Using Goals and Experience to Guide Abduction[1]

David B. Leake
leake@cs.indiana.edu

Technical Report #359
Department of Computer Science, Indiana University
Lindley Hall 215, Bloomington, IN 47405

Abstract

Standard methods for abductive understanding are neutral to prior experience and current goals. Candidate explanations are built from scratch by backwards chaining, without considering how similar situations were previously explained, and selection of the candidate to accept is based on its likelihood, without considering the information needs beyond routine understanding. Problems arise when applying these methods to everyday understanding: The vast range of possible explanations makes it difficult to control the cost of explanation construction and to assure that the explanations generated will actually be useful.

We argue that these problems can be overcome by using goals and experience to guide both explanation generation and evaluation. Our work is within the framework of case-based explanation, which builds explanations by retrieving and adapting prior explanations stored in memory. We substantiate our model by describing mechanisms that enable it to effectively generate good explanations. First, we demonstrate that there exists a theory of anomaly and explanation that can guide retrieval of relevant explanations. Second, we present a plausibility evaluation process that efficiently detects conflicts and confirmations of an explanation's assumptions by prior patterns, making it possible to focus explanation adaptation when retrieved explanations are implausible. Third, we present methods for judging whether explanations provide the information needed to satisfy explainer goals beyond routine understanding. By reflecting experience and goals in the search for explanations, case-based explanation provides a practical mechanism for guiding search towards explanations that are both plausible and useful.

[1]The work described here was 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.

Contents

1 Introduction

Abduction is often described as "inference to the best explanation" (Harman, 1965). This description suggests two fundamental issues facing abductive reasoning systems: performing the inference to generate candidate explanations and selecting which of the resulting candidates to accept. In the growing body of research on automated abduction, standard methods have emerged for addressing both of these issues. For generating candidate explanations, the standard method is to build candidate explanations by backward chaining, starting from scratch, from the fact to be explained. For selecting the best explanation, the standard method is to compare candidates by context-independent criteria that focus on the explanation's structure.

This chapter argues that despite the successes of standard methods in the domains where they have been applied, they are insufficient for abductive reasoning in the rich world of everyday explanation. First, explanation construction by chaining from scratch is in principle extremely expensive, making it impractical for processing in unconstrained domains. Second, because standard explanation evaluation methods are neutral to specific explainer knowledge and goals, they may not result in the most plausible explanations or the most useful ones.

We advocate an alternative model that addresses both these problems by using goals and experience to guide abduction. The framework for our model is case-based explanation (Schank, 1986; Kass, 1986; Leake & Owens, 1986; Kass & Leake, 1988; Schank & Leake, 1989). Case-based explanation replaces explanation by chaining with a memory-based approach: It builds new hypotheses by retrieving stored explanations for prior cases, evaluating their applicability and usefulness for the new situation, and adapting them as needed to repair problems. The potential benefits of case-based explanation are twofold: generation of better candidate explanations, by favoring explanations supported by prior experience, and increased efficiency of explanation generation, by re-using the results of prior effort.

Realizing the benefits of case-based explanation depends on effective retrieval of relevant prior cases in memory: If retrieval is excessively expensive, or if the cases retrieved are inappropriate, retrieval and adaptation costs will nullify the advantages of reasoning from prior experience. The first major focus of this chapter is a theory of how to guide retrieval; we describe an indexing vocabulary for explanations of anomalies, examine its function, and argue that the vocabulary enables a case-based explanation system to focus search on relevant explanatory information.

The chapter's second main focus is how to evaluate candidate explanations. No matter how effective a case-based explanation system's retrieval may be, the retrieved explanation may need adaptation to fit the new circumstances, which depends on identifying specific flaws needing repair. Consequently, unlike standard explanation evaluation schemes that compare the *structure* of candidate explanations to decide a single global value for their goodness, our evaluation scheme focuses on the *content* of individual assumptions, allowing judgements both to provide a better estimate of likelihood, by reflecting system knowledge, and to identify specific points needing repair.

Our evaluation scheme also departs from traditional views in that context is crucial in its evaluation decisions. Standard methods seek a single "best" explanation, but in everyday situations, many valid explanations can be generated for a given phenomenon, each providing different information. Explanations that are useful for one purpose may be useless for others; evaluation must reflect the information that the explainer needs, which in turn depends on how the explanation will be used (Leake, 1988; Ram, 1990; Leake, 1990; Krulwich, Birnbaum, & Collins, 1990; Ram & Leake, 1991; Leake, 1992). Reflecting explainer needs in explanation evaluation depends on identifying the goals that can drive explanation, and we describe our theory of goal-based influences in the later part of the chapter. The theory described here is implemented in the testbed system ACCEPTER (Leake, 1992), a story understanding program that detects anomalous events in the stories it processes, characterizes the anomalies to facilitate explanation retrieval and adaptation, and evaluates candidate explanations for a range of tasks.

2 Issues in Everyday Abduction

Before discussing specifics of our model, we illustrate some of the general issues it addresses. Consider the problem of explaining the following statement: "John used a blowtorch to break into an automatic teller machine." Many candidate explanations could be generated, including:

- John needed money to pay back a loan shark for gambling debts.[2]
- John thinks that ATMs are easy targets.
- Mark couldn't do the break-in because he was sick.
- Crowbars aren't strong enough to open modern ATMs.
- The bank's security camera had been removed for repairs.
- New high-temperature torches can quickly melt the alloys used in ATMs.

[2]In our model, explanations are belief dependency chains tracing plausible reasoning; they derive the event or state being explained from a set of antecedents. However, we will often state the antecedents alone as a shorthand for the entire explanation. In this example, the entire explanation could include factors such as the fact that loan sharks place their victims under extreme duress to pay their debts, generating a very high-priority goal to obtain money; that robbery is a plan for obtaining money; that a step in robbery is gaining access to the money to be robbed, etc.

The range of possible explanations to generate and evaluate results in a twofold problem: focusing explanation search through the enormous number of possible candidates, and judging the explanations that are found in light of the very different focuses that explanations may take.

2.1 Controlling Explanation Construction

Abductive understanding systems generally build explanations by backwards chaining from the fact to be explained, with each new explanation constructed from scratch. The enormous cost of undirected chaining is well known, and many methods have been proposed for reducing the cost. One method is schema-based processing, which enormously constrains inference but which sacrifices the ability to account for events outside system schemas (Minsky, 1975; Schank & Abelson, 1977; Charniak, 1978; Cullingford, 1978). To accommodate novelty, this method can be augmented with a chaining process to allow the system to explain events outside of its schemas and to generate new schemas (Mooney, 1990).

Other methods concentrate on constraining the chaining process, by means such as combining of top-down and bottom-up processing (Wilensky, 1983), limiting the amount of chaining allowed (Mooney, 1990), heuristics to limit the branching factor of search (Hobbs, Stickel, Appelt, & Martin, 1990), and using marker-passing to propose candidate paths (Charniak, 1986; Norvig, 1989).

Case-based explanation is a third alternative. Case-based reasoning systems solve new problems by retrieving stored cases reflecting similar problems, and adapting the solutions of those problems to fit the new circumstances (see (Kolodner, 1988; Hammond, 1989; Bareiss, 1991) for a sampling of case-based reasoning research). A central motivation for case-based explanation is the idea that in a regular world, experience can facilitate explanation of novel events: explanations of similar prior situations can serve as a starting point when building new explanations.

For example, suppose an explainer wishes to explain why John performed the break-in instead of another criminal named Mark. There are many candidate explanations: John might have asked to do the break-in to prove himself to the gang; he might have asked to do it to earn the higher percentage of the profits his gang gave to members in risky roles; or the gang leader might have asked him to do it when he discovered that the ATM was a model for which John had special expertise. Many other explanations could be hypothesized as well, leaving a vast field of possible candidates.

However, explainers with experience in similar situations can control the explosion of possibilities by first considering what *does* happen, based on prior experience, rather than immediately attempting to consider everything that *might possibly* happen. This explanation process depends on memory to suggest explanations: For example, if the explainer knows that yesterday John replaced Mark because Mark was being followed by his parole officer, that explanation seems a very likely candidate for the current substitution as well. Psychological experiments provide support for the view that people use a reminding process to generate explanations, and that they give increased plausibility ratings for explanations suggested by remindings (Read & Cesa, 1991).

In addition to being applicable when very similar situations have been explained before, case-based explanation can apply in novel situations. With appropriate indexing of stored explanations, it can suggest relevant explanations even in situations quite different from those previously explained. For example, if no similar changes have been explained for previous crimes, retrieval of explanations generated for other domains may be useful. The explanation for a last-minute substitution for an actor in a play—that the scheduled actor is sick—could suggest illness as an explanation for the change in how the robbery is carried out. The ability to retrieve relevant explanations from other domains, combined with the ability to adapt prior explanations to new situations, provides flexibility unavailable in schema-based approaches while maintaining efficiency (Kass, 1990). In addition, as the system's library of explanations grows, it learns new cases that allow it to draw on a wider range of stored explanations and increasing the likelihood of having a relevant explanation in memory.

A final motivation for the case-based approach is that it facilitates generation of relevant explanations. Chaining-based approaches are neutral to how their results may be used, but different goals for using explanations require generating explanations providing different information. For example, although each explanation in our original list is a possible answer to the question "why?", it is clear that different alternatives will be needed for different understanders. We have already described some explanations appropriate to someone who is surprised that John did the robbery rather than Mark. For someone surprised by John taking the risk of robbery, an appropriate explanation might be "John needed the money to pay back a loan shark"—that explanation shows that John was forced to take the risk in order to avoid the greater risk of physical harm. Someone surprised by John's deviation from his usual *modus operandi* of using a crowbar would be satisfied by the explanation "crowbars aren't strong enough to open modern ATMs." In all these examples, the same event is being explained, but the focus is a different anomaly. Consequently, different explanations are needed. Because the anomaly to be addressed can be reflected in the indices used during explanation retrieval, case-based explanation makes it possible to focus explanation search on relevant explanations.

2.2 Judging candidate explanations

Choosing the most likely explanation: In abductive understanding and diagnosis systems, selection of the most likely explanation is based on Occam's razor, with explanations ranked by "minimality" according to some syntactic minimality criterion (Wilensky, 1983; Charniak, 1986; Kautz & Allen, 1986; Peng & Reggia, 1990). These approaches are neutral to the content of particular assumptions; they instead focus on structural considerations such as the lengths of alternative explanations' derivations or the number of abductive assumptions they require. Other approaches use different syntactic criteria, ranking explanations by their structural "coherence" (Thagard, 1989; Ng & Mooney, 1990).

Such methods suffer from two problems. First, comparing explanations according to syntactic properties alone is not sufficient to reliably rank their likelihood. For example, counting assumptions may be misleading because two commonplace assumptions may be more likely than one unusual one. Second, even in situations for which syntactic criteria do generate a reasonable comparative ranking of explanations' plausibility, a comparative ranking alone cannot give sufficient information. The problem is that unless the set of candidate explanations is complete, the best explanation among the current candidates may still be insufficient. In the everyday explanation task, incomplete sets of candidates are the norm—rather than selecting the best of a set of explanations presented to them, everyday explainers generate a stream of explanations, stopping when they are satisfied with an alternative.

Solving the aforementioned problems requires evaluation that can judge the reasonableness of explanations' content based on world knowledge. We propose a method that uses a highly constrained search process to search memory for information that supports or undermines belief in an explanation's assumptions, and that judges explanations in terms of that information.

Beyond validity: Validity considerations alone may be insufficient to choose between explanations. *All* the explanations in our list of alternatives for the break-in could be valid simultaneously, but the information they provide is different, so that their usefulness is different for different tasks. For example, the ATM manufacturer may wish to explain the break-in in order to design a better ATM, requiring an explanation of how the ATM's armor was defeated; John's parents may wish to understand the break-in in order to dissuade him from criminal acts, requiring an explanation of his motives. For some purposes, goals may even override validity considerations: A good explanation in a humorous context may be one that is farfetched or obviously false.

Both psychological research (Lalljee & Abelson, 1983; Snyder, Higgens, & Stucky, 1983) and philosophical investigations (e.g. (Mackie, 1965; Van Fraassen, 1980)) have pointed to the need for different explanations depending on overarching goals. However, abductive understanding systems take a goal-neutral view. Artificial Intelligence investigation of goal-based explanation evaluation has been concentrated in explanation-based learning (EBL) research (Mitchell, Keller, & Kedar-Cabelli, 1986; DeJong & Mooney, 1986), and focuses on the task of concept recognition (see (Keller, 1988) for an overview of some of this work). Rich models of usefulness have been developed for the recognition task (e.g., (Keller, 1987; Minton, 1988)), and other research has considered how the recognition task is motivated by other goals (Kedar-Cabelli, 1987), but other tasks have received little attention. Developing a rich model of goal-sensitive evaluation depends on first identifying the range of goals that explanation can serve, in order to develop criteria for the information needs arising from each one. We address this question with a theory of the goals that affect explanation and their information requirements.

3 A Model of Case-Based Explanation

Case-based explanation replaces the chaining used in traditional explanation construction with a memory-based process. The case-based explanation framework described here was developed in the SWALE system, a story understanding system that explains novel events when schema-based understanding fails (Kass, 1986; Leake & Owens, 1986; Schank, 1986; Kass & Leake, 1988; Schank & Leake, 1989); specific aspects of the framework have been examined in stand-alone systems descended from the modules of the original system (Kass, 1990; Leake, 1992).

The SWALE system's primary example is the story of the racehorse Swale. Swale was in peak form, and had recently won two major victories, when a shocking article appeared on the front page of the *New York Times:* Swale was dead. The death prompted huge interest in its explanation, both within and outside the racing community.

The problem of explaining Swale's death illustrates some of the problems of everyday explanation. First, it illustrates the dependence of explanations on context. To most of those who hear of the Swale example, explanation is motivated by a specific anomaly—that Swale died in his prime—and a good explanation must address that anomaly. For example, a relevant explanation would be "Swale was poisoned by the owner of the horse scheduled to compete with him." However, that explanation would not be relevant in another context: If the competing owner had thought the dose of poison would only be enough to make him too sick to compete in the next race, a good explanation would have to address the discrepancy, as in the explanation "Swale had an allergy to the type of drug used." Both of the explainers are explaining the same event—Swale's death—but the explanations they require are quite different.

The example also illustrates a second property of everyday explanation: the difficulty of controlling explanation search. A vast range of possible causes could be hypothesized for Swale's death, and little information was available to constrain the alternatives generated. However, human explainers had little trouble generating plausible hypotheses, and prior experiences were often cited as the reasons for the hypotheses proposed. For example, one veterinarian's immediate reaction to the news was "this sounds like an aneurysm. I've seen this sort of thing before" (Cress, 1984). A student hypothesized a heart attack after being reminded of the death of the runner Jim Fixx, who died when the exertion of recreational jogging overtaxed a hereditary heart defect. This explanation does not apply directly—Swale was unlikely to do recreational jogging—but minor adaptation of the explanation, substituting horse racing for jogging, produces the plausible explanation that the stress of running in a race overtaxed a hereditary heart defect. The process of retrieving and adapting the Jim Fixx explanation is used by SWALE to generate one of its candidate explanations for Swale's death.

Swale's death reminded another student of the death of another young superstar, Janis Joplin, and that student's reminding and adaptation process was also modeled by SWALE. Joplin was driven to recreational drug use by the stress of being a star, and died from accidentally taking an overdose of recreational drugs. Little of that explanation applies to Swale—racehorses do not take recreational drugs. However, a kernel of the explanation is applicable: that Swale might have died of a drug overdose. This explanation is not plausible in itself, because it is unsubstantiated, but it forms the starting point for additional explanation accounting for how the overdose could have been administered. Using its world knowledge, SWALE seeks standard explanations for racehorses receiving drugs, which suggests the hypothesis that Swale died of an overdose of performance-enhancing drugs.

3.1 The Case-Based explanation algorithm

The Swale examples illustrate the basic process of case-based explanation of anomalous events:

- **Anomaly detection:** Detect that current understanding is insufficient and that further explanation is needed.
- **Anomaly characterization:** Generate indices reflecting the situation to be explained and the focus of explanation.
- **Explanation retrieval:** Attempt to retrieve a candidate explanation from memory. If no relevant candidates are found, stop—explanation effort fails.[3]
- **Explanation evaluation:** Evaluate candidate's goodness.

  If the explanation is satisfactory, update beliefs to reflect explanation, store new explanation for future use, and stop.
- **Explanation adaptation:** If applicable repair strategies are available, revise explanation and return to explanation evaluation phase.

  If no applicable repair strategies are available, return to retrieval phase to attempt retrieval of another candidate.

[3] Indexing may use a wide range of strategies to increase the field of candidate explanations. See (Leake & Owens, 1986; Schank, 1986; Schank & Leake, 1989).

3.2 Issues in the Case-Based Approach

Storing explanations in memory requires a memory structure to represent explanations. That structure must include information to support direct application of the relevant explanations to anomalies, by providing a mapping from the anomaly to the explanation, and must have a sufficiently rich internal structure to allow the explanation to be adapted to new situations. SWALE's explanations are stored in memory as explanation patterns (XPs) (Schank, 1986). XPs are dependency networks tracing how a state or event can be inferred from a set of assumptions according to plausible inference rules; they represent a variablized version of the explanatory chain from a prior explanation.

Many issues arise in triggering explanation of anomalies and controlling the processing of XPs, such as how to detect anomalies, how to guide retrieval, how to evaluate candidate XPs, and how to control adaptation. Here we focus on techniques developed for retrieval and evaluation, highlighting their relationship to explanation generation and evaluation methods traditionally used in abductive systems.

Although the methods presented here are very different from standard approaches to abductive understanding, they could be integrated into traditional abductive systems in a straightforward way. A chaining-based explainer could take retrieved explanations as its starting point for additional backwards chaining, in that way avoiding some of the inference otherwise needed (Hammond, 1987). In addition, our evaluation process could be used to replace or augment standard minimality criteria for choosing between candidate explanations. However, the methods presented here also facilitate application of more sophisticated strategies that use the evaluation phase to direct incremental explanation construction. Kass (1990) describes methods for using problem descriptions generated by explanation evaluation to direct explanation adaptation, providing additional focus when modifying retrieved explanations.

3.3 The program ACCEPTER

The explanation retrieval and evaluation methods we describe are implemented in ACCEPTER, a story understanding system that detects anomalies, retrieves explanations for them, and evaluates the alternatives. ACCEPTER was initially developed as the main module of SWALE; the current version is a stand-alone system in which a human user replaces SWALE's adaptation component (that component has also been investigated in an independent system; see (Kass, 1990)). ACCEPTER processes about 20 simple (1–4 line) stories about anomalous events, primarily stories from the news, and evaluates the goodness of a total of about 30 candidate explanations from the perspectives of five types of overarching goals.

ACCEPTER's routine understanding is schema-based, which provides efficiency for routine situations but is inflexible in the face of novel situations. ACCEPTER overcomes that problem by monitoring its understanding to detect situations in which its schema-based understanding is insufficient—situations in which actual events conflict with prior beliefs and expectations—in order to decide when new explanations must be built. Although the stand-alone version of ACCEPTER is not a learning system, the version of ACCEPTER in SWALE stores new explanations in memory, making them available to facilitate future explanation of similar anomalies. A complete description of ACCEPTER's implementation is available in (Leake, 1992); here we concentrate on the underlying theory.

4 Focusing Explanation Retrieval

If a case-based explanation system can retrieve relevant prior explanations when it encounters a new situation, it can facilitate the explanation process by re-using the applicable portions of prior experiences. However, if the system cannot find the most relevant cases in its memory, adaptation of retrieved cases is unnecessarily costly, and efficiency advantages of case-based reasoning are reduced or nullified.

Effective explanation retrieval depends on developing an indexing vocabulary to link situations requiring explanations to explanations stored in memory. The goal of the vocabulary is to group explanations that are relevant to particular anomaly types, and to connect new anomalies to relevant classes of explanations. Accomplishing this goal depends on describing *what needs to be explained* in a new situation in the same way as *what is explained* by relevant explanations. If this can be done, explanations relevant to a new anomaly can be retrieved by assigning the vocabulary element for the anomaly and retrieving explanations stored under the same vocabulary element. Additional focus can be obtained by combining the vocabulary elements (which represent broad classes of anomalies) with descriptions of the specifics of the situation at hand, in order to retrieve explanations from similar situations. Additional flexibility can be achieved by defining a similarity metric for anomaly characterizations and using it to allow explanations with near-miss characterizations to be retrieved when no stored explanation precisely matches stored candidates. We begin by describing the vocabulary elements themselves, and later consider how to add specific information and how to retrieve explanations with near-miss characterizations.

4.1 A Vocabulary for Anomalies

Our anomaly vocabulary is designed to aid retrieval of explanations to resolve anomalies. As shown previously, resolving an anomaly depends not only on explaining the event in isolation but on accounting for the *conflict* between the event and prior beliefs or expectations. In order for an anomaly vocabulary to suggest explanations that are relevant to the conflict, the vocabulary elements must reflect the type of conflict involved. An important factor is the domain of knowledge that was involved in the reasoning that went astray. For example, if an event contradicts expectations about decision-making in planning, the explanation must focus on factors relevant to the understander's model of decision-making; if the event contradicts expectations about the way a plan will proceed after it has been selected, the anomaly characterization must describe the type of knowledge used to predict the course of the plan.

To reflect the domain of knowledge underlying failed expectations, an anomaly vocabulary must include a distinct category for each domain of knowledge that can form the basis of expectations. We have identified eight major categories of knowledge affecting expectations and defined anomaly categories for each one. The types of knowledge and related anomaly types are:

- Planning choice: We might have expected John to use his usual plan for getting money—borrowing from a relative—rather than doing the robbery. Here the anomaly is SURPRISING-PLAN-CHOICE; an explanation must account for the deviation from his expected plan selection process.

- Plan instantiation: We might have expected John to use a crowbar instead of a torch. Here the anomaly is SURPRISING-PROP-CHOICE; an explanation must account for why his action deviated from the expected one for choosing objects to use in plans.

- Plan and action execution: We might have seen John set out to rob another bank instead, making the change in target surprising. The category for failures of plans to proceed successfully is PLAN-EXECUTION-FAILURE. Failures can also occur when plans are successful contrary to expectations: If we previously expected the plan to be blocked (*e.g.*, by new security measures at the bank), and it succeeded anyway, the explanation would also have to account for that deviation, a BLOCKAGE-VIOLATION anomaly.

- Information transmission: We might have previously read that the robbery took place at a different bank, making the anomaly a problem of BAD-INFORMATION.

- Models of physical, biological or chemical processes: We might not have expected the blowtorch to be able to melt the ATM door quickly enough. The category for deviations from models of processes, such as unusually rapid melting, is PROCESS-EXECUTION-FAILURE.

- Models of device function: We might have expected the ATM alarm to alert the bank before the break-in could be completed. The failure to do so is an instance of DEVICE-FAILURE.

- Inductive generalizations about object features: We might think that all ATMs are run by other banks, in which case it is anomalous that the owner of the ATM is First National Bank. The category for such anomalies is UNUSUAL-OBJECT-FEATURE.

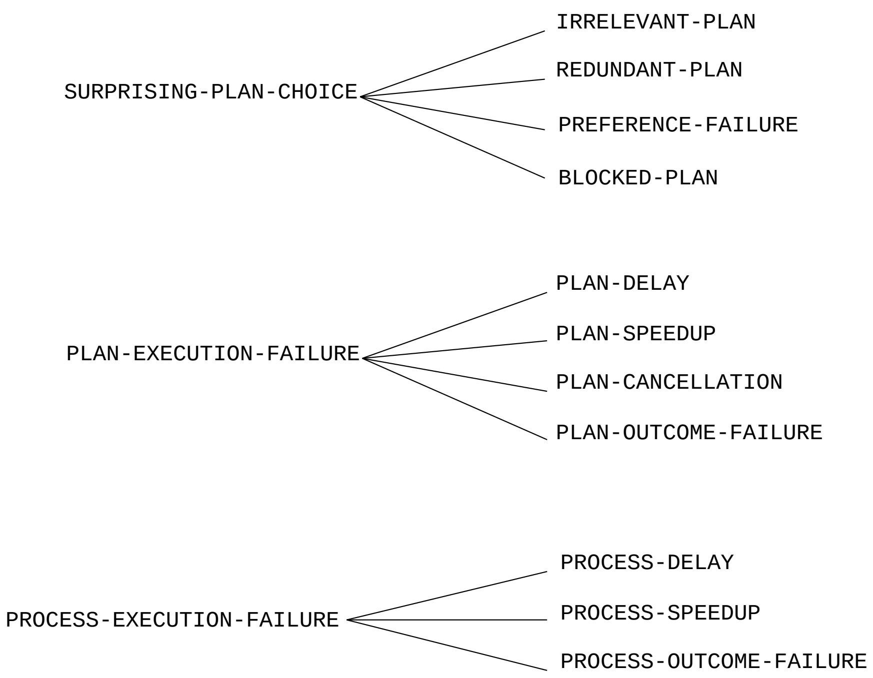

Figure 1: Subcategories for three anomaly categories.

- Generalizations about persistence of features (McDermott, 1982): Successful theft from an ATM would be surprising if we believed that the machine contained no money, and would not be refilled until the next day. The anomaly of money being in the ATM is STRANGE-FEATURE-CHANGE.

The nine preceding anomaly categories apply to a wide variety of everyday anomalies. Although they are not exhaustive, developing of these categories has convinced us that a complete account of everyday anomalies would require a fairly small number of additional categories.

Specifying the categories into subcategories: The preceding categories describe the domain of knowledge used in reasoning leading to a failed expectation. Consequently, they allow retrieval of explanations that focus on why reasoning in that domain of knowledge may go wrong. Additional guidance is provided by extending the vocabulary to include subcategories of the initial categories reflecting how expectations failed; examples of some of these subcategories are shown in figure 1.

For example, although the category PLAN-EXECUTION-FAILURE can guide explanation search towards explanations relevant to knowledge of planning, there are many different

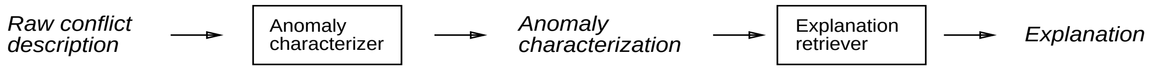

Figure 2: Focusing explanation retrieval.

ways that the execution of a plan can fail, and these ways correspond to explanations with different focuses. By generating subcategories corresponding to different types of failures, it is possible to concentrate attention on explanations relevant to the specific problem at hand. One such subcategory is PLAN-DELAY, which describes those instances of PLAN-EXECUTION-FAILURE in which plans conflict with expectations because their execution is slower than expected. The category PLAN-DELAY suggests standard explanations specific to delays, such as: plans are often delayed because of problems with competition for resources (a car trip may be delayed by heavy traffic), or because preconditions that are usually satisfied in advance needed to be established before the plan could begin (a car trip may take longer because of needing to fill up the gas tank before starting). Thus anomaly subcategories provide additional focus for explanation search. We consider these categories and subcategories the top levels of an abstraction hierarchy that could be specified further to provide a discrimination tree to direct search in a large library of explanations.

4.2 How the Vocabulary Guides Explanation Search

The anomaly vocabulary guides explanation retrieval in a two-step process, as sketched by figure 2. First, the anomaly is characterized to build up a description of its important features. Second, the characterization is used as an index into memory.

Using the vocabulary to guide anomaly characterization: When ACCEPTER detects an anomaly, an initial description of the anomaly—what conflicted with expectations, and a pointer to those expectations—is passed to its anomaly characterizer. Based on the expectation source and the type of conflict, the characterizer selects the associated anomaly category. Once the category has been selected, the characterization process selects features of the current situation to use as more specific indices.

In principle, it is impossible to identify all important features of an anomalous situation until it has been explained. However, knowledge of the anomaly category can suggest the features to consider, based on experience of the types of features that have proven important in the past. For example, if an understander is surprised by the plan that an actor chooses for its goal, experience suggests that features likely to be relevant to the explanation include the actor, the plan, and the way the plan deviated from expectations. For each anomaly

Slot	Filler
Assumed goal	Open ATM
Actor	John
Plan	Melt door
Conflict description	Melting door instead of prying

Table 1: Components of the anomaly characterization for a SURPRISING-PLAN-CHOICE anomaly.

category, knowledge of the features likely to be relevant is encoded in a knowledge structure that the characterizer instantiates to reflect specifics of the current situation that are likely to be relevant to its explanation.

Table 1 shows an example of the characterization for the SURPRISING-PLAN-CHOICE anomaly of John choosing to melt the ATM door rather than to pry it open. The components of this structure direct retrieval toward unusual plans that the actor (or similar actors) selected in similar circumstances (e.g., that another thief switched to torches because crowbars are hard to maneuver in cramped ATM enclosures). For more details, see (Leake, 1992).

Using the characterization to guide explanation retrieval: In ACCEPTER's explanation memory, explanations are organized by the anomaly vocabulary: the primary index for explanation retrieval is the vocabulary element for the anomaly it explains. Within the set of explanations grouped by a single element of the vocabulary, search is guided by the specifics of how the roles in their anomaly characterization structures are filled. ACCEPTER tries to retrieve the stored explanations with the most similar characterizations, with similarity of characterizations measured by similarity of slot-fillers in the anomaly characterizations, judged by their distance in the system's abstraction hierarchy of objects and events. Thus explanations in memory can be considered to be organized in an implicit net, with the possibility of a single characterization having multiple abstractions. This implicit net is searched breadth-first.

As an example of how the anomaly categories constrain explanation search, consider again the characterization in table 1, which describes the anomaly "John was expected to pry the ATM door open, but melted it instead." Based on this characterization, explanation search would first try to retrieve an explanation with exactly the same anomaly characterization, retrieving explanations for other instances of John choosing to melt a door instead of prying it (e.g., he may have done it before when he was trying to impress someone with his high-tech methods). If no explanation is found under that specific index, ACCEPTER tries to find explanations under the same anomaly type, but with generalizations of the fillers in the current instance. For example, John could be generalized to "thief" to look for other episodes of thieves melting rather than prying doors, which might index the explanation of the door of a safe that was sealed too tightly to insert a crowbