume, D. (1955). *An Inquiry Concerning Human Understanding*. (C.W. Hendel, Ed.). New York: Macmillan. (Original work published 1748).

Hume, D. (1978). *A Treatise of Human Nature*. (L. A. Selby-Bigge Ed.). New York: Oxford University Press. (Original work published 1739).

Kant, I. (1929). *Critique of Pure Reason* (N.K. Smith, Trans.). London: Macmillan. (Original work published 1781).

Kehoe, E.J. (1979). The role of CS-US contiguity in classical conditioning of the rabbit's nictitating membrane response to serial stimuli. *Learning and Motivation, 10*, 23-38.

Koslowski, B. & Okagaki, L. (1986). Non-Humean indices of causation in problem-solving situations: Causal mechanism, analogous effects, and the status of rival alternative accounts. *Child Development, 57*, 1100-1108.

Koslowski, B., Okagaki, L., Lorenz, C. & Umbach, D. (1989). When Covariation is not enough: The Role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development, 60*, 1316-1327.

Mackie, J.L. (1974). *The Cement of the Universe: A Study of Causation*. Oxford: Clarendon Press.

Reed, P. (1992) Effect of a signalled delay between an action and outcome on human judgement of causality. *The Quarterly Journal of Experimental Psychology, 44B*, 81-100.

Shanks, D.R. and Dickinson, A. (1987). Associative accounts of causality judgment. In G.H. Bower (Ed.), *The Psychology of Learning and Motivation, Vol. 21*. San Diego: Academic Press.

Shultz, T.R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47*(1, Serial No. 194).

Shultz, T.R. & Kestenbaum, N.R. (1985). Causal reasoning in children. *Annals of Child Development, 2*, 195-249.

Shultz, T.R. & Mendelson, R. (1975). The Use of covariation as a principle of causal analysis. *Child Development, 46*, 394-399.

Siegler, R.S. & Liebert, R.M. (1974). Effects of contiguity, regularity and age on children's causal inferences. *Developmental Psychology, 10*, 574-579.

Waldmann, M.R. & Holyoak, K.J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*, 222-236.

Wasserman, E. (1990). Detecting response-outcome relations: Toward an understanding of the causal structure of the environment. In G.H. Bower (Ed.), *The Psychology of Learning and Motivation, Vol. 26*. San Diego: Academic Press.

White, P.A. (1988). Causal processing: Origins and development. *Psychological Bulletin, 104*, 36-52.

White, P.A. (1989). A Theory of causal processing. *British Journal of Psychology, 80*, 431-454.

White, P.A. (1990). Ideas about causation in philosophy and psychology. *Psychological Bulletin, 108*, pp. 3-18.

Young, M.E. (1992). A Simple recurrent network model of serial conditioning: Implications for temporal event representation. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum.