This article appears in *Cognitive Science*, Volume 15, Number 4, 1991

Goal-based Explanation Evaluation[1]

David B. Leake

Computer Science Department
Indiana University
101 Lindley Hall
Bloomington, IN 47405
leake@cs.indiana.edu

[1]I would like to thank my dissertation advisor, Roger Schank, for his very valuable guidance on this research, and to thank the *Cognitive Science* reviewers for their helpful comments on a draft of this paper. The research described here was conducted primarily at Yale University, supported in part by the Defense Advanced Research Projects Agency, monitored by the Office of Naval Research under contract N0014-85-K-0108 and by the Air Force Office of Scientific Research under contract F49620-88-C-0058.

Correspondance and requests for reprints should be sent to David B. Leake, Computer Science Department, Indiana University, Bloomington, IN 47405-4101.

Abstract

Many theories of explanation evaluation are based on context-independent criteria. Such theories either restrict their consideration to explanation towards a fixed goal, or assume that all valid explanations are equivalent, so that evaluation criteria can be neutral to the goals underlying the attempt to explain. However, explanation can serve a range of purposes that place widely divergent requirements on the information an explanation must provide. We argue that understanding what determines explanations' goodness requires a dynamic theory of evaluation, based on analysis of the information needed to satisfy the many goals that can prompt explanation; this view conforms to the common-sense idea that people accept and apply explanations precisely if those explanations give the information they need. We examine a range of goals that can underly explanation, and present a theory for evaluating whether an explanation provides the information an explainer needs for these goals. We illustrate our theory by sketching its implementation in the computer program ACCEPTER, which does goal-based evaluation of the goodness of explanations for surprising events in news stories.

1 Introduction

The use of explanation is central to theories in many areas of artificial intelligence, such as text understanding (*e.g.* (Granger, 1980; Schank, 1982; Wilensky, 1983; Schank, 1986; Hobbs et al., 1990)), plan repair and indexing (*e.g.*, (Hammond, 1989)), and guiding generalization (*e.g.*, (Mitchell et al., 1986; DeJong & Mooney, 1986)). In addition, recent experiments have supported some of these explanation-based processes as psychological models (Ahn et al., 1987). However, the benefits of explanation-based processing depend on the explanations taken as starting point. In real-world situations, many candidate explanations may be available for a single event, prompting the question of how to select the explanation on which to base further processing.

Since explanations can be used in many ways, it seems reasonable that rather than seeking a universal "best" explanation, an explainer should tailor explanation towards serving the goal for which it is intended. Nevertheless, the influence of context on explanation has received little study in psychology and artificial intelligence. In psychology, the central current for research on people's choice of explanations is attribution theory (Heider, 1958), which generally accounts for choice of explanations without reference to what motivated the explanation effort. Likewise, in artificial intelligence, criteria for explanations' goodness tend to take a context-independent view, basing their judgements on criteria that make no reference to the goals overarching the explanation effort (*e.g.*, (Granger, 1980; Wilensky, 1983; Pazzani, 1988; Rajamoney & DeJong, 1988; Thagard, 1989; Ng & Mooney, 1990; Hobbs et al., 1990)). These theories focus on choosing the explanations most likely to be valid. However, even if they succeed in identifying explanations that are valid, they may not give sufficient information to select the explanation that best satisfies the explainer's goals.

For example, suppose that the event to account for is a company bankruptcy. Suppose further that the company was bankrupt as the result of two factors, either of which would have been sufficient to cause the bankruptcy: the company's bad management, and excessive local taxes. Consequently, for this example there are at least two valid explanations: "the bankruptcy was caused by bad management" and "the bankruptcy was caused by high taxes."

Bankruptcies have adverse effects on the communities in which they occur, so a local politician might try to explain the bankruptcy, with the underlying goal of preventing its recurrence. For this goal, the relative value of the explanations would be quite different. "The bankruptcy was caused by bad management" is unhelpful to the politician, since it suggests no course of action— government has no control over management in local industries. However, "the bankruptcy was caused by high taxes" could be the source of the generalization that extreme tax rates are likely to force companies into bankruptcy, which could be useful as motivation for the politician to lower taxes. For a shareholder in a company in a low-tax state, the usefulness of the two explanations is reversed: the shareholder benefits more from considering the dangers of bad managers, since bad management might cause company stock to lose value.

We might think that either explainer could simply require that an explanation identify *all* causally relevant factors, in that way assuring that all important conditions are taken into account. However, real-world events are simply too complicated for complete explanation. For example, any company bankruptcy actually depends on countless factors, such as competing products, consumer demand, the company's labor situation, its past history (affecting its cash reserves), the owners' management style, and the current bankruptcy laws. In our view, good explanations must highlight a few important factors from the many that are causally involved: those that give the explainer the information it needs. Since no explanation can be exhaustive, goal-based explanation evaluation is needed for explainers to recognize whether candidate explanations address the appropriate factors. Insufficient explanations can be rejected, or elaborated until they provide the needed information.

In what follows, we develop a theory of goal-based explanation evaluation. We elaborate on the role of overaching goals in explanation, demonstrating how different goals lead to different requirements for explanations' goodness, and how these can be characterized along a small set of dimensions. We then describe the implementation of goal-based evaluation in a computer model, ACCEPTER, which illustrates our theory (Leake, 1988a, 1988b, 1989b, in press). We argue that goal-based evaluation improves an explanation-based system's capability to focus explanation, to guide explanation construction, and to function effectively despite some imperfections in its domain theory.

2 Overview

We begin by sampling prior explanation evaluation approaches from psychology, philosophy, and artificial intelligence. In both psychology and in artificial intelligence these theories tend either to look at the evaluation process as context-independent, or to examine it within a single fixed context. Some philosophical accounts have taken a view close to ours, arguing that explanation depends on context, and we give a sampling of some of those perspectives.

We then investigate how a specific aspect of context, the overarching goal of the explainer to use an explanation, determines the information that the explainer requires when confronted with an anomalous situation. We show that goal-based evaluation permits a more perspicacious choice between candidate explanations than prior approaches, and also extends the range of explanations an explainer can use, by enabling principled use of partial explanations.

We concentrate on evaluation of explanations of anomalous events; those events often prompt new goals because they reveal unexpected aspects of a situation. The new goals prompt plans, whose success depends on appropriate information, in turn prompting *explanation purposes* to extract particular information from an explanation. In order to characterize the information that explanation purposes require, we develop a small set of *evaluation dimensions* for testing the usefulness of hypothesized causes of an event, and that are used to build checks for the explanation purposes. Thus our model accounts for explanation evaluation as follows:

$$\text{Anomaly} \Longrightarrow \text{Goals} \Longrightarrow \text{Plans} \Longrightarrow \begin{array}{l}\text{Explanation}\\ \text{purposes}\end{array} \Longrightarrow \begin{array}{l}\text{Dimension}\\ \text{checks}\end{array} \Longrightarrow \begin{array}{l}\text{Evaluation}\\ \text{decision}\end{array}$$

In this paper, we concentrate our discussion on explanation purposes and evaluation dimensions. We show that checks along different combinations of those evaluation dimensions are sufficient to evaluate explanations for a wide range of explanation purposes. We sketch ACCEPTER's preliminary implementation of checks for certain evaluation dimensions, and show how its evaluation dynamically reflects changes in context and in overarching goals. Finally we argue that the effectiveness of any explanation-based processing depends on the ability to assure that explanations satisfy the needs for information that arise from system goals.

3 Previous perspectives

3.1 Psychological approaches

Much psychological research on choice of explanations originated in Heider's seminal work on attribution theory (Heider, 1958). Attribution theory investigates how people decide whether to explain an action in terms of features of its actor, or features of the environment. (Most work in attribution theory assumes that either personal or situational factors will apply, but not both.) One important result is Kelley's *covariation principle*, which gives a hypothesis for how people make the decision between attributing an outcome to personal or situational factors (Kelley, 1967). The covariation principle suggests that people look at covariation across different time, people, and other entities in order to decide which type of factor applies. For example, if John dislikes a movie, but most other viewers are enthusiastic, Kelley's covariation principle suggests that John's dislike should be explained by aspects of John, rather than aspects of the movie.

Although attribution theory gives criteria for deciding which class of factors to implicate, it does not suggest how to decide which particular personal or environmental factors are important. In the example of the movie, it says that a good explanation must involve *some* aspect of John, but deciding which is beyond its scope. This problem was pointed out by Lalljee and Abelson, who observe that people would usually try to find a more specific reason for the dislike (Lalljee & Abelson, 1983). For example, someone who had invited John to the movie, expecting him to like it, would probably try to determine John's particular objections. The more specific information is more useful: it allows predicting John's reaction next time, to avoid inviting him to another inappropriate movie.

Attribution theory also fails to consider the effect of context on preferences for explanations. The covariation principle does not take an explainer's prior expectations into account, or the reason for explaining. However, (Lalljee et al., 1982) shows that the explanations people seek, rather than being determined by abstract criteria, vary with circumstances: unexpected behavior requires more complex explanations than expected behavior, and is likely to require more of *both* situational and personal elements.

A knowledge structure approach: Lalljee and Abelson question attribution theory's basic approach, and suggest that its problems can be overcome by adopting a knowledge structure approach to attribution. They distinguish two types of explanation: *constructive* and *contrastive*. In constructive explanation, people explain events in terms of knowledge structures such as scripts and plans (Schank & Abelson, 1977). In contrastive explanation, they explain them by showing why events deviated from expectations provided by knowledge structures. For example, "John left his bicycle unlocked" might be explained in terms of *goal reversal*: perhaps rather than having the normal goal of wanting to protect it, he actually wanted to get rid of it.

