COGNITIVE SCIENCE 15, 401–424 (1991)

A Computational Theory of Learning Causal Relationships

MICHAEL PAZZANI
University of California, Irvine

I present a cognitive model of the human ability to acquire causal relationships. I report on experimental evidence demonstrating that human learners acquire accurate causal relationships more rapidly when training examples are consistent with a general theory of causality. This article describes a learning process that uses a general theory of causality as background knowledge. The learning process, which I call *theory-driven learning* (TDL), hypothesizes causal relationships consistent both with observed data and the general theory of causality. TDL accounts for data on both the rate at which human learners acquire causal relationships, and the types of causal relationships they acquire. Experiments with TDL demonstrate the advantage of TDL for acquiring causal relationships over similarity-based approaches to learning: Fewer examples are required to learn an accurate relationship.

1. INTRODUCTION

There are many tasks that require an understander to reason about causal relationships. Indeed, it is hard to imagine how one could survive without the ability to reason about actions and their effects. To illustrate the importance, and ubiquity of causal reasoning, consider the following three key reasoning tasks:

- *Prediction:* Foreseeing what will happen if a balloon is pricked by a pin
- *Planning:* Specifying an action to achieve the goal of bursting a balloon; or, specifying an action to avoid if the balloon is to remain inflated
- *Abductive inference* (Peirce, 1932): Inferring what (unobserved) action(s) may have occurred to account for a balloon bursting.

I thank Cliff Brunk, Michael Dyer, Caroline Ehrlich, David Littman, David Ruby, and Tom Shultz for helpful comments regarding the organization, content, and style of this article. Comments by two anonymous referees helped to clarify ideas and improve the organization of the article. Mort Friedman provided facilities and encouragement for the experiments. Discussions with Keith Holyoak were useful in designing the experiments. This work is supported by National Science Foundation Grant IRI-8908260.

Correspondence and requests for reprints should be sent to Michael J. Pazzani, ICS Department, University of California, Irvine, CA 92717.

Because of the importance of causal reasoning, the primary goal in this article is to describe and demonstrate a computational learning procedure that accounts for the following critical fact: *Human learners acquire causal relationships more rapidly than would be expected if the learning mechanism relied solely on correlations between actions and state changes.* That is, a learner somehow brings to bear on the task of learning causal relationships knowledge that facilitates the acquisition of accurate causal relationships. This knowledge is called a theory of causality.

This article first defines the learning task in terms of the knowledge acquired by the learner and the theory of causality the learner starts with. Next, it is argued that human learning of causal relationships is facilitated by a theory of causality. A new learning procedure is introduced, called *theory-driven learning* (TDL), and compared to similarity-based and explanation-based approaches on the task of acquiring causal relationships.

1.1 A Theory of Causation

In order to predict state changes, the learner must acquire a *theory of causation.* A theory of causation is a collection of domain-specific causal relationships indicating the state changes that result from a particular class of actions. Each causal relationship consists of a description of a class of actions and a description of a class of state changes connected by *causal links* (Schank & Abelson, 1977). Causal relationships state things such as striking a balloon with a sharp object results in the balloon bursting.

The theory of causation enables the reasoner to predict the effects of actions, to generate plans that result in state changes, and to infer what actions may have occurred to explain state changes. For the purposes of these tasks, it does not matter how the causal relationships comprising the theory of causality are acquired. Any learning procedure, including a neural network-learning algorithm (e.g., (Rumelhart, Hinton, & Williams, 1986), could acquire the theory of causation by associating classes of state changes with classes of actions.

1.2 A Theory of Causality

A human learner brings to this learning task a set of domain-independent principles allowing the learner to conclude that a particular class of actions necessarily results in a state change (Bullock, Gelman, & Baillargeon, 1982; Shultz, 1982). This knowledge is called a theory of causality, and is distinguished from a theory of causation. A theory of causality indicates the conditions under which an action *appears* to result in a state change. For example, the theory of causality may include the condition that when an action on an object precedes a state change for that object, then the action appears to cause the state change. Note that the theory of causality is not able to predict what would happen if a balloon were poked with a pin. Rather,

when an example of a balloon bursting after being poked with a pin has been observed, the theory of causality is able to attribute the state change of the balloon to the fact that the balloon was poked with a pin, rather than an arbitrary action that may have occurred at the same time (e.g., a child eating a lollipop).

In contrast to a theory of causality, a theory of causation indicates the conditions under which an action results in a state change. For example, a reasoner's theory of causation may include a causal relationship indicating that striking a balloon with a sharp object results in the balloon bursting. This causal relationship can be used to infer that a particular red balloon will burst if it is poked with a pin. Note that for this action, a variety of state changes including the balloon catching fire and the pin shattering are consistent with the theory of causality, but not the theory of causation.

In the computer implementation, the theory of causality is represented by a set of *causal patterns.* A total of 30 causal patterns have been identified and implemented. Appendix A lists the patterns for physical causality. The causal patterns of TDL encode constraints that have been empirically determined to influence the acquisition of causal relationships. These constraints include:

- *Regularity:* Since a cause must necessarily result in an effect, the cause and the effect must co-occur (Shultz & Mendelson, 1975). Note that causality does not demand a perfect correlation.
- *Temporal order:* Children as young as 4 require a potential cause to precede an effect (Shultz & Mendelson, 1975). Although this may seem like a trivial constraint, existing learning systems (Lebowitz, 1986a; Salzberg, 1985) that predict the outcome of actions do not make use of temporal information.
- *Temporal contiguity:* An effect must immediately follow a cause (Michotte, 1963). When all other factors are equal, people select a cause closest in time to an effect.
- *Spatial contiguity:* An effect must be in contact with (or near) a cause (Bullock, 1979). When all other factors are equal, people select a cause that is closest in space to an effect.

The first constraint, regularity, is not explicitly represented by the causal patterns. Instead, the TDL procedure directly insures that causal relationships obey the regularity constraint. The remaining constraints are explicitly represented in the theory of causality. There are two ways that theory of causality constrains the search for a hypothesis:

- *Determining the true cause in an ambiguous situation.* For example, consider the following observation. First, two actions occur at the same time: Karen is eating a lollipop and Chris pokes a balloon with a pin.

Next, the balloon bursts. By ruling out eating the lollipop as a cause for the balloon bursting, the search space for the problem of determining what causes balloons to burst can be reduced.

- *Selecting relevant features.* A theory of causality can focus attention on the potentially relevant features of the objects involved in an action. For example, the features of the object that pokes a balloon, and the features of the balloon itself may determine whether or not the balloon bursts. However, it unlikely that the features of the person that pokes the balloon are significant.

1.3 The Learning Task

TDL's objective is to construct a theory of causation, given a theory of causality and a number of observations. The input to the TDL procedure is a sequence of observations. Each observation consists of several *actions* and *state changes* connected by temporal links. The learning task is complicated by the fact that several actions may occur at the same time. Therefore, the learner must be able to distinguish between actions that temporally preceded a state change, and actions that resulted in a state change, as follows:

GIVEN: 1. A series of observations.
2. A theory of causality (i.e., a set of causal patterns).

CREATE: A theory of causation (i.e., a set of causal relationships).

1.4 Learning Causal Relationships: An Example

In this section, an example of learning causal relationships is presented in order to provide an overview of TDL. The following is a protocol of Lynn (3 years, 11 months), trying to figure out when she can inflate balloons and when she cannot.

1. Mike is blowing up a red balloon.
2. Lynn: "Let me blow it up."
3. Mike lets the air out of the balloon and hands it to Lynn.
4. Lynn blows up the red balloon.
5. Lynn picks up a green balloon and tries to inflate it.
6. Lynn cannot inflate the green balloon.
7. Lynn puts down the green balloon and looks around.
8. Lynn: "How come they only gave us one red one?"
9. Mike: "Why do you want a red one?"
10. Lynn: "I can blow up the red ones."
11. Mike picks up a green balloon and inflates it.
12. Mike lets the air out of the green balloon; hands it to Lynn.
13. Mike: "Try this one."
14. Lynn blows up the green balloon.
15. Lynn gives Mike an uninflated blue balloon.
16. Lynn: "Here, let's do this one."

It appears from the first observation (lines 1–4), that Lynn has acquired a causal relationship indicating that she can inflate any balloon by blowing air into the balloon. A causal pattern that states *an action on a particular object followed by a state change of the object, suggests that the action results in the state change,* focuses TDL to hypothesize this same causal relationship.

After the second observation (lines 5–10), a counterexample to the initial hypothesis is seen and the learner must generate a new hypothesis to account for a different result. The two balloons differed in color and the hypothesis can be accounted for by a causal pattern that states *two actions that have different results and that are performed on different objects, suggests that a feature that differs between the two actions enables the action to produce the result.*

The second hypothesis is contradicted by the third observation (lines 10–16), when Lynn determines that the color of the balloon is not important. Instead, she attributes the different result to a different action that preceded her successfully inflating a balloon. In TDL, this hypothesis would be produced by a causal pattern that states *an initial action on an object preceding a subsequent action that precedes a state change for the object, suggests that the initial action results in a state change that enables the subsequent action to result in the state change.* In the next section, an experiment is reported, whose goal is to determine whether learning new causal relationships is facilitated when a causal relationship conforms to this causal pattern.

2. CONSTRAINTS ON LEARNING CAUSAL RELATIONSHIPS: EXPERIMENTAL RESULTS

The purpose of this experiment was to investigate how a theory of causality affects the number of trials required to learn to make accurate predictions. I investigated the last causal pattern from the previous section. This pattern postulates an intermediate enabling state, where a prior action on one object is present when a subsequent action on the same object results in a state change. Section 3.3 discusses this pattern in more detail (see Figure 6). It was predicted that it would take fewer trials for subjects to learn a new causal relationship conforming to this pattern than a similar causal relationship not conforming to this pattern.

One group of subjects had to learn that a child would be able to inflate a balloon only if she dipped the balloon in water before blowing air into it. Another group of subjects had to learn that a child would be able to inflate a balloon only if she snapped her fingers before blowing air into the balloon. The former relationship is consistent the causal pattern tested. The latter is not consistent with any causal pattern used by TDL.

To eliminate cue salience (Bower & Trabasso, 1968) as a possible explanation for the increased learning rate, I ran two control groups that performed a concept-identification task (Bruner, Goodnow, & Austin, 1956) rather

than a prediction task with the same stimuli. Instead of predicting whether a balloon could be inflated, the control groups had to associate the category name, "alpha," with the action sequence of dipping a balloon into water followed by blowing air into it (or snapping fingers followed by blowing air into balloon). To summarize, there were four conditions in this experiment:

- *Dip/Inflate:* This condition requires predicting that a balloon could be inflated only after it was dipped in water.
- *Snap/Inflate:* This condition requires predicting that a balloon could be inflated only after the actor first snaps her fingers.
- *Dip/Alpha:* This condition requires learning that examples of a person dipping a balloon in water and attempting to inflate a balloon belong to a category called "alpha."
- *Snap/Alpha:* This condition requires learning that examples of a person snapping her fingers and attempting to inflate a balloon belong to a category called "alpha."

It was predicated that the Dip/Inflate inflate condition would be easiest for subjects to learn. This is the only condition in which TDL is applicable. The subjects in the remaining conditions must rely solely on correlation because the theory of causality is either irrelevant to the task or contradicted by the training data.

2.1 Method

2.1.1 Subjects. The subjects were 80 male and female undergraduates attending the University of California, Los Angeles, who participated in this experiment in partial fulfillment of course requirements for an introductory psychology course. Each subject was tested individually. Subjects were randomly assigned to one of the four conditions.

2.1.2 Stimuli. The stimuli consisted of four videotapes, one for each condition. Each tape consisted of a series of observations (e.g., dipping a balloon in water and attempting to inflate the balloon). In the inflate conditions, each observation was followed by a continuation in which the balloon was either inflated successfully or not inflated. In the alpha conditions, each observation was followed by a display of the word *alpha* or the phrase *not alpha*. Action sequences differed according to the prior action (either the actor dipped a balloon in water, put a necklace on, or snapped her fingers), the color of the balloon (orange or yellow), or the size of the balloon (small or large). The sequential nature of videotape made it necessary to control the order of the presentation of observations: The dipping and snapping actions were interchanged between the dip and snap conditions. Appendix B contains the order in which examples were presented. The ordering insures

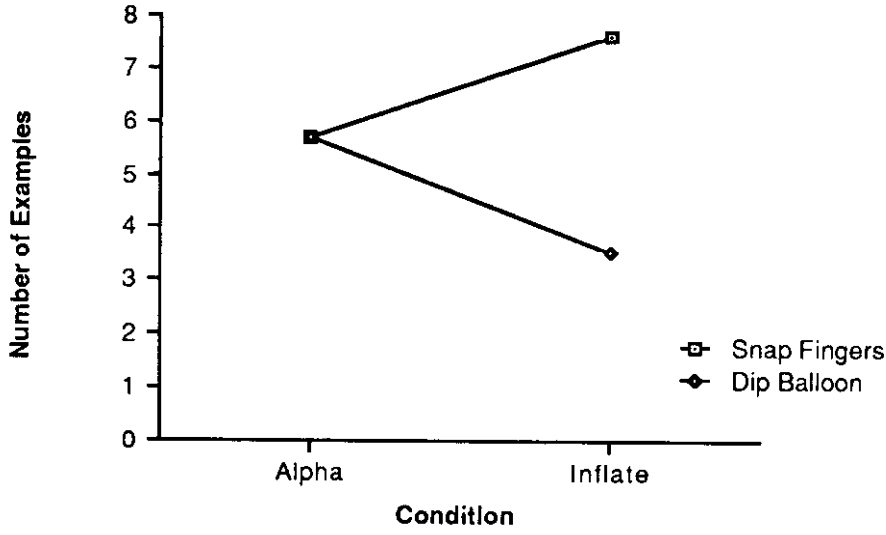

Figure 1. The results of the experiment indicate that subjects require fewer trials to learn a relationship that conforms to a common pattern of causal relationships.

that the subjects saw positive and negative examples in the same order. The observations in the alpha tapes were in the same order as the observations on the corresponding inflate tapes.

2.1.3 Procedures. Each subject was shown an observation on a color television screen. The tape was paused and the subject was asked to make a prediction. Then the tape was resumed and the subject was able to judge the accuracy of the prediction. This process was repeated until the subject was able to predict correctly on six consecutive observations. The number of the last trial on which the subject made an error was recorded.

2.2 Results

The result of this experiment (illustrated in Figure 1) confirmed the prediction, $F(3, 76) = 8.88$, $p < .05$. Subjects required fewer trials to learn to predict that a balloon that had been dipped in water could be inflated (3.5 trials) than to predict that a balloon could be inflated after the child snapped her fingers (7.6 trials). One possible explanation for this result is that dipping a balloon in water is perceptually more salient than snapping fingers. If this were the case, we would expect the same preference to hold when associating a name with an class of observations. However, subjects required approximately the same number of trials to determine that a balloon being dipped in water is an alpha (5.7 trials) and to determine that the child snapping her fingers is an alpha (5.9 trials).

2.3 Discussion

These results support the hypothesis that subjects first focus on relationships consistent with a general theory of causality. In the dip/inflate condition, subjects can ignore correlations between the size or color of the balloon and the result. In the snap/inflate condition, the observations of the subject do not conform to any causal pattern and the subject must consider the size and color of the balloon as well as the type action as a possible cause.

This experiment suggests that certain structural configurations of actions are cues for causal relationships. When observations conform to common patterns of causal relationships, it is easier to induce a causal relationship. Although there may be many regularities in the observed data, TDL focuses on those regularities likely to play a part in a causal relationship.

This experiment provides evidence for just one of the causal patterns in the computer implementation. The remaining causal patterns encode principles that have been empirically determined to influence the attribution of causality (see Section 1.2). In the next section, the process of TDL is described in more detail.

3. THEORY-DRIVEN LEARNING (TDL)

In this section, I describe the representation of observations, causal relationships, and causal patterns, and elaborate on the TDL process.

3.1 Observations

In the computer implementation of TDL, observations—the training data for TDL—are represented in conceptual dependency (Schank & Abelson, 1977). Figure 2 illustrates one sequence of actions and state changes. This figure contains the representation of two actions that occur at the same time: John is eating a Life Saver in the kitchen and John touches a red balloon with a pin. A state change occurs immediately after these two actions: The balloon bursts. Each action and state change is described by a number of roles (e.g., actor, object, type, to, etc). Roles are indicated by lowercase letters in figures. The values of roles, indicated in capital letters, may be simple objects (e.g., BROWN) or composite objects that have additional roles (e.g., PP color RED).

3.2 Causal Relationships

Conceptual dependency is also used to represent causal relationships. Causal links (Schank & Abelson, 1977) are used to specify the relationships between actions and states. The following causal links are used:

- An action can **result** in a state change.
- A state can **enable** an action to occur.

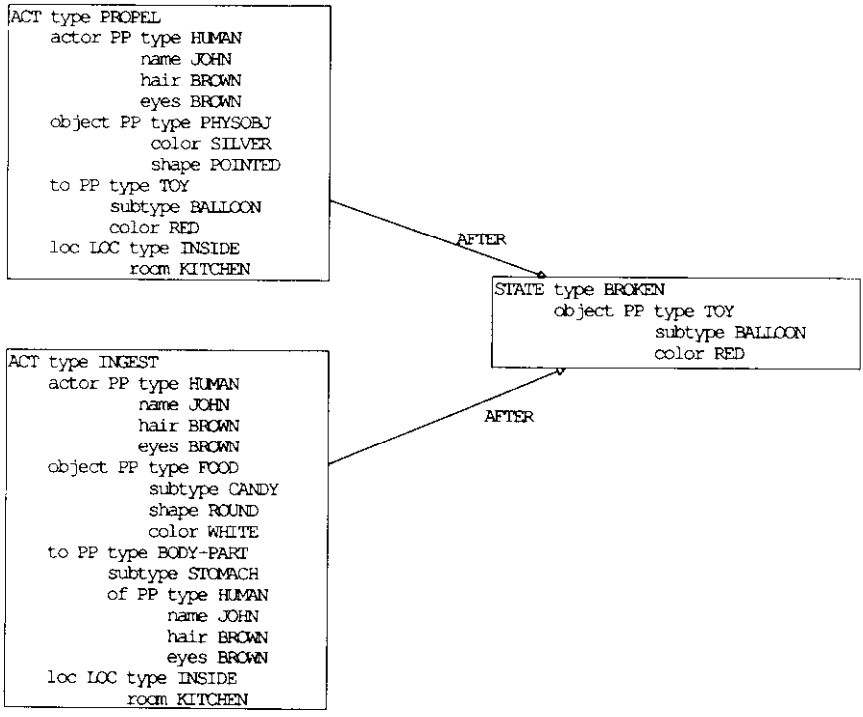

Figure 2. An example of an observation used as a training example for theory-driven learning. This is the Conceptual Dependency representation for: "John, who has brown hair and brown eyes is eating a Life Saver candy in the kitchen when he touches a red balloon with a pin. The balloon bursts." Role names are in lower case letters. Values of roles are capitalized.

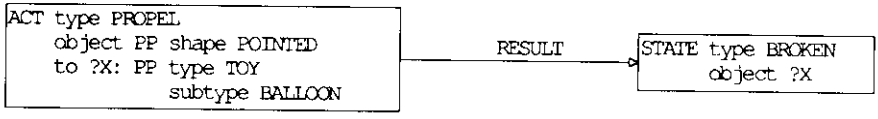

Figure 3. An example of a causal relationship that is acquired via theory-driven learning: Touching a balloon with a pointed object results in the balloon bursting. Variables are preceded by a question mark.

One causal relationship is illustrated in Figure 3. This relationship indicates that striking a balloon with a sharp object results in the balloon bursting. The variable in the relationship, preceded by a question mark in the figure, ensures that the balloon that is struck is identical to the balloon that bursts.

3.3 Causal Patterns

The TDL procedure can only learn causal relationshps that conform to one of the causal patterns comprising the theory of causality. Figure 4 illustrates

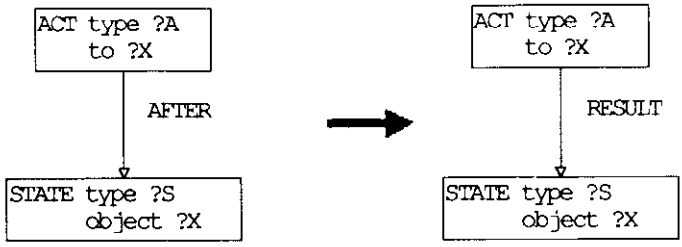

Figure 4. A causal pattern: An action with a particular destination (indicated by the variable ?X) followed by a state change for the destination suggests that the action results in the state change.

one causal pattern. A causal pattern matches an observation and proposes a causal relationship. This pattern states: *An action with a particular destination followed by a state change for the destination, suggests that the action results in the state change.* The antecedent (on the left side of the bold arrow) is matched against observations to produce a causal relationship (on the right side of the bold arrow). The bold arrow does not mean logical implication. It can be read as "suggests that."

Although the pattern in Figure 4 appears to be very simple, it encodes three important assumptions about causal relationships. First, it encodes the constraint that the destination of action (i.e., the object that fills the to role of an action) must be the object whose state has changed. This constraint would rule out a wide variety of arbitrary actions from being considered as potential causes. Second, it indicates that the only important roles of the action are the type of action and the object. The actor who performs the action, and the time that the action is performed are not relevant. Third, the pattern also contains the temporal ordering constraint, i.e., the action must precede the state change.

When this particular pattern is applied to the observation of the balloon being stuck with a pin by someone eating a Life Saver candy (see Figure 2), it results in the causal relationship that applying a force to a balloon results in the balloon bursting. Note that the causal relationship created by this pattern does not require that the object applying the force be pointed. Learning this additional constraint requires observations of balloons not bursting when being touched by blunt objects.

The causal pattern in Figure 4 is called an *exceptionless* causal pattern because it applies when similar actions result in the same state change. Other causal patterns focus on reasons why similar actions have different results. With the representations used for observations and causal relations, there are two reasons why similar actions have different results:

1. A role of the action differs. For example, the action may be performed on a different object or have a different destination.
2. A prior action is needed to change the state of an object so that a subsequent action may result in a state change.

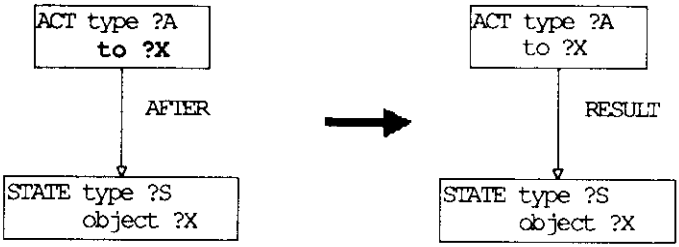

Figure 5. A dispositional causal pattern: Similar actions with different destinations followed by different state changes for the destination suggests that a role of the destination is required for an action to result in the state change. The bold type indicates that the search for a difference is constrained to the destination role.

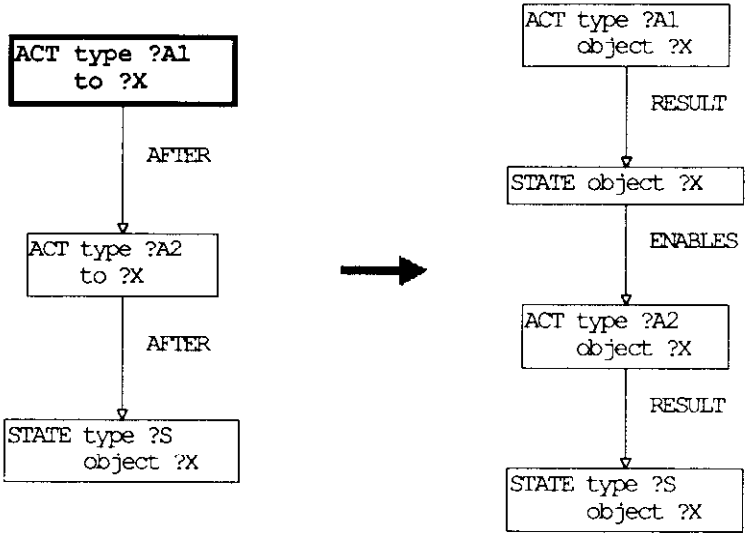

Figure 6. A historical causal pattern: An initial action with a destination that precedes a subsequent action with the same destination suggests that the initial actions results in an intermediate state change that enables the subsequent action to result in the final state change. The bold type for the initial action indicates that a different initial action will not result in the same final state change.

There is a separate type of causal pattern for each reason. For the first reason, the causal patterns are called *dispositional* causal patterns because they attribute a different result to differing dispositions (i.e., potential or capacities) of actors or objects. The causal patterns accounting for the second reason similar actions have different results are called *historical* causal patterns because they attribute a different result to different histories of the objects involved.

The protocol in Section 1.4 illustrates all three types of causal patterns. The first causal relationship (that the child can inflate all balloons) can be produced by the exceptionless pattern illustrated in Figure 4. The second

causal relationship (that the child can inflate red balloons) can be produced by the dispositional causal pattern illustrated in Figure 5. The fact that the search for a difference is constrained to differences in the destination is indicated in boldface for the destination in the figure. In the protocol in Section 1.4, the two destinations differ only in the color of the balloon. If there were several differences, it would be necessary to correlate roles of the destination over additional examples to discover a reliable causal relationship.

The historical pattern displayed in Figure 6 attributes the difference in a result to an initial action that results in a state enabling the second action to result in the state change. The fact that the search for a difference is constrained to differences in a prior action is indicated in boldface for the prior action in the figure. This pattern would find relationships such as that stretching a balloon results in a state the enables blowing air into the balloon to inflate the balloon. This causal pattern was tested in the experiment reported in Section 2.

The types of causal patterns are ordered by the simplicity of the causal relationship they create. Exceptionless patterns produce the simplest relationships. Causal relationships produced by dispositional patterns are more complex because they require additional conditions to be true for the action to have an effect (e.g., that the balloon be red). Historical patterns produce the most complex causal relationships (i.e., relationships that postulate unseen intermediate states).

3.4 The TDL Procedure

The TDL procedure has two functions. First, it ensures that a proposed causal relationship obeys the regularity principle. Second, if more than one causal pattern applies to a set of observations, it determines what causal relationship will be created. A current best hypothesis (Mitchell, 1982) for a causal relationship is created from one causal pattern rather than maintaining a set of consistent hypotheses (e.g., Mitchell, 1982; Vere, 1975). Psychological evidence (e.g., Bower & Trabasso, 1968; Levine, 1967) indicates that only one, or a small number of hypotheses, are considered at one time. Simplicity is the criteria used for selecting the best hypothesis.

In a given situation, more than one type of causal pattern may apply. For example, the exceptionless pattern in Figure 3, the dispositional pattern in Figure 6, and the historical pattern in Figure 7 (p. 415) all match the first observation of the learning task in Section 1.3. TDL orders the types of causal patterns by simplicity. Within each type of pattern, it is also possible that more than one pattern may match an observation. In this case, one pattern is arbitrarily chosen to create a causal relationship. If further observations prove the causal relationship to be inaccurate, the relationship will be discarded, and an alternative pattern may be applied in this situation.

Learning occurs whenever an unpredicted state change is observed. The following algorithm describes the learning procedure:

1. **Collect similar observations:** Observations with actions similar to the action of the current observation, are retrieved from memory (if there are any).[1]
2. **Partition observations:** Two sets of observations are created: The positive examples are those similar observations with the same state change as the new observation and the negative examples are those observations with different state changes (or no state change at all).
3. **Match observation and causal patterns:** If there are no observations with a different state change, the exceptionless patterns are applicable. Otherwise, dispositional or historical patterns might apply. To apply a causal pattern, first a generalized observation is created by finding all roles common to the set of observations with the same state change. Next, the generalized observation is matched against the antecedents of the causal patterns.
4. **Instantiate causal relationship:** If the antecedent of causal pattern matches (and the proposed causal relationship is consistent with observation retrieved from memory), a new causal relationship is created using a procedure that depends on the type of pattern.
   - *Exceptionless:* A new causal relationship is constructed by replacing each of the variables in the causal relationship of the pattern with the corresponding binding with all roles removed in the generalized observation.
   - *Dispositional:* Dispositional causal patterns restrict the search for a condition that differs between the positive and negative examples to an object that plays a specified role (indicated in boldface in the figures) in the action. The features common to all objects that play this role in all positive examples are collected. Next, those features present in any object that plays this role in negative examples are eliminated from consideration. One feature is selected from the candidates at random, and hypothesized to be responsible for the different state change (i.e., the causal relationship only holds when that feature is present).[2] The causal relationship is created in a manner identical to the exceptionless patterns except it also contains the feature (or conjunction of features) hypothesized to be responsible for the different state change.
   - *Historical:* The causal pattern indicates that the causal relationship is conditionally dependent on some previous action (indicated in boldface and a boldface box in the figures). These patterns are processed

[1] In this article, I do not address the issue of memory retrieval. The important point is that some, but not all prior actions and subsequent state changes, are recalled. An action and the state change is retrievable if the action is indexed in memory by a unique role that is present in the current observation (Kolodner, 1984; Lebowitz, 1980). The interested reader is referred to Pazzani (1990) for a description of the memory-retrieval process.

[2] If there are no candidate features, then the conjunction of all candidate features is tried.

in a manner similar to the dispositional pattern except that a prior action instead of a role is blamed for the different state change.

The matching process used by TDL is a strict match that succeeds only if the antecedent of the pattern subsumes a generalized observation. The antecedents of causal patterns are matched against generalized observations to enforce the regularity principle by ensuring that the causal relationship is consistent with the previous recallable observations. If the antecedent of a pattern does not match the generalized observation, a causal relationship cannot be created from that pattern. Note that the generalized observations only summarizes those prior observations that can be retrieved from memory.

A causal relationship is created from as few as one observation in TDL. Such a causal relationship is subject to revision when more examples are observed. A causal relationship constructed by TDL contains a counter that is incremented when a successful prediction is made, and another counter that is incremented when an incorrect prediction is made. When the ratio of correct predictions and total predictions is lower than a certain value,[3] then the causal relationship is eliminated. TDL does not use a backtracking mechanism to generate a new causal relationship when an erroneous relationship has been eliminated. Rather, the observations that can be recalled will prevent the system from applying the causal pattern that created the inaccurate causal relationship. An alternative causal pattern will apply to the new set of recalled observations and generate a new hypothesis. In the next section, an example is presented of the TDL procedure and the evaluation of causal relationships when new observations are encountered.

3.5 TDL: An Example

In this section, the execution of OCCAM is traced, a learning program that implements the TDL procedure. The program is presented with the following three observations:

- John, who has brown hair and brown eyes, is eating a Life Saver in the kitchen and John touches a red balloon with a silver pin. A state change occurs immediately after these two actions: The balloon bursts. The representation for this observation was given in Figure 2.
- John touches a red balloon with his finger in the living room and the balloon does not burst.
- Bob, who has black hair and brown eyes, touches a yellow balloon in the kitchen with the blade of a silver knife and the balloon bursts.

It is assumed here that OCCAM starts off with no causal relationships and with all of the causal patterns listed in Appendix A. Of particular importance are the causal patterns in Figures 4 and 7.

[3] This is a parameter in OCCAM. The current value of the parameter is 0.8.

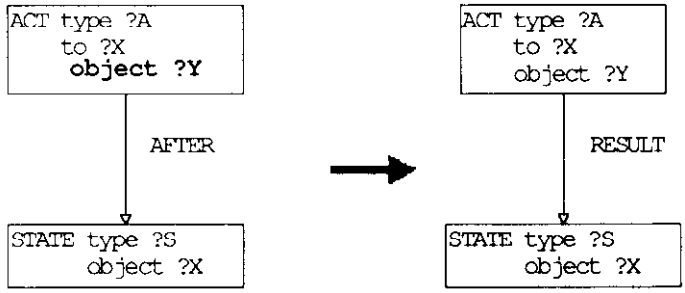

Figure 7. A dispositional causal pattern: Similar actions with the same destinations and with different objects followed by different state changes for the destination suggests that a role of the object is required for an action to result in the state change.

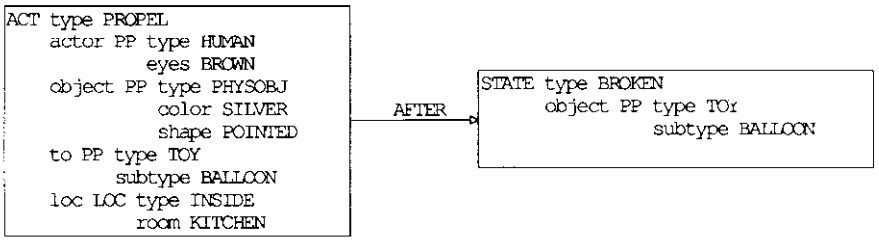

Figure 8. A generalized observation formed from finding the common roles of two observations. This generalized observation indicates that after a person with brown eyes touched a balloon with a silver pointed object in the kitchen, the balloon popped.

When the first observation is encountered, an unexpected state change occurs. Because there are no recalled situations in which a balloon did not burst when a force was applied, only exceptionless patterns will be matched against this situation. There is only one pattern that matches any part of this observation (see Figure 4). This pattern suggests the causal relationship that applying a force to a balloon results in the balloon bursting. (This causal relationship is not shown in any figure. However, it is identical to the situation in Figure 3 except it does not contain the constraint that the object be pointed.)

In the next observation, a balloon is touched with a finger. OCCAM uses the causal relationship to predict that the balloon will burst. However, the balloon did not burst. The causal relationship formed from the first example is deleted. Because OCCAM focuses on explaining unexpected state changes, and there is no state change for the current observation, the two observations are simply stored in memory.

When the third observation is encountered, an unexpected state change occurs. OCCAM retrieves the previous two observations from memory, and creates a generalized observation by finding the roles common to the first and third observations because these precede the same type of state change. The generalized observation is shown in Figure 8. By correlating over the

recalled observations, many roles need not be considered as being potential candidates to account for the difference in the outcome. These include the color of the actor's hair and the color of the balloon. When the generalized observation is matched against the causal patterns, additional roles will be considered irrelevant.

This generalized observation matches the dispositional causal pattern in Figure 7. This pattern encodes the knowledge that some difference in the object used in an action can result in a different state change. There are two differences between the object of the generalized observation and the object of the observation with a different state change: color and shape. OCCAM randomly selects one of these roles and creates a causal relationship.[4] In this example, OCCAM makes a fortuitous selection and chooses the shape. The resulting causal relationship is shown in Figure 3. If color were chosen instead of shape, then OCCAM would create an inaccurate causal relationship that would make an error if a balloon were touched with a pointed object that was not silver or a silver object that are not pointed. In either case, the inaccurate causal relationship would be retracted and replaced by the more accurate relationship in Figure 2.

Notice that there are many roles shared by the two observations with the same state. TDL does not consider many of these similarities, such as the location of the action or the color of the actor's eyes, to be relevant to the causal relationships. This permits TDL to converge rapidly on an accurate causal relationship consistent with the theory of causality. The price it pays for this increased speed is the inability to learn a relationship inconsistent with the theory of causality.

4. THE SCOPE AND LIMITATIONS OF TDL

To gain an understanding of the limitations and scope of TDL, it is necessary to describe OCCAM (Pazzani, 1990), the learning architecture that includes TDL as one component. OCCAM's other learning components are an explanation-based learning (EBL) component (DeJong & Mooney, 1986; Mitchell, Keller, & Kedar-Cabelli, 1986) and a similarity-based learning (SBL) component (Lebowitz, 1986b; Mitchell, 1982).

The SBL component creates causal relationships from several observations with the same state change. The conditions under which an action will result in a state change are learned incrementally by finding all roles in common to observations with the same state change (cf. Bruner et al., 1956). SBL does make use of either the theory of causality or the theory of causation to guide the learning process. This is both an advantage (i.e., SBL is not restricted to acquiring concepts that are consistent with existing knowledge) and a disadvantage (i.e., the prior knowledge of the learner does not constrain the learning process (DeJong & Mooney, 1986; Mitchell et al., 1986).

[4] In Pazzani, Dyer, and Flowers (1987) an extension to OCCAM that learns conditions under which one role should be favored is discussed.

The EBL component creates new causal relationships by using the theory of causation (as opposed to the theory of causality) to explain a single observation that has an unpredicted state change. A state change is explained by chaining together two or more causal relationships. For example, consider what happens when vinegar and baking soda are put in a bottle and a balloon is placed on the opening of the bottle. In this situation, the balloon will expand. This can be explained by chaining together three simple causal relationships. First, mixing vinegar and baking soda results in the production of carbon dioxide gas. Second, the production of gas results in an increased pressure in the bottle and the balloon. Third, an increase of pressure in the balloon results in the balloon expanding. EBL creates a new causal relationship by finding the most general conditions that this same explanation will apply. This new causal relationship can be created analytically from just one example by taking advantage of an interaction between existing causal relationships.

When an unpredicted state change is observed, OCCAM must determine which of its three learning methods to apply. OCCAM uses EBL if it can explain the state change by chaining together existing causal relationships. If an explanation cannot be produced and the observation is consistent with the theory of causality (i.e., the observation matches a causal pattern), then OCCAM uses TDL. As a last resort, OCCAM attempts SBL.

An important implication of this architecture is that SBL and TDL create the causal relationships needed by EBL. This permits OCCAM to use the results of its initial data-intensive learning in its later knowledge-intensive learning. For this reason, it is not necessary to have historical causal patterns of arbitrary length in OCCAM. Rather, several simple causal relationships created by TDL can be chained together by EBL to learn complex causal relationships with several intermediate states.

TDL's role is restricted to those observations that cannot be explained by the current theory of causation (otherwise, EBL would be used), and that meet the constraints of a potential causal relationship. Regularities between observations that cannot be explained, and that do not match a causal pattern (e.g., the opening of a garage door by pressing the button on a remote control) can be detected and generalized by the SBL component of OCCAM. The data on human learners does not show that people are incapable of learning causal relationships inconsistent with their theory of causality. Rather, people learn more slowly when observations do not conform to common patterns of causal relationships.

Figure 9 (p. 418) shows the result of running OCCAM on the observations from the experiment in Section 2. In this simulation, OCCAM was able to use TDL only on the *dip/inflate* condition. In the *alpha* conditions, only SBL can beused. In the *snap/inflate* condition, the first few observations may fit a causal pattern and a relationship such as "blowing into a yellow balloon results in a balloon being inflated" will be created. Later observations prove this relationship to be inaccurate, and the series of observations do not

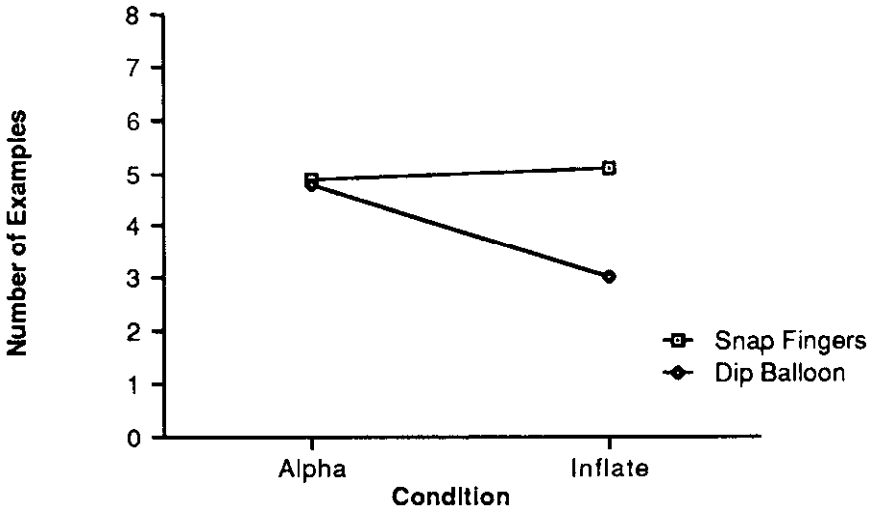

Figure 9. Data from running OCCAM on observations simulating the experiment in Section 2. The data points are averaged over 30 simulations. As in the experiment, fewer trials were needed to learn a relationship that conforms to a common pattern of causal relationships.

match any causal pattern. Therefore, OCCAM uses SBL to learn the accurate relationship for the *snap/inflate* condition.

The data from the simulation indicate that learning causal relationships is facilitated when the relationships are consistent with the theory of causality. In the data with human subjects, learning causal relationships inconsistent with the theory of causality takes longer than simple concept identification. Although there is a difference in the computer simulation, it is not statistically significant. It has been speculated that the smaller magnitude of the difference is caused by the fact that OCCAM may be able to retrieve accurately more prior observations than the human subjects.

4.1 Comparison to EBL

There are two primary differences between EBL and TDL approaches to learning causal relationships. First, the causal relationships produced by EBL deductively follow from the existing knowledge (i.e., the theory of causation). EBL does not increase the set observations that can be explained by the deductive closure of the theory of causation (Dietterich, 1986), but it does increase the set of observations whose state change can be predicted by the application of a single causal relationship. In contrast, the causal relationships produced by TDL follow from both the existing knowledge (i.e., the theory of causality) and the set of observations. EBL uses the observations only to focus the search for an explanation. TDL needs the observations to determine which states result from each type of action. For this reason, standard EBL algorithms (e.g., DeJong & Mooney, 1986; Mitchell et al., 1986) cannot make use of a theory of causality to create causal relationships.

A second difference between EBL and TDL is the method used to learn the conditions under which an action results in a state change. EBL produces generalizations by summarizing an inference chain. The preconditions acquired by EBL are operational descriptions of the preconditions the rules used to produce the inference chain. TDL learns the conditions under which an action results in a state change by a focused correlation among several observations.

4.2 Determinations

TDL is, in some ways, similar to learning with determinations (Russell, 1986). Determinations, like causal patterns, are not sufficient to make a prediction about an unobserved example. Once an observation has been encountered, properties of additional observations can be deduced. For example, one determination states that nationality determines language. After an example of an American speaking English has been seen, a generalization that all Americans speak English can be created. There are two differences between TDL and learning with determinations. First, TDL does not require that the observations and the causal pattern logically entail the causal relationship. Rather, the causal pattern is a heuristic suggesting a relationship subject to empirical validation. Second, more than one determination cannot apply to a given observation. Therefore, learning with determinations does not require a mechanism to select among alternatives.

5. IMPLICATIONS OF TDL

The distinction between a theory of causation and a theory of causality is useful for interpreting the results of experiments assessing the causal reasoning capabilities of human subjects. For example, in one study (Ausubel & Schiff, 1954), kindergarten students and sixth-grade students were asked to predict which side of a teeter-totter would fall when the correct side was indicated by a relevant role (length) or an irrelevant role (color). They found that the kindergarten children learned to predict on the basis of relevant or irrelevant roles at approximately the same rate (3.7 trials for relevant, 3.4 trials for irrelevant). However, the older children required significantly fewer trials to predict on the basis of a relevant role than an irrelevant one (.83 vs 3.1 trials). The relevant and irrelevant conditions are identical with respect to a theory of causality. Both conditions require finding a difference in an object that is responsible for the difference in a state change. Without any prior knowledge of causation, either role is equally likely. This experiment does not demonstrate that older children have a better theory of causality than younger children; rather, the experiment shows that older children have a more complete theory of causation. This theory includes relationships such as "a teeter-totter falls on the heavier side" and "the longer side is likely to be the heavier side." In the relevant condition, the correct prediction deductively follows from this knowledge.

Some experiments have shown that younger children have a less complete theory of causality than older children. For example, Bullock (1979) presented evidence that 3-year-old children do not make use of the spatial contiguity principle as much as 5-year-old children do. This difference is not due to knowledge of causation (e.g., familiarity with materials in the experiment). Rather, the difference is attributable to a difference in the 3-year-olds' and 5-year-olds' theory of causality.

It has been found (Shultz, Fisher, Pratt, & Rulf, 1986) that when subjects have a detailed knowledge of the causal mechanism (i.e., a theory of causation sufficient to explain a state change), they do not require temporal and spatial contiguity. I view this as evidence that supports the decision in OCCAM to prefer EBL to TDL is both apply. Furthermore, I view the principles of temporal and spatial contiguity as heuristics that allow an observer to infer a causal mechanism. For example, the causal pattern in Figure 7 changes temporal links in the observation into causal links in the causal relationship and postulates a state as an intermediate result.

6. EXTENSIONS TO TDL

In addition to causal patterns that guide the search for relationships of physical causation, causal patterns for social causation have also been developed. In physical causality, a state change occurs as a consequence of the transmission of some sort of force. In contrast, transmission of forces does not play a major role n determining human behavior. Instead, human behavior is considered to be a consequence of intentions to achieve some goal. The social causal patterns postulate intentional relationships (Dyer, 1983) between goals, plans and actions. For example, one social pattern is: *An event (?e) that motivates a goal (?g) for one person (?p1) is observed by another person (?p2) who performs an action (?a) that achieves ?g for ?p1, suggests that ?e motivates ?g for ?p2.*

OCCAM uses this pattern when it is given a series of observations of parents helping their children and strangers not assisting a child. OCCAM hypothesizes that parents have a goal of preserving the health of their children. (Of course, before being ruled out by additional examples, OCCAM also entertained a number of incorrect hypotheses such as persons with brown hair have a goal of preserving the health of children). Once OCCAM has constructed the social relationship that parents have a goal of preserving the health of their children, it can use it as background knowledge for EBL. This particular relation is useful in explaining and generalizing an observation in which a parent plays the ransom in a kidnapping episode.

Causal patterns that facilitate learning about electronic devices have also been experimented with. For example, the following causal pattern may be used to focus the search for causal relationships: *Pressing a switch followed by a state change of an electronic device suggests that pressing the switch*

results in a state change of the electronic device. This causal pattern encodes the fact that, as adults, we are more likely to attribute a change in an electronic device to the pushing of a button or the flicking of a switch than to some other random action (such as a cat meowing). This is true even if the wires are hidden (as in a light switch) or the connection is not observable (as in the remote control for a television). Similarly, one would be surprised if pressing a button resulted in a change in some nonelectronic device, such as the inflaton of a balloon.

This last example raises the question of how a theory of causality might be acquired. Early work on children's understanding of causality (Piaget, 1930) pointed out many differences in causal explanations among various age groups. In spite of more recent evidence that very young infants are able to perceive causal relationships (Leslie & Keeble, 1987), there is no question that older children are better at attributing causality than younger children (Bullock. 1979).

Currently, in OCCAM, there is a fixed set of causal patterns that never change as the program learns. When the program starts, it has its complete theory of causality. I believe that a general theory of causality can be acquired and refined from experience. I am currently developing an extension that would find higher order regularities (Goodman, 1983) among the causal relationships created by SBL. These higher order regularities become the causal patterns needed by TDL.

7. CONCLUSION

I have argued that learning causal relationships is facilitated by a general theory of causality constraining the set of possible causal relationships. I have presented a process called theory-driven learning, which proposes that causal hypotheses are consistent both with observed data and the general theory of causality. A computer implementation of the theory is one component of OCCAM. Simulation of the theory provides empirical support for the advantage of TDL over purely correlational approaches to learning: Fewer examples are required to learn a causal relationship. I have provided experimental evidence that people possess the kind of causal knowledge encoded in OCCAM's causal patterns.

REFERENCES

Ausubel, D.M., & Schiff, H.M. (1954). The effect of incidental and experimentally induced experience on the learning of relevant and irrelevant causal relationships by children. *Journal of Genetic Psychology, 84,* 109–123.

Bower, G., & Trabasso, T. (1968). *Attention in learning: Theory and research.* New York: Wiley.

Bruner, J.S., Goodnow, J.J., & Austin, G.A. (1956). *A study of thinking.* New York: Wiley.

Bullock, M. (1979). *Aspects of the young child's theory of causality.* Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. Friedman, (Ed.), *The developmental psychology of time.* New York: Academic.

DeJong, G., & Mooney, R. (1986). Explanation-based learning: An alternate view. *Machine Learning, 1,* 145–176.

Dietterich, T. (1986). Learning at the knowledge level. *Machine Learning, 1,* 287–315.

Dyer, M. (1983). *In-depth understanding.* Cambridge: MIT Press.

Goodman, N. (1983). *Fact, fiction and forecast* (4th ed.). Cambridge, MA: Harvard University Press.

Kolodner, J. (1984). *Retrieval and organizational strategies in conceptual memory: A computer model.* Hillsdale, NJ: Erlbaum.

Lebowitz, M. (1980). *Generalization and memory in an integrated understanding system.* Unpublished dissertation, Yale University, New Haven, CT.

Lebowitz, M. (1986a). Integrated learning: Controlling explanation. *Cognitive Science, 10,* 219–240.

Lebowitz, M. (1986b). Not the path to perdition: The utility of similarity-based learning. *Proceedings of the National Conference on Artificial Intelligence.* Palo Alto, CA: Morgan Kaufmann.

Leslie, A., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition, 25,* 265–288.

Levine, M. (1967). The size of the hypothesis set during discrimination learning. *Psychology Review, 74,* 428–430.

Michotte, A. (1963). *The perception of causality.* New York: Basic Books.

Mitchell, T. (1982). Generalization as search. *Artificial Intelligence, 18,* 203–236.

Mitchell, T., Keller, R., & Kedar-Cabelli, S. (1986). Explanation-based learning: A unifying view. *Machine Learning, 1,* 47–80.

Pazzani, M.J. (1990). *Creating a memory of causal relationships: An integration of explanation-based and empirical methods.* Hillsdale, NJ: Erlbaum.

Pazzani, M., Dyer, M., & Flowers, M. (1987). Using prior learning to facilitate the learning of causal theories. *Proceedings of the 10th International Joint Conference on Artificial Intelligence.* Palo Alto, CA: Morgan Kaufmann.

Peirce, C. (1932). *The collected papers of Charles Sanders Peirce* (Vol. 2). Cambridge, MA: Harvard University Press.

Piaget, J. (1930). *The child's conception of physical causality.* London: Kegan Paul.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations.* Cambridge, MA: MIT Press.

Russell, S. (1986). Preliminary steps toward the automatization of induction. *Proceedings of the National Conference on Artificial Ingelligence.* Palo Alto, CA: Morgan Kaufmann.

Salzberg, S. (1985). Heuristics for inductive learning. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence.* Palo Alto, CA: Morgan Kaufmann.

Schank, R.C., & Abelson, R.P. (1977). *Scripts, plans, goals, and understanding.* Hillsdale, NJ: Erlbaum.

Shultz, T. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47.*

Shultz, T., Fisher, G., Pratt, C., & Ruff, S. (1986). Selection of causal rules. *Child Development 57,* 143–152.

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

Vere, S., (1975). Induction of congress in the predicate calculus. *Proceedings of the Fourth International Joint Conference on Artificial Intelligence.* Palo Alto, CA: Morgan Kaufmann.

APPENDIX A

A List of OCCAM's Causal Patterns

This appendix contains a listing of causal patterns for postulating causal relationships. The corresponding dispositional patterns follow each exceptionless pattern.

- An action on an object by a state change for the object suggests that the action results in the state change.
  - actor disposition
  - object disposition
- An action with a particular destination followed by a state change for the destination suggests that the action results in the state change.
  - actor disposition
  - object disposition
  - destination disposition.
- An action on a component of an object followed by a state change for the object suggests that the action results in the state change.
  - actor disposition
  - object disposition
- An action with a particular component as a destination followed by a state change for the destination suggests that the action results in the state change.
  - actor disposition
  - object disposition
  - destination disposition
- An initial action on an object preceding a subsequent action that precedes a state change for the object suggests that the initial action results in a state change that enables the subsequent action to result in the state change.
  - There are three other variations of this pattern that permit the object that changes to be the destination of the initial action, the subsequent action, or both. In Conceptual Dependency, the destination of the object can be affected by an action.
- An initial action on an object preceding a subsequent action that does not precede a state change for the object suggests that the initial action results in a state change that disables the subsequent action to result in the state change.
  - There are three other variations of this pattern that permit the object that changes to be the destination of the initial action, the subsequent action, or both.

APPENDIX B

Order of Stimuli

The stimuli used in the experiment discussed in Section 2 were shown to subjects from videotape. The following table lists the ordering of the observations on the videotapes:

Balloon Size	Balloon Color	Action Dip/Inflate	Action Snap/Inflate	Result
Small	Yellow	Dip	Snap	Inflated
Large	Yellow	Snap	Dip	Not Inflated
Large	Orange	Necklace	Necklace	Not Inflated
Small	Yellow	Necklace	Necklace	Not Inflated
Small	Orange	Dip	Snap	Inflated
Large	Yellow	Necklace	Necklace	Not Inflated
Large	Orange	DIp	Snap	Inflated
Small	Orange	Snap	Dip	Not Inflated
Large	Orange	Snap	Dip	Not Inflated
Large	Yellow	Dip	Snap	Inflated
Small	Orange	Necklace	Necklace	Not Inflated
Small	Yellow	Snap	Dip	Not Inflated