Our theory follows the basic lines of this approach, accounting for anomalies through both constructive explanation (to understand the surprising information) and contrastive (to account for the failure in previous expectations or beliefs). In addition, we look at how preference for explanations is affected by goals beyond the general desire to fill knowledge gaps, to further specific aims overarching the understanding process.

Excuse theory: Research on attribution has examined the influence of one class of overarching goal on the attribution process: the goal to absolve the explainer from blame. Excuse theory studies how the desire to form excuses makes people manipulate the types of factors to use in attribution, to blame external influences for their own bad performance (for example, (Mehlman & Snyder, 1983) or (Snyder et al., 1983)). Excuse theory's demonstration that goals influence explanation is consistent with our view. However, we consider evaluation within a knowledge structure framework, rather than attribution theory, in order to address the specific knowledge required from an explanation, and consider the influence of a wider range overarching goals.

3.2 Philosophical views

Much of the philosophical investigation of explanation has attempted to establish formal requirements for explanations, independently of context, but some approaches have taken a view much closer to our goal-based perspective. For example, Hanson points out that different explainers would focus on different aspects of a fatal vehicle accident:

> There are as many causes of x as there are explanations of x. Consider how the cause of death might have been set out by a physician as 'multiple haemorrhage', by the barrister as 'negligence on the part of the driver', by a carriage-builder as 'a defect in the brakeblock construction', by a civic planner as 'the presence of tall shrubbery at that turning'. (Hanson, 1961, page 54)

Likewise, Mackie (1965) discusses the need of explanations to aid in making the distinctions important to a particular explainer. In this view, explanation is conducted against the *causal field* of situations to be distinguished by the explanation, and varies with that field. This makes requirements for explanation strongly context-dependent: an explanation of "why did the car crash going around the turn?" would be different if the causal field were other turns (in which case the explanation might be the turn's sharpness), or if it were other instances of the car going around the same turn (in which case the explanation might be that this time, the driver was sleepy). Van Fraassen (1980) argues in favor of Hanson's view, and accounts for the choice of causes with an idea similar to the causal field. However, our approach differs in examining the needs that generate the background for explanation, and showing how particular goals determine concrete requirements that an explanation must satisfy.

3.3 AI approaches

In artificial intelligence, the question of explanations' goodness has been investigated in three main areas. Research in expert systems explanation has concentrated on the question of explanations' goodness for explaining system behavior, for the benefit of the system user; research in explanation for text understanding has concentrated on how to select valid explanations from a range of hypotheses; and explanation-based learning research has primarily considered the problem of determining explanations' goodness for learning to improve performance on concept recognition and search. We discuss below the approaches of each area.

3.3.1 Expert systems explanation

Research on expert systems explanation has investigated the problem of generating explanations of expert system behavior that are sufficient either to educate the user about the task domain, or to justify system decision-making (Shortliffe, 1976). The desire to provide the needed information has prompted development of many explanation systems, including systems that can show not just their decision paths, but the reasoning underlying those paths (Swartout, 1983); systems to devise explanations that are appropriate to the user's prior knowledge level (Paris, 1987); and systems that treat the explanation process as a continuing dialogue, allowing clarifications and elaborations to be offered in response to follow-up questions (Moore & Swartout, 1989). Our investigation takes a complementary view, somewhat closer to that of the *user* of such systems than the systems themselves: we examine how an understander with particular knowledge and goals can decide whether a given explanation is sufficient, or whether to ask additional questions.

3.3.2 Selecting explanations for understanding

As discussed in the introduction, much AI research on explanation evaluation has been addressed towards deciding explanations' likely validity. For example, research on explanation in story understanding has relied primarily on fixed structural criteria for choosing between competing explanations, such as favoring explanations involving short explanatory chains, or favoring explanations with the most structural coherence (*e.g.*, (Granger, 1980), (Wilensky, 1983), (Ng & Mooney, 1990)).[2]

However, the examples in our introduction show that validity alone is not enough to assure an explanation's goodness: an explanation may be valid without being useful. In fact, validity may actually be undesirable in the context of certain explainer goals, as is shown by the following scene from the movie *Breaking Away*:

> A used-car salesman is taking a prospective buyer out on a test drive. He stops suddenly to avoid a bicyclist, and the car dies. He tries frantically to start it,

[2] In (Leake, in press), we argue for evaluating explanations' validity in terms of content-based criteria rather than solely structural ones, and present a content-based approach to deciding validity.

> but he can't. He explains to the buyer: "I must have put expensive gas in it by mistake. This baby *just hates* expensive gas."

Although the salesman knows his explanation is invalid, he selects it because it serves his purpose better than the true explanation of the car's bad condition. Not only does it divert blame from the car's condition to the salesman's mistake, but it also introduces a new factor to sway the prospective buyer towards buying it: that the car is inexpensive to operate.

3.3.3 Operationality in explanation-based learning

Explanation-based learning, in which an explanation is used to guide generalization, is a powerful means of category formation, often allowing learning from single examples (Mitchell et al., 1986; DeJong & Mooney, 1986). The desire to assure that useful concepts are formed has prompted study of what determines an explanation's operationality— its suitability for use by the system performance element. (The notion of operationality originated with a somewhat different formulation in (Mostow, 1983).)

In most EBL research, operationality criteria aimed towards efficient recognition of concept instances, or control of search in problem solving. Criteria used for judging the operationality of a concept formulation range from static annotation of predicates with their ease of evaluation (Mitchell et al., 1986), to dynamic techniques based on system knowledge (DeJong & Mooney, 1986), to techniques directly reflecting the utility of an explanation, using estimates and actual measurements of recognition cost versus benefit to the system (Minton, 1988).

Research on operationality for concept recognition has led to the identification of significant characteristics of operationality for that task, such as the operationality/generality tradeoff (*e.g.*, (Segre, 1987)). Because operationality has been studied for so few tasks, there has been some tendency to assume that such properties apply to operationality judgements for *any* task. However, (Keller, 1988) points out that existence of an operationality/generality tradeoff depends on specific assumptions underlying the recognition task, and that the tradeoff does not exist for every use of explanations— a more general explanation may actually be more operational as well. For example, a teacher generating an explanation for a student may find a more general explanation more operational, for the task of extending the range of assignments a student can solve. Thus in order to understand operationality, we need to consider the wide variety of tasks for which explanations are used.

Keller's program MetaLEX (Keller, 1987) reflects the dynamic nature of operationality by not relying on fixed operationality criteria. Instead, its operationality criterion is an input to the system, and can reflect current goals and goal priorities. Consequently, the spirit of this work is quite similar to ours, though it does not address the question of identifying the range of possible purposes for explanation. (Souther et al., 1989) also presents an argument close in spirit to ours— that it is essential to be able to generate explanations from a given viewpoint— and identifies classes of explanations that students might seek when studying college-level botany. However, since that work concentrates on tutoring applications, it does not need to connect the classes to over-arching goals that make them important, beyond simply doing well in a course.

The above approaches concern the level of elaboration to use when explaining a given concept. Although many explanation-based systems take as input the concept to explain, approaches have also been advanced to select useful concepts to acquire. They are beyond the scope of this paper; for a discussion of some of those issues, see (Kedar-Cabelli, 1987), which examines how the purpose of finding objects to use in plans guides selection of concepts to explain, or (Schank, 1982; Riesbeck, 1981; Hammond, 1989; Leake, 1991), which address how the need to repair an expectation failure prompts the explanation effort.

4 A theory of goal-based evaluation

Our work concentrates on evaluating explanations generated for anomalous states or events that an understander notices in everyday understanding— conflicts with its beliefs and expectations. These anomalies show that the understander's world model is flawed in some way, since the world differs from its expectations or beliefs. In order to respond appropriately to protect its goals, and to take advantage of changes in the world, the understander needs additional information. For example, someone fired from a job may have many reasons to try to discover why the firing occurred, such as avoiding future firings, protesting the current one, or preparing to counter bad references when applying for new employment. To illustrate the range of purposes that may arise, and their effect on explanation, we consider the following example:

> Company X was beleaguered by high taxes, foreign competition, and out-dated equipment, despite low labor costs due to being non-union. Rumors of problems spread, and the company's stock plummeted, but its managers announced their decision not to have layoffs. The next week, it was rumored that they would lay off 20 percent of their work force.

Because the layoffs violate the mangers' pledge, they are anomalous, and might prompt explanation. Actors involved in the incident would be likely to have goals affected by it, and achievement of those goals would often require particular information. These information requirements determine which explanations are sufficient:

1. Someone who knew about the managers' pledge, and consequently did not believe that there would be layoffs, would want to verify that the layoffs would really occur. The goal of maintaining accurate beliefs prompts a plan to substantiate the layoffs. This prompts the explanation purpose of convincingly connecting the layoffs to trusted information. For this purpose, a satisfactory explanation might be *a newspaper found a secret company memo from the company president, describing the timetable for the layoffs.*

2. The same person might have the goal of avoiding future incorrect predictions in similar situations. One plan to achieve that goal is to explain why the current belief went wrong, and repair the source of the problem. Here the purpose of explanation would be accounting for the bad prediction in terms of false prior beliefs. If the problem was that the manager had been believed trustworthy, but was actually dishonest, that explanation would account for the expectation failure, and prevent the explainer from being mislead by the manager again.

3. A worker who was surprised to be layed off might want to avoid being unemployed, by finding a new job before being layed off next time. For this purpose, a suitable

explanation might be *the layoffs were inevitable because of pressure to reduce costs to shore up the company's falling stock*: given that this is a valid explanation, the danger of future layoffs could be predicted by watching the stock price.

4. In a local politician, the layoffs might prompt the goal of improving the area's economic health, by ending the layoffs. This gives rise to the explanation purpose of finding feasible repair points for the current situation, by explaining how the layoffs are caused or enabled by factors that are under government control, and whose removal will restore the desired situation. For example, if layoffs resulted from high taxes, a possibility would be to lower taxes in order to make the factory again profitable, and have the workers called back.

5. A worker who wondered whether to look for a new job, or simply wait to be called back, would want to know how long the layoffs were likely to last. This would prompt the explanation purpose of clarifying the situation to help form predictions of the layoffs' duration, which might involve finding if the layoffs resulted primarily from long-term factors, such as foreign competition, versus short-term ones, such as temporary overstocks.

6. A worker still employed, who wanted his job to be more secure, might want to prevent future layoffs. This would prompt the explanation purpose of finding potential causes of the layoffs that the worker can affect. If the factory's previous lack of unionization resulted in a contract that gave employees no security, enabling the layoffs, the worker could respond by starting a union. (The requirements for preventing an event are not necessarily the same as for its repair: unionization is likely to be an effective preventative for layoffs, but after workers have already been layed off, a new union is unlikely to make the company rescind the layoffs.)

7. The manager who ordered the layoffs might want to avoid negative publicity, and decide to do so by deflecting blame. One explanation purpose for this goal would be identifying other actors' contributions to the outcome, as by explaining the role that an outside expert had in shaping the decision.

8. The owner of another factory might want to improve its profitability. This goal might trigger the plan of analyzing other managers' decisions, to learn better management strategies. This could prompt the explanation purpose of deciding why the other managers chose to have layoffs, as opposed to alternative responses to the company's problems, such as increasing advertising to increase demand for products. Accounting for their surprising choice might show important factors that the owner would need to consider in similar future decisions. To determine those factors, the owner would have the explanation purpose of finding particular types of causes: the goals and goal priorities that entered into the decision of layoffs versus advertising.

9. A business consultant might want to develop a theory of how demographic trends force layoffs in different industries. A plan for that goal is to show how the theory accounts for particular episodes of layoffs, which would prompt the explanation purpose of finding causes of the current layoffs that are relevant to the theory.

Even though all the above actors are explaining the same event, each has different needs for information, prompting a different explanation purpose. Although we can imagine single explanations that would be usable for multiple purposes, an explanation's usefulness for multiple purposes is not assured: none of the sample explanations are sufficient for any purposes beyond the one each illustrates.

Our examples identify nine major explanation purposes. Although we do not claim that they form an exhaustive list, these purposes are sufficient to demonstrate that there is a wide range of goal-based purposes for explanation, including:

1. Connect event to expected/believed conditions.

2. Connect event to previously unexpected conditions.

3. Find predictors for anomalous situation.

4. Find repair points for causes of an undesirable state.

5. Clarify current situation to predict effects or choose response.

6. Find controllable (blockable or achievable) causes.

7. Find actors' contributions to outcome.

8. Find motivations for anomalous actions or decisions.

9. Find a within-theory derivation.

Our basic model of goal-driven explanation is that anomalies show that the world is different from expected, so that goals and plans may need to be re-considered. The desire to profit from the situation, or to avoid bad effects that might result, triggers selection of new goals, which trigger plans to achieve them. These plans generate needs for information. Based on those needs, an explainer has an explanation purpose to construct an explanation reflecting certain aspects of the situation, and evaluation must confirm that those aspects are included in the explanation. Table 1 sketches this process, showing how each of the purposes above can arise from a general goal, and a strategy to achieve the goal. Each explanation purpose determines specific types of information that an explanation must provide.

Sometimes an anomaly will suggest more than one goal, causing multiple explanation purposes to be active simultaneously. For example, if something happens that is both surprising and desirable, the explainer might want to find how to recognize when it is likely to occur again, to profit from it, and to know whom to credit with bringing it about, to encourage that person to bring it about again. Our theory deals only with evaluation of explanations once the purpose has been selected; the sequence of steps from anomaly detection to new goals, and plans giving rise to requirements for information to achieve them, is a topic for future research.

The following sections discuss the implementation of evaluation criteria in a computer system that judges explanations for a range of explanation purposes, accepting them if they include the information needed for those purposes.

5 A computer model

ACCEPTER is a story understanding program that requests explanations when it detects anomalies— conflicts between new information and its prior beliefs or expectations. Since

Table 1

A sketch of how the nine explanation purposes arise. Goals to deal with a surprising situation trigger plans for which information is needed, prompting an explanation purpose.

Goal $\Longrightarrow$	**Plan** $\Longrightarrow$	**Explanation purpose**
Prevent bad effects of acting on false information.	1. Confirm reasonableness of new information.	Connect event to expected/believed conditions.
	2. Find source of flaw in previous information.	Connect event to previously unexpected conditions.
Minimize bad effects/maximize good effects in similar future situations.	Predict similar events in time to prepare.	Find predictors for anomalous situation.
Use malfunctioning device.	Execute repair.	Find repair points.
Protect current plans.	Deal with important ramifications of anomaly.	Clarify situation to predict effects.
Re-achieve the good effects caused by anomalous event.	Re-cause event.	Find achievable causes.
Prevent recurrence of surprising bad state.	Punish current actors to deter future perpetrators.	Find actors' contributions to outcome.
Counter adversary.	Predict and respond to his actions.	Find motivations for adversary's unexpected actions.
Refine/demonstrate a theory.	Use theory to account for unexpected data.	Build within-theory derivation.

an anomalous event indicates that the system's understanding is incorrect, it signals that there may be unexpected risks or opportunities, and prompts new goals for an understander. Goals to guide evaluation are given to ACCEPTER as input, and it evaluates explanations' appropriateness for those goals. The system processes about 20 simple stories of anomalous events, including stories of the space shuttle Challenger's disaster, the warship Vincennes' accidental shootdown of an Iranian airliner, celebrity deaths, and automobile defects. For these anomalies, it evaluates a total of about 30 explanations. The appropriateness of each explanation can be evaluated for predicting similar events, preventing their recurrence, repair of the current situation, and assignment of blame or responsibility.

The program has been used both as a stand-alone system, and as the explanation evaluation component in SWALE, a story understanding system that uses information from ACCEPTER to guide explanation of novel events in the stories it processes (Schank & Leake, 1989; Leake & Owens, 1986; Kass, 1986; Kass & Leake, 1988). SWALE uses the resultant explanations to account for novel events, to form new explanatory schemas, and to guide indexing of those schemas in memory.

After discussing ACCEPTER's evaluation criteria, we return to ACCEPTER's role in SWALE, to examine the advantages of goal-based evaluation to an explanation-based understanding system.

5.1 ACCEPTER's basic algorithm

ACCEPTER takes as input a story represented in terms of Conceptual Dependency theory primitives (Schank, 1972), or in terms of schemas packaging sequences of those primitives to represent stereotyped sequences of actions. For example, a schema might represent the events involved in eating at a restaurant, such as entering, being seated, being brought menus, ordering, etc. These schemas are represented in ACCEPTER's memory as memory organization packets (MOPs) (Schank, 1982).

ACCEPTER processes stories one fact at a time, updating its beliefs and generating expectations for later inputs from the story. For this routine understanding, ACCEPTER uses schema-based understanding process modelled on (Cullingford, 1978), integrating new information into a dynamic memory (Schank, 1982; Lebowitz, 1980; Kolodner, 1984). As it integrates input facts into memory, it checks for anomalies— conflicts between the inputs and its beliefs or expectations. ACCEPTER implements a theory of pattern-based anomaly detection that allows problems to be detected with controlled inference, but that process is beyond the scope of this paper. See (Leake, 1989a) for an overview, or (Leake, in press) for a more complete account.

When an anomaly is found, ACCEPTER presents the anomaly to the user, along with possible explanations. These explanations are formed by retrieving and instantiating explanation patterns (XPs) (Schank, 1986) from an XP library in ACCEPTER's memory. Explanation patterns represent both specific explanations of prior episodes, and more general stereotyped commonsense explanations, such as *If a car won't start, it may have a dead battery*, or *If a student fails a test, it may be because of not studying enough*, or *If a car has a defect, it may be because of manufacturer's bad quality control.* (We describe the structure of explanation patterns below.)

From the list presented by ACCEPTER, the user selects a candidate explanation pattern to instantiate, and inputs an explanation purpose to be taken into account when judging the information that the explanation provides. ACCEPTER then performs the evaluation, signalling any problems it encounters. In the SWALE system, ACCEPTER's problem characterizations are used to index into a library of explanation modification strategies, to select a strategy appropriate to the problem (Kass, 1986). The strategy is then applied, and its result re-evaluated to identify problems that remain, or new problems introduced by the repair. In the stand-alone system, the user can respond to ACCEPTER's problem report by choosing an alternative explanation that avoids the problem, or selecting different role-fillers for any user-instantiated roles in the explanation. For example, one explanation ACCEPTER evaluates for the Challenger explosion is that Russia might have sabotaged the launch. ACCEPTER signals a plausibility problem, since Russia would not risk such a dangerous confrontation with the United States. After ACCEPTER points out that problem, the user can re-instantiate the explanation with another saboteur. If the user selects Libya, ACCEPTER considers the new explanation more reasonable because of Libya's willingness to take risks.

Figure 1 summarizes ACCEPTER's basic understanding process.

Figure 1: ACCEPTER's basic understanding process.

5.2 Representation of explanations

Explanation patterns trace the reasoning needed to account for an event. They represent explanations' structure in belief-support chains, which are belief dependency networks that include four components: initial states or events, hypothesized as leading to an event; internal beliefs, inferred from the initial hypotheses on the way to the conclusions to be derived; the conclusions themselves; and plausible inference links tracing the inference process from initial beliefs to conclusions.

For example, the XP *early death from life in the fast lane* traces the circumstances leading to the death of stars such as Janis Joplin and John Belushi. Its belief-support chain starts with two basic hypotheses: that the deceased was a performer, and was very successful. One of ACCEPTER's inference rules encodes the tendency for star performers to be under considerable stress. Starting with the inferred belief that the performer is under high stress, the desire to reduce stress can be inferred, and from this desire the tendency to take recreational drugs as an escape. Taking recreational drugs sometimes leads to an accidental drug overdose, from which death is a likely outcome. Figure 2 is a schematic diagram of the belief-support network associated with this reasoning chain.

We can imagine situations in which any link in the above reasoning would fail to hold. For example, stress does not necessarily lead to taking drugs, not does recreational drug use necessarily lead to an overdose. Each of the links is simply a plausible inference link, showing how the hypotheses make the conclusion more likely. For a detailed description of the structure of belief-support chains, and the types of links used, see (Schank & Leake, 1989).

6 Explanation evaluation for routine understanding

ACCEPTER's explanation evaluation is guided by two types of goals. First, its evaluation serves the basic goal of an understander to account for events in the stories it processes, and to maintain accurate predictions. Second, it serves dynamically changing goals beyond routine understanding, by evaluating whether explanations provide the information necessary to achieve them. In this section, we discuss how ACCEPTER's evaluation criteria

Figure 2: ACCEPTER's belief-support chain for the XP *early death from life in the fast lane.*

judge whether an explanation provides the information needed to satisfy the goals of routine understanding. The following section discusses evaluation criteria for explanations serving other goals.

When an understander encounters anomalies, they prompt knowledge goals (Ram, 1989) to reconcile the anomalies with other knowledge, both to correct the understander's picture of the current situation, and to avoid forming faulty expectations in the future. Correcting its picture of the current situation requires determining why the surprising event (or state) was reasonable, in order to integrate the anomalous event into previous beliefs; avoiding forming future bad expectations requires identifying the flaw in prior reasoning (Leake, 1988b; Collins & Birnbaum, 1988).

Evaluation for showing an event's reasonableness: Finding why the surprising event (or state) was reasonable requires showing why the event should have been expected, given the prior situation. To verify that an explanation provides the needed substantiation, ACCEPTER checks that the explanation's antecedents are consistent with previous beliefs, expectations, or known patterns (Leake, 1989a), and that the belief-support links deriving the event are also consistent with its predictive knowledge. For example, if we are surprised by a car's high price, a salesman might show that the high price is caused by quality manufacturing and unusual amenities.

Evaluation for identifying flaws in prior reasoning: Avoiding forming future bad expectations requires explaining the flaw in prior reasoning. If the expectation was generated by applying a standard knowledge structure, the explanation must show why that knowledge structure was inapplicable, by pointing to aspects of the situation that were distinctive from the norm, but previously overlooked. (In real-world situations, it is simply impossible to consider all potentially relevant factors in advance.) If the expectation arose from reasoning about a causal chain or physical process, the explanation can identify blockages in the chain, or unusual conditions interfering with the process. For example, if we expect bread to rise, we might explain its failure to do so by the yeast being put into hot tap water, which was much hotter than we had believed, and being killed by the high temperature. In this case, the cook probably thought the temperature of hot tap water was lower, and the explanation shows that this belief must be revised. We call the factors that were previously unexpected, or even disbelieved, *distinctive causes* of the event— they distinguish the real situation and what was previously considered.

In order to verify whether an explanation connects an event to distinctive causes, a system needs to be able to determine whether a fact is distinctive in the current context. For example, many causes lead to the Challenger explosion, such as design of the solid rocket boosters (which relied on O-rings to hold a seal), launch approval procedures (which allowed the engineers' warnings to be overridden), the high temperature of the boosters' flames (that helped them burn through the seals), and the coldness of the launch day (which made the seals brittle, interfering with the seals' positioning, and making it easier for flames to penetrate). However, most of these factors were routine: no problems had arisen from the design and launch procedures, or the booster flame temperatures, in past launches. Consequently, an explanation based on them would not say why the shuttle exploded on that particular launch but not on previous launches.

However, the cold weather on the launch day *did* differ from circumstances of previous launches. Since the cold day was distinctive, it would be reasonable to fear another explosion on cold days, and to expect successful launches on warmer days. In fact, when the Space Shuttles were grounded after the Challenger explosion, many people argued that the launches should be allowed to continue on any days that were above a certain temperature.

ACCEPTER's evaluation criteria are expressed in terms of a set of requirements that the antecedents of an explanation must satisfy in order for the explanation to provide the needed information. We call the classes of factors *evaluation dimensions*; the distinctiveness of causes from prior expectations is one of these dimensions. We discuss the evaluation dimensions in detail below.

7 Beyond routine understanding: evaluation for more specific goals

The previous section sketched evaluation for two explanation purposes: connecting an event to expected/believed conditions, in order to confirm its reasonableness, and connecting an event to unexpected or disbelieved conditions, in order to repair faulty knowledge. Both these purposes arise from ACCEPTER's basic understanding goal of maintaining an accurate model of the stories it reads.

ACCEPTER's second set of explanation purposes is provided externally, and reflects other goals. In a planning system that used ACCEPTER to maintain its world model, the purposes would be triggered by the goals and plans of the overarching planner. In the current stand-alone version of ACCEPTER, the purposes are selected by a human user.

ACCEPTER evaluates whether explanations include the information needed for finding predictors, finding repair points, finding controllable causes, and finding actors' contributions to an outcome. The system also combines its evaluation for other purposes to judge explanations for a higher-level goal: blocking future occurrence of an undesirable outcome.

ACCEPTER's evaluation process is to trace through the belief-support network explaining an event, examining the causes in the network to see if some subset of those causes provides the needed information. If some subset of those causes includes all needed types of factors (*e.g.*, if ACCEPTER's purpose is to connect an event to unexpected conditions, and it finds a distinctive cause), the explanation is accepted. All relevant causes the system finds, and a description of the tests they satisfy, are output by the system, to facilitate applying the explanation's information in plans for overarching goals.

While it would be possible to simply devise independent evaluation procedures for each purpose, parsimony suggests analyzing the purposes to find shared components of the tests for different explanation purposes, and using those components as building blocks for evaluation procedures. ACCEPTER builds its purpose-specific evaluation procedures from tests that evaluate causes along nine *evaluation dimensions*:[3] timeliness, knowability, distinctiveness, predictive power, causal force, independence, repairability, blockability, and desirability. In what follows, we will demonstrate how explanation purposes relate to particular information needs, and how the needed information is characterized in terms of combinations of the evaluation dimensions. For example, we will show that in order to repair an undesirable device state, we need to find any causes of the problem that are still in effect (timeliness),

[3]Although we simplify the discussion by presenting the values on these dimensions as yes/no decisions, all actually fall along a continuum.

that we can repair (repairability), and that are unusual compared to the normal device state (distinctiveness).

After an initial set of dimensions was defined for a few purposes, we found that needs for many of our other nine purposes could be described by simply using different combinations of dimensions from the initial set, and we believe that only a small set of additional dimensions would be needed for adding other purposes to the system. The sections below describe some of the dimensions implemented for ACCEPTER's explanation purposes, and sketch their implementation.

7.1 Evaluation dimensions for prediction

When an event surprises us, and we want to anticipate it it in the future, we need to explain what caused its occurrence, in predict it in similar future circumstances. In order for a group of causes to be useful for prediction, it must hold that (1) occurrence of the causes makes the event likely, (2) the cause happens long enough in advance of the event for the prediction it triggers to be useful, (3) one of the causes is unusual compared to the expected situation, so that it gives evidence for the surprising event *as opposed to* the previously expected one, and (4) the causes are factors that we are likely to be aware of in the future. These requirements correspond to four evaluation dimensions: (1) predictive power, (2) timeliness, (3) distinctiveness, and (4) knowability.

Predictiveness: Even if something caused an event, it may not be predictive of that event. For example, even if the blowout of a tire was caused by driving at high speed, we would not expect a blowout the next time we see a speeding driver. To determine whether causes are predictive of an outcome, ACCEPTER relies on annotations of the causal rules used to connect the causes to the outcome: if all the rules connecting the causes to the outcome are predictive, the causes are considered predictive.

Timeliness: A prediction is only useful if it gives enough warning to let us deal better with the predicted event. For example, suppose a NASA engineer explains the explosion of the Space Shuttle Challenger by "the boosters burned through, allowing flames to reach the main fuel tank, causing explosion." That explanation lets us expect an explosion the next time the boosters burn through, but by then it will be too late to abort the launch. However, an engineer who realizes that the burnthrough was caused by the booster seals being brittle, due to cold weather, could predict difficulties in cold weather, and avoid future explosions by refusing to approve launches in weather below a certain temperature.

For some goals, timeliness requirements depend on information within the explanation. For example, if the goal is prevention of a bad outcome, the explainer needs to predict the outcome while there is still time to block one of the causes. In this case, ACCEPTER determines from the explanation the amount of warning needed, by examining the explanation's belief-support chain to find the earliest causes the explainer can prevent. For other goals, factors external to the explanation determine the needed amount of warning. For example, someone hoping to profit from predicting stock prices will need to predict their fluctuations in time to buy or sell before the change takes place. The warning needed is unrelated to why the stock price changed; it only depends on how fast orders are processed. Although ACCEPTER has no mechanism for deciding needed timeliness for these goals, the user can make ACCEPTER's judgement reflect their needs by simply inputing the desired amount of warning.

To calculate how much warning one of the causes gives, ACCEPTER adds up temporal separations along the links of the explanation's belief-support chain, to find out how far in advance of the outcome the cause probably occurred.

Distinctiveness: As discussed above, distinctiveness judges whether a surprising event was itself caused by something surprising. If so, that surprising cause may be useful as a predictive feature of the outcome.

ACCEPTER judges a cause's distinctiveness by checking whether it deviates from stereotype-based expectations that were in effect before the anomaly. These expectations include expectations triggered by application of schemas for standard events in a given context, such as the MOP for standard events in a restaurant meal, and stereotypes for the activities of broad classes of actors (*e.g.*, that athletes in training should avoid wild parties). (For a description of these stereotypes and how they are applied, see (Leake, 1989a) or (Leake, in press).) To determine whether a fact is distinctive in a context, the program uses its routine understanding process to generate the expectations for that context, and compares the fact to those expectations.

The following example shows output from ACCEPTER's formation of distinctiveness judgements.[4] The explanation being considered is that Challenger's explosion was caused by the combination of the launch (which enabled the booster flames) and cold weather (which caused the seals' brittleness). The launch is routine, so it is not a useful predictive feature, even though it is causally relevant:

```
Checking whether CHALLENGER'S SPACE-LAUNCH satisfies
test(s) for "DISTINCTIVENESS".

   Applying test for DISTINCTIVENESS to CHALLENGER'S SPACE-LAUNCH.

      Using routine understanding to check whether CHALLENGER'S
      SPACE-LAUNCH is standard in context of CHALLENGER'S
      ROCKET-STRUCTURE.

         Building up new memory context with expectations
         from CHALLENGER'S ROCKET-STRUCTURE.

         Integrating CHALLENGER'S SPACE-LAUNCH into
         that context.

      CHALLENGER'S SPACE-LAUNCH satisfies the role-filling
      pattern "ROCKETs routinely fill role SPACECRAFT
      in SPACE-LAUNCH", so it's routine.

   ... test failed.
```

However, the cold weather is unexpected, so it is judged distinctive, allowing problems to be predicted on future cold days.

Knowability: No matter how early a predictive event may be, it will not help us to predict unless we can find out that it occurred. For example, we know that takeover announcements

[4]Minor editing of the output has been done for readability.

cause stock price increases, so it would be helpful to predict them long enough in advance to buy stock. The decision to start a takeover happens at least a few hours before the announcement, so it is timely, and it is also predictive of the announcement. Unfortunately, it is usually useless for prediction: the takeover decision is kept secret, so the investor will not be able to find out about it until the announcement takes place. *Knowability* measures how easy it is to know when an event occurs. Predictive knowledge needs to be indexed under features that the system is likely to know about, or that it can check for routineness with reasonable cost.

ACCEPTER distinguishes between three levels of knowability: *observable* causes, which are likely to be noticed in routine understanding, *testable* causes, that could be known by carrying out testing plans in the system's memory, and *undetectable* causes, which cannot be detected by known tests, even if the outcome is important enough to make testing worthwhile. (Of course, different tests can require very different levels of effort, and this could be added to the tests' representation, to allow ACCEPTER to make finer-grained distinctions.) Depending on the importance of predicting the outcome, the user selects which level of knowability to require. ACCEPTER uses the heuristics below to decide whether causes are observable or testable. Causes that are neither observable nor testable are assumed to be undetectable.

Judging observability: Nodes in ACCEPTER's memory net are marked with information on their usual observability to someone nearby. For example, most actions are observable to people who are present, but thoughts are not. Observability information about specific object and event features is also stored, indexed under the objects and events. For example, a person's hair color is usually observable, but his blood pressure is not. When no observability information is indexed under the object or event type, observability information is inherited from higher-level abstractions in ACCEPTER's memory.

One of the program's examples concerns the death of the racehorse Swale, who died unexpectedly at the peak of his career. When ACCEPTER evaluates usefulness of *death from horse race + heart defect* for predicting future deaths, it accepts the horse race as observable, because, in its memory net, horse racing is a specification of public performance, and most public performances can be observed. However, hearts are specifications of internal organs, and the physical state of internal organs is not usually observable, so not all the causes are observable— the explanation is not sufficient to allow prediction. (We note that this criterion is equivalent to past approaches for static operationality criteria (Mitchell et al., 1986), and suffers from the same problems; we point towards an alternative direction in the summary section below.)

Judging testability: Sometimes predicting an outcome is worth the effort of performing tests. If we know a low-cost plan for checking a predictive feature, we can perform that plan periodically to give warning of problems ahead. For example, an employee at Yale changed her behavior after explaining the cause of engine damage to her car:

> X burned out an engine by driving when her car was low on oil. After that incident, she started checking the oil level whenever she got gas, to correct low oil before damaging another engine.

People have standard tests for features of situations that have important effects, but are not directly observable. If we want to know the temperature outside, we might go to a window that has a thermometer mounted on its ledge. If we think that the battery of a car is dead, we might try turning on the radio or the lights. If we want to know whether a steak is sufficiently done, we can press it, and see how it springs back. ACCEPTER judges testability of an explanation's causes by searching its memory for testing plans that can determine whether the cause holds. (The tests in its memory are simply place holders; how to carry out the tests is not represented in the system). If ACCEPTER finds a plan, it judges the cause testable.

The output below shows ACCEPTER judging the explanation *death from horse race + heart defect* for Swale's death, to see if it could be used by an owner to predict and avoid deaths in other horses. The heart defect is not observable, but is testable by doing an electrocardiogram. Consequently, the explanation shows that an owner could predict problems: by having a vet do an EKG on the horses the owner buys, to find any heart defects.

```
Applying test for KNOWABILITY to SWALE'S HEART'S
HEREDITARY-DEFECTIVE ORGANIC-STATE.

   Searching up abstraction net for observability information.

   SWALE'S HEART'S HEREDITARY-DEFECTIVE ORGANIC-STATE
   is probably not observable, since it is a(n) ORGANIC-STATE
   of a(n) INTERNAL-ORGAN.

   Searching up abstraction net for pointers to standard tests.

   SWALE'S HEART'S HEREDITARY-DEFECTIVE ORGANIC-STATE
   is testable, since ORGANIC-STATEs of HEARTs can
   be detected by the standard test ELECTROCARDIOGRAM.

   ... test passed.
```

7.2 Evaluation dimensions for repair

If the anomaly that prompts explanation is a device failure, we may wish to repair the device aspects that caused the failure. This prompts the explanation purpose of finding repair points. Four evaluation dimensions are important when checking if an antecedent is a good repair point. The most obvious requirements are that the antecedent have *causal force* (it must have caused the bad device behavior), and that it be *repairable* by the explainer. However, not all causes are worth repairing. If someone was carrying a television, and tripped on some uneven steps, causing him to drop and damage it, the condition of the steps would be a cause of the damage, and might be repairable. However, repairing the steps would not fix the television. The cause to repair must be one whose presence is *predictive* of the device failure's recurrence, according to the predictiveness criteria described above.

Finally, even if there is a continuing problem that we can repair, it will not help if something else will cause the problem again as soon as we fix it. For example, dropping the television might have caused a power supply defect, burning out a fuse. Fixing the fuse by itself is pointless: the bad power supply will simply burn out the replacement. Thus we need to fix the power supply as well. This example shows that the explanation needs to trace back to a cause with *independence* from prior causes. To judge independence, ACCEPTER simply assumes that a cause is independent from prior causes unless it is caused by the current state of some object. Causal force and repairability are discussed in more detail below.

Causal force: ACCEPTER's inference rules distinguish two classes of connections between events (or states). One event *causes* a second if it actually brings about the second event, as when a heart attack causes death. An event is *predictive* of another, without necessarily being a cause, if the first prompts an expectation for the second. For example, if we are shocked by an awful meal in a restaurant, someone might explain it by saying "all New Haven restaurants are bad." The explanation may let us predict future bad meals, but it doesn't say what causes the low quality. ACCEPTER's inference rules are annotated with whether they describe a causal connection, or a predictive one (causal connections that hold in a given situation may be non-predictive, if they trace an unlikely outcome: drinking milk causes an allergic reaction in some people, but we do not normally predict that reaction).

Repairability: Finding the cause of a bad state only helps repair if we also know how to fix it. For example, tracing an engine failure back to some obscure component of the transmission will be little use to most people who are trying to do a road-side emergency repair. However, if they find that a wire has shaken loose, or a hose has come unattached, they can make the repair.

To judge repairability of a device, ACCEPTER searches memory for a repair plan indexed under the device, and the state being considered as a potential repair point. If a television had an anomalously bad picture, ACCEPTER would look at the causes of the bad picture given by the explanation, and see if it could retrieve a plan to fix any of them. For example, if the explanation attributed the problem to bad atmospheric conditions, it would be unlikely to find a repair plan. However, if the problem were caused by the antenna pointing the wrong way, a plan for fixing orientations of small objects— grabbing them and moving them— could be retrieved and applied.

The output below gives a sample of ACCEPTER's repairability judgements. When it evaluates the explanation that John's "brand A" car was defective because of a bad transmission, it first checks to see whether it can retrieve a plan for repairing the transmission.

It finds no stored repair plans for transmissions, or for any of the abstractions of transmissions in its memory:

```
Applying test for REPAIRABILITY to TRANSMISSION-743'S
LOW MECHANICAL-CONDITION.

      Searching up abstraction net for pointers to standard
      repair plans.

   ... test failed.

   TRANSMISSION-743'S LOW MECHANICAL-CONDITION is probably
   not repairable, since no standard repair plans are
   stored under any of its abstractions.
```

Since ACCEPTER finds no standard repair plans, it checks whether later steps in the belief-support chain are repairable. The XP's belief-support chain shows that the engine defect is caused by both the fact that the transmission is defective, as above, and that the transmission is a component of the engine. Although ACCEPTER finds no repair strategy specifically directed towards fixing the fact that a transmission is an engine component, it does find a strategy for repairing any component relationship that causes problems: replace the component. This strategy suggests that replacing the transmission is a possible repair plan:

```
Applying test for REPAIRABILITY to TRANSMISSION-743'S
PART-OF-RELATIONSHIP to BRAND-A'S ENGINE.

   Checking repairability of features of TRANSMISSION-743'S
   PART-OF-RELATIONSHIP to BRAND-A'S ENGINE.

      Searching up abstraction net for pointers to standard
      repair plans.
```

```
      BRAND-A'S ENGINE AS CONTAINER OF TRANSMISSION-743'S
      PART-OF-RELATIONSHIP to BRAND-A'S ENGINE is repairable, since
      CONTAINERs of PART-OF-RELATIONSHIPs can usually be repaired
      by the standard plan REPLACE-COMPONENT.

   ... test passed.

... Detail is acceptable.
```

7.3 Evaluation dimensions for control

When an anomalous state or event is undesirable, an understander may want to prevent it in the future. One plan for this goal is to block some of the event's causes directly; another is to discourage any actors responsible for the situation from contributing to its recurrence. (Likewise, if an outcome is desirable, an explainer may wish to find how to achieve it, either directly or by influencing others, but we will not consider that purpose here.)

Directly preventing future occurrence of an event involves finding premises with *causal force*— that cause the outcome— and that the explainer can block (*blockability*). An explanation that shows blockable causes can be used in two ways: first, the explainer may be able to permanently disable one of the causes, so that any repetition of the same causal chain is impossible. (For example, if a house was burglarized because a thief could enter through an open window, the victim can block recurrence by keeping the windows shut.) Second, if it is impractical to permanently block any of an event's causes (it may be too unpleasant to keep windows closed in the summer), explanation needs to show not only how the outcome could be blocked, but also how to predict a specific instance of the outcome long enough in advance to block the problem in the current instance. This procedure is basically the *anticipate and avoid* strategy suggested by Hammond for avoiding failures in case-based planning (Hammond, 1989). We discuss below some heuristics for judging blockability, and how they are applied to judge explanations for anticipating and avoiding an undesirable outcome.

Blockability: Deciding what an explainer can prevent is difficult; things that seem uncontrollable at first glance may actually be easy to influence. For example, if a picnic is rained out, it is natural to accept the rain as beyond our control, but it might might be avoidable by scheduling the picnic during the dry season. Nevertheless, a very simple heuristic is sufficient for judging an actor's ability to control many events: we can assume that actors who voluntarily fill actor roles in events, or provide other role-fillers for them, can probably block those events.

A real-world example of the importance of actor roles in blocking an outcome comes from the Challenger explosion. Each area of NASA attempted to find ways to prevent similar situations arising, and different divisions focused on different causes of the problem, depending on the explainers' influence. After the explosion, according to the astronaut John Creighton:

> [Everyone had a different idea] of what we didn't think worked. If you're an engine man, you want the engine fixed; if you're in charge of something else, you want that fixed.[5]

ACCEPTER's blockability checks consider three ways a person can be involved in an event: as an *actor* who is immediately involved; as a *director* of the action, who is not immediately involved, but who has authority over its progression; and as a *supplier* of the objects or actors that the director selects.

Actors in an event may be able to block it by refusing to participate. Directors may block it by ordering actors under their control not to participate, by controlling the setting for events (locale, time of occurrence, or features of the environment), by changing ways of selecting the objects used, to avoid using objects that are particularly likely to contribute to a bad outcome, or by changing object suppliers to avoid such objects. Suppliers can also block outcomes by controlling the objects they make available. Table 2 sketches how each of these means of explainer control entered into strategies for preventing another space shuttle disaster.

ACCEPTER's basic procedure for deciding blockability by a given person is to use the explanation to find a person's involvement in the outcome, and then see if that person is

[5] *Newsweek*, October 10, 1988.

Table 2

Strategies for blocking recurrence of the Challenger disaster

Real-life application of the prevention strategies for control by actors, directors and suppliers.

Control by actors

- **Refuse to participate**
  After the explosion, astronauts said that they wouldn't fly until the shuttles were repaired.

Control by directors

- **Change setting for the event**
  - **Change locale for the event**
    Some people suggested changing the launch site to Hawaii, where the weather is warmer.
  - **Change the time of the event**
    Engineers advised delaying launching when the weather was below 53 degrees.
  - **Change features of the environment**
    NASA installed heaters on the launch pad to warm boosters before launch.
- **Change role-filler choice**
  - **Apply tests to rule out bad objects**
    Some people suggested NASA should inspect the booster seals after delivery of the boosters. NASA rejected this because tests would require disassembly that might introduce new defects.
  - **Change supplier of object**
    Some advocated stopping using boosters made by Morton Thiokol.

Control by suppliers

- **Block access to a class of role-filler**
  New shuttle manufacturing was frozen by congress while the Challenger investigation went on, in order to avoid repeats of the explosion.
- **Change design**
  Morton Thiokol redesigned the boosters, with better seals.

involved as an actor, director, or supplier of an object (which ACCEPTER checks by seeing if that person is its owner). If so ACCEPTER assumes that the person could block the event. It then checks whether timely prediction is possible, to allow the person to anticipate the outcome and exert control to block its occurrence.

For example, the output below shows how ACCEPTER decides that the space shuttle's astronauts could prevent future explosions. The antecedent being considered is Challenger's launch, and ACCEPTER first verifies that the launch is a cause of the explosion, to see if blocking the launch could be an effective way to prevent the outcome:[6]

```
Checking whether some antecedent satisfies the following tests:
CAUSAL FORCE TEST (does fact cause consequent?), and
BLOCKABILITY + TIMELINESS (can CHALLENGER'S SPACE-LAUNCH's
ASTRONAUT block after predicting outcome?).

   Applying test for CAUSAL FORCE (does fact cause
   consequent?) to CHALLENGER'S SPACE-LAUNCH.

      Checking the connection between CHALLENGER'S SPACE-LAUNCH
      and CHALLENGER'S EXPLOSION.

         Testing if CAUSAL-MOP-SCENE:LAUNCH->IGNITION
         satisfies test for LINK CAUSATION.
         ... test passed.

         Testing if CAUSAL-MOP-SCENE:IGNITION->HIGH-PRESSURE
         satisfies test for LINK CAUSATION.
         ... test passed.

         Testing if
         BRITTLE-SEAL+CONTAINER-SEAL+CONTENTS-PRESSURE=>CONTAINER-EXPLOSION
         satisfies test for LINK CAUSATION.
         ... test passed.
```

[6]Each causal link is represented as an XP in ACCEPTER's memory; the output only shows the names of the links.

```
        Testing if EXPLOSION-IN-COMPONENT=>EXPLOSION-IN-WHOLE
        satisfies test for LINK CAUSATION.
        ... test passed.

    ... All links are acceptable.

  ... test passed.
```

Since the explanation shows that the launch causes a rocket booster's explosion, which in turn causes the shuttle as a whole to explode, blocking the launch would block the explosion. This identifies it as a cause that would be worthwhile to block, so ACCEPTER checks whether an astronaut could block it. It first checks whether the astronaut controls the availability of objects required to fill roles in the launch (which the astronaut does not), or whether the astronaut has control over whether to participate in the launch:

```
Checking if CHALLENGER'S SPACE-LAUNCH is blockable.

   Checking if CHALLENGER'S SPACE-LAUNCH's ASTRONAUT
   controlled outcome by controlling a needed object.
   ... No actor-controlled objects found.

   Checking if CHALLENGER'S SPACE-LAUNCH's ASTRONAUT
   controlled outcome through a role he filled.

   CHALLENGER'S SPACE-LAUNCH's ASTRONAUT might
   have been able to prevent CHALLENGER'S SPACE-LAUNCH,
   by refusing to be its ASTRONAUT, since that
   is a voluntary actor role.
```

From this information, ACCEPTER concludes that the launch is a blockable cause. The next question is whether an explosion can be predicted before the launch, so the astronaut can stop the launch in time. Because the last cause of the explosion that the astronaut controls is the launch, preventing future explosions is only possible if the explanation shows how to predict explosion before the launch occurs. To know how much warning would be needed, ACCEPTER first uses the explanation to determine how far in advance of the explosion the launch occurs, and then checks whether the explanation allows prediction with that amount of warning. (ACCEPTER uses a coarse-grained temporal representation, with temporal separations represented as NONE, MINUTES, HOURS, DAYS, WEEKS and YEARS.)

```
Checking if XP allows prediction of outcome before
CHALLENGER'S SPACE-LAUNCH occurs.

   Calculating amount of warning needed to predict
   before CHALLENGER'S SPACE-LAUNCH occurs.

      CHALLENGER'S SPACE-LAUNCH leads to ROCKET-IGNITION-41
      immediately.

      ROCKET-IGNITION-41 leads to GAS-42'S HIGH PRESSURE
      MINUTES afterwards.

      GAS-42'S HIGH PRESSURE leads to SOLID-ROCKET-43'S
      EXPLOSION MINUTES afterwards.

      SOLID-ROCKET-43'S EXPLOSION leads to CHALLENGER'S
      EXPLOSION immediately.

Predicting before CHALLENGER'S SPACE-LAUNCH requires
finding a predictive feature at least MINUTES
before CHALLENGER'S EXPLOSION.

Checking detail for predicting CHALLENGER'S
EXPLOSION MINUTES before it happens.
```

Since the launch is minutes before the explosion, an explanation for prevention must allow prediction to be done at least minutes before the explosion as well. To decide whether the explanation does so, ACCEPTER applies the tests for predictiveness described in the previous section. The coldness of the seal occurs early enough to provide warning, and is distinctive, predictive, and knowable. Consequently, ACCEPTER accepts the explanation as being useful for an astronaut's prediction, allowing him to monitor and intervene: on cold days, the astronaut can refuse to fly.

7.4 Evaluation dimensions for actors' contributions to an outcome

To maintain effective performance, any actor in the world needs ways of judging the appropriateness of its actions. When a surprisingly good or bad outcome occurs, a system may benefit by explaining who contributed to the event. By identifying the actors involved, and ascribing praise or blame for their roles, it may be possible to influence their future behavior. Also, by learning about the (possibly unexpected) good and bad ramifications of others' behavior, the system can learn about possible problems and opportunities to consider when planning its own future actions, forming new strategies for guiding its own behavior.

These goals prompt the explanation purpose of identifying actors' contributions to an action, and ascribing praise or blame. This can be done according to the *desirability* of the outcome, and of actions leading to it. In addition, an actor can also be blamed, even if it was impossible to predict or control an outcome, if that actor contributed to the outcome through an undesirable act. The bad result strengthens the act's original proscription: for example, we might blame a drug dealer for an addict's death by overdose, even if deaths from overdose are relatively unlikely.

The output below shows ACCEPTER's evaluation of blame for the explanation that the racehorse Swale died because a trainer accidentally gave him an overdose of performance-enhancing drugs. It first checks whether the actor should have anticipated the problem, and could have prevented it:

```
Checking PERFORMANCE-DRUG-OVERDOSE-BY-TRAINER for ASSIGNING-BLAME.

   Applying test for BLOCKABILITY + TIMELINESS (can ATHLETIC-TRAINER-1497
   block after predicting outcome?) to ATHLETIC-TRAINER-1497'S
   M-PERFORMANCE-ENHANCEMENT-DRUGS.

      ATHLETIC-TRAINER-1497 could have prevented initiation
      of ATHLETIC-TRAINER-1497'S M-PERFORMANCE-ENHANCEMENT-DRUGS
      since its DRUGGER controls initiation.
```

However, ACCEPTER determines that performance enhancing drugs are not predictive of fatal overdose, so the trainer would not have expected the outcome. Consequently, the trainer cannot be blamed with intentionally causing Swale's death.

ACCEPTER then checks whether giving performance-enhancing drugs is in itself undesirable. In ACCEPTER's memory net, administering performance enhancing drugs is a specification of the category for illegal actions, so it judges the drugging as undesirable. Giving performance enhancing drugs is annotated as usually being a voluntary action, so ACCEPTER decides that the explanation gives cause for blame.

7.5 Summary of evaluation dimensions

Table 3 summarizes our evaluation dimensions, and the simple heuristics ACCEPTER uses to test along them. For a discussion of the strategies not discussed here, and how they combine to satisfy the information requirements for other purposes, see (Leake, in press).

It should be noted that while the system uses dynamic criteria to evaluate for some dimensions (*e.g.*, timeliness, distinctiveness, and repairability), its criteria for other dimensions are static (*e.g.*, the observability component of knowability). We did not investigate dynamic characterizations of all dimensions, simply because our main effort was devoted to investigating the relationship between evaluation goals and the needed dimensions. However, we strongly agree with the arguments in (DeJong & Mooney, 1986) that such criteria must be able to dynamically take into account current system knowledge. Richer and more dynamic characterizations of ACCEPTER's dimensions involve many issues for future research. For example, a richer characterization of knowability criteria would have to involve reasoning about competing ways to gather information, their likely costs, and their chance of success; (Hunter, 1990) discusses some of these directions in the context of knowledge planning.

8 The value of goal-based evaluation

The preceding sections sketched our approach to goal-based explanation evaluation. In this section, we put that approach in perspective, discussing how a goal-based evaluation module can contribute to an overarching explanation-based system. We advance three main

Table 3.

ACCEPTER's heuristics for testing an event or state A, stated as a cause in an explanation, along each evaluation dimension.

Timeliness	Trace along the XP's derivation of outcome from A, summing the standard delays from antecedent to consequent of each inference rule.
Distinctiveness	Use routine understanding mechanism to build up standard expectations from the background situation, and integrate A into that context. A is distinctive if it is unexpected or anomalous in that context.
Knowability	Check whether A, or its abstractions in memory, is annotated as usually observable, or search for a plan in memory that can be applied to determine whether A has occurred.
Predictive power/causal force	Check whether the XP derives the outcome from A by a sequence of predictive links (respectively, causal links).
Independence	Check if A is caused by a state still in effect. If not, assume A is independent of prior causes.
Repairability	Search memory for a standard repair plan for any abstraction of A.
Controllability	Look for direct involvement of the actor as an actor, director, or (for blockability only) supplier of A.
Desirability	Check if action described by A is a specification of illegal-activity.

points: First, that goal-based evaluation allows learning to be focused more effectively than in previous approaches. Second, that it facilitates construction of useful explanations in a case-based explanation, and, more generally, makes it possible to reliably build explanations for systems with multiple types of goals. Third, that it offers a valuable new way to deal with the imperfect theory problem in explanation-based learning, by allowing an explanation-based system to learn effectively despite certain imperfections in its domain theory.

Focusing explanation towards knowledge gaps: Our explanation process centers around explaining anomalies. In our discussion of evaluation for routine understanding, we described ACCEPTER's requirement that explanations account not only for the event, but for why expectations failed. This differs from approaches such as (Mooney, 1990), which concentrate on accounting for why the event occurred. The difference is important to deciding which beliefs to repair, and what to learn from a new situation. If we want to explain why basketball team A defeated team B in a close game, the explanations we will seek will be quite different if we expected B to win (in which case we might refer to injuries of B's players), or if we expected A to win by a large margin (in which case we might refer to team A being over-confident). In general, goal-based evaluation allows learning to focus on the aspects of a situation that are important: those relevant to system goals. ACCEPTER can judge a single explanation acceptable for one purpose, but not for another. for example, one of ACCEPTER's stories involves the recall of a defective car. When the system's purpose is to find ways of predicting defects, it accepts the explanation that the manufacturer has bad quality control; when its purpose is to find repair points, it rejects that explanation as insufficient, and accepts an explanation that points to the specific part that failed.

Guiding explanation construction in case-based explanation: How to control explanation construction is a difficult issue that has received surprisingly little attention. Many explanation-based systems construct explanations by undirected chaining, which risks overwhelming inference cost for any but the most simple explanations (Rieger, 1975). The SWALE system addresses this problem by using case-based reasoning to build new explanations. Rather than explaining from scratch, it applies prior experience by retrieving and adapting explanations from similar past episodes. The case-based approach facilitates construction of complicated explanations, by re-using prior reasoning whenever possible. In addition, when new and old situations are quite similar, the case-based approach can generate explanations that are more likely to be valid: rather than being arbitrary rule combinations, the hypotheses it builds are supported by prior experience.

Unlike traditional explanation-based systems, which can rely on all candidate explanations being in the proper form for their single task (*e.g.*, showing sufficient conditions for concept membership for the concept recognition task in (Mitchell et al., 1986)), a case-based explainer reuses explanations that may have been constructed in very different contexts, and for very different goals. Consequently, unlike single-purpose systems that can tailor explanation construction to their goals, and rely on initial explanation construction to provide an appropriate type of explanation, a case-based explainer cannot be assured of starting with an explanation that includes appropriate information (Leake, 1989b). Consequently, it must have a means for determining whether an explanation contains the specific information it needs, and to identify gaps to fill through explanation adaptation. Our evaluation process provides that guidance, enabling a case-based explainer to reliably use cases from a multi-purpose case library. This guidance could also be applied to any explanation construction system that must build explanations for multiple purposes.

Dealing with the imperfect theory problem: Explanation-based learning research traditionally considers explanations to be deductive proofs of concept membership (Mitchell et al., 1986). In this framework, the structure of the proof assures that the explanation points to a sufficient set of factors for concept membership. However, if we seek to explain real-world events, no explanation can include *all* the causally-relevant factors. As Mitchell et al. observe, real-world domain theories are often both incomplete and intractable.

Responses to the imperfect theory problem propose methods for repairing the theory's defects. For example, (Dietterich & Flann, 1988) suggests that when a domain theory allows multiple incompatible explanations, induction over explanations for a set of training examples can be used to find a specialized domain theory that explains the positive examples, but none of the negative ones. (Rajamoney, 1988) advocates experimentation to determine how to extend or repair a domain theory in response to problems such as multiple incompatible explanations, or the inability to construct any complete explanation. Unfortunately, these approaches are not always practical: it may be necessary to act without having observed enough examples to repair a domain theory using induction over explanations, and may be infeasible to perform the experiments needed to do experimentation-based theory revision.

However, the inability to construct a complete explanation, or to rule out incompatible alternative explanations, does not necessarily interfere with human use of explanation. People often accomplish their goals using fragmentary explanations. Someone with little automotive knowledge may buy a used car, and notice that it often fails to start on cold days. Although the buyer would probably be unable to generate a complete explanation, it would still be possible to form the partial explanation that cold is one of the causes the problem. This hypothesis does not give sufficient conditions for the failure, since the car sometimes starts despite the cold, and it would probably not be feasible for a novice to identify the other factors that are involved. Nevertheless, the partial explanation is still useful: it can be used to decide to keep the car in the garage on cold nights.

Likewise, choosing between competing incompatible explanations may not be possible, or may not be worthwhile even if it is possible: a policeman arresting a murderer does not need to choose between the incompatible explanations that the murderer was motivated by jealousy, or by financial gain, as long as the murderer proposed by all competing explanations is the same person. Thus imperfections in theory do not have to interfere with the use of an explanation. By giving a principled account of which aspects of explanations are important, and which are not, our goal-based evaluation criteria determine when an imperfect explanation can simply be applied, without having to extend the theory to resolve defects or choose between competing alternatives. Allowing explanation-based processing of partial and imperfect explanations significantly extends the circumstances in which a system can apply explanation-based techniques.

9 Conclusion

Evaluation of explanations is a dynamic process: an explanation's goodness depends on whether it satisfies the explainer's current needs for information. An explanation that is good for one purpose may be irrelevant to another, or may give inadequate information for it. For some purposes, a very vague or incomplete explanation may be sufficient; for others, certain aspects will need to be described in great detail.

In order to judge the goodness of an explanation, we need to know how it will be used, and what information that use requires. This paper sketches how goals prompt plans that in turn trigger basic explanation purposes, and traces their information requirements. To judge whether explanations satisfy these requirements, ACCEPTER uses heuristics for judging causes along combinations of simple evaluation dimensions.

The information required for the explanation purposes we have described can be characterized in terms of nine evaluation dimensions (timeliness, knowability, distinctiveness, predictive power, causal force, independence, repairability, controllability, and desirability), giving a compact way of describing needs for information, and suggesting classes of evaluation heuristics to refine in future research.

By providing a goal-sensitive way to judge the information contained in a particular explanation, ACCEPTER's approach extends considerably the applicability of explanation-based processing, giving a means for deciding whether to accept and learn from an imperfect explanation. It is a first step towards a model of explanation as a dynamic, goal-driven process, integrated fully with the system tasks it serves.

References

Ahn, W., Mooney, R., Brewer, W., & DeJong, G. (1987). Schema acquisition from one example: Psychological evidence for explanation-based learning. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 50–57). Seattle: Cognitive Science Society.

Collins, G., & Birnbaum, L. (1988). An explanation-based approach to the transfer of planning knowledge across domains. *Proceedings of the 1988 AAAI Spring Symposium on Explanation-based Learning* (pp. 107–111). Stanford: AAAI.

Cullingford, R. (1978). *Script Application: Computer Understanding of Newspaper Stories* (Technical Report 116). New Haven: Yale University Computer Science Department.

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

Dietterich, T., & Flann, N. (1988). An inductive approach to solving the imperfect theory problem. *Proceedings of the 1988 AAAI Spring Symposium on Explanation-based Learning* (pp. 42–46). Stanford: AAAI.

Granger, R. (1980). *Adaptive Understanding: Correcting Erroneous Inferences* (Technical Report 171). New Haven: Yale University Computer Science Department.

Hammond, K. (1989). *Case-Based Planning: Viewing Planning as a Memory Task.* San Diego: Academic Press.

Hanson, N. (1961). *Patterns of Discovery.* Cambridge: Cambridge University Press.

Heider, F. (1958). *The Psychology of Interpersonal Relations*, volume XV of *Current Theory and Research in Motivation.* New York: Wiley.

Hobbs, J., Stickel, M., Appelt, D., & Martin, P. (1990). *Interpretation as abduction* (Technical Report 499). Menlo Park: SRI International.

Hunter, L. (1990). Planning to learn. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 261–268). Cambridge: Cognitive Science Society.

Kass, A. (1986). Modifying explanations to understand stories. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 691–696). Amherst: Cognitive Science Society.

Kass, A., & Leake, D. (1988). Case-based reasoning applied to constructing explanations. Kolodner, J. (Ed.), *Proceedings of the Case-Based Reasoning Workshop* (pp. 190–208). Clearwater Beach: DARPA.

Kedar-Cabelli, S. (1987). Formulating concepts according to purpose. *Proceedings of the Sixth Annual National Conference on Artificial Intelligence* (pp. 477–481). Seattle: AAAI.

Keller, R. (1987). *The Role of Explicit Contextual Knowledge in Learning Concepts to Improve Performance.* PhD thesis, Rutgers University.

Keller, R. (1988). Operationality and generality in explanation-based learning: Separate dimensions or opposite endpoints? *Proceedings of the 1988 AAAI Spring Symposium on Explanation-based Learning* (pp. 153–157). Stanford: AAAI.

Kelley, H. H. (1967). Attribution theory in social psychology. Levine, D. (Ed.), *Nebraska Symposium on Motivation.* Lincoln, NE: University of Nebraska Press.

Kolodner, J. (1984). *Retrieval and Organizational Strategies in Conceptual Memory.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lalljee, M., & Abelson, R. (1983). The organization of explanations. Hewstone, M. (Ed.), *Attribution Theory: Social and Functional Extensions.* Oxford: Blackwell.

Lalljee, M., Watson, M., & White, P. (1982). Explanations, attributions, and the social context of unexpected behavior. *European Journal of Social Psychology*, *12*,17–29.

Leake, D. (1988a). Evaluating explanations. *Proceedings of the Seventh National Conference on Artificial Intelligence* (pp. 251–255). Minneapolis: AAAI.

Leake, D. (1988b). Using explainer needs to judge operationality. *Proceedings of the 1988 AAAI Spring Symposium on Explanation-based Learning* (pp. 148–152). Stanford: AAAI.

Leake, D. (1989a). Anomaly detection strategies for schema-based story understanding. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 490–497). Ann Arbor: Cognitive Science Society.

Leake, D. (1989b). The effect of explainer goals on case-based explanation. Hammond, K. (Ed.), *Proceedings of the Case-Based Reasoning Workshop* (pp. 290–294). Pensacola Beach: DARPA.

Leake, D. (in press). *Evaluating Explanations.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Leake, D., & Owens, C. (1986). Organizing memory for explanation. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 710–715). Amherst: Cognitive Science Society.

Lebowitz, M. (1980). *Generalization and Memory in an Integrated Understanding System* (Technical Report 186). New Haven: Yale University Computer Science Department.

Mackie, J. (1965). Causes and conditions. *American Philosophical Quarterly*, (4).

Mehlman, R., & Snyder, C. (1983). Excuse theory: A test of the self-protective role of attributions. *Journal of Personality and Social Psychology, 49*(4),994–1001.

Minton, S. (1988). Quantitative results concerning the utility of explanation-based learning. *Proceedings of the seventh national conference on artificial intelligence* (pp. 564–569). Saint Paul: AAAI.

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

Mooney, R. (1990). *A General Explanation-based Learning Mechanism and its Application to Narrative Understanding.* San Mateo, CA: Morgan Kaufmann.

Moore, J., & Swartout, W. (1989). A reactive approach to explanation. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* (pp. 1504–1510). Detroit: IJCAI.

Mostow, D. J. (1983). Machine transformation of advice into a heuristic search procedure. Michalski, R., Carbonell, J., & Mitchell, T. (Eds.), *Machine Learning: An Artificial Intelligence Approach*. Cambridge: Tioga Publishing Company.

Ng, H., & Mooney, R. (1990). On the role of coherence in abductive explanation. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 337–342). Boston: AAAI.

Paris, C. (1987). Combining discourse strategies to generate descriptions to users along a naive/expert spectrum. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (pp. 626–632). Milan, Italy: IJCAI.

Pazzani, M. J. (1988). Selecting the best explanation for explanation-based learning. *1988 Spring Symposium Series: Explanation-Based Learning* (pages 165–169). Stanford: AAAI.

Rajamoney, S. (1988). Experimentation-based theory revision. *Proceedings of the 1988 AAAI Spring Symposium on Explanation-based Learning* (pp. 7–11). Stanford: AAAI.

Rajamoney, S., & DeJong, G. (1988). Active explanation reduction: An approach to the multiple explanations problem. *Proceedings of the Fifth International Conference on Machine Learning* (pp. 242–255). Ann Arbor: Machine Learning.

Ram, A. (1989). *Question-driven understanding: An integrated theory of story understanding, memory and learning* (Technical Report 710). New Haven: Yale University Computer Science Department.

Rieger, C. (1975). Conceptual memory and inference. *Conceptual Information Processing.* Amsterdam: North-Holland.

Riesbeck, C. (1981). Failure-driven reminding for incremental learning. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence* (pp. 115–120). Vancouver, B.C.: IJCAI.

Schank, R. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, *3*(4),552–631.

Schank, R. (1982). *Dynamic Memory: A Theory of Learning in Computers and People.* Cambridge: Cambridge University Press.

Schank, R. (1986). *Explanation Patterns: Understanding Mechanically and Creatively.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R., & Abelson, R. (1977). *Scripts, Plans, Goals and Understanding.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R., & Leake, D. (1989). Creativity and learning in a case-based explainer. *Artificial Intelligence*, (40). Also in Carbonell, J. (Ed.), 1990, *Machine Learning: Paradigms and Methods*, Cambridge: MIT Press.

Segre, A. M. (1987). On the operationality/generality tradeoff in explanation-based learning. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (pp. 242–248). Milan, Italy: IJCAI.

Shortliffe, E. (1976). *Computer-based medical consultations: MYCIN.* New York: American Elsevier.

Snyder, C., Higgens, R., & Stucky, R. (1983). *Excuses: Masquerades in Search of Grace.* New York: Wiley.

Souther, A., Acker, L., Lester, J., & Porter, B. (1989). Using view types to generate explanations in intelligent tutoring systems. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 123–130). Ann Arbor: Cognitive Science Society.

Swartout, W. (1983). XPLAIN: a system for creating and explaining expert consulting programs. *Artificial Intelligence*, (21),285–325.

Thagard, P. (1989). Explanatory coherence. *The Behavioral and Brain Sciences*, *12*(3),435–502.

Van Fraassen, B. (1980). *The Scientific Image*, chapter 5. Oxford: Clarendon Press.

Wilensky, R. (1983). *Planning and Understanding.* Reading, MA: Addison-Wesley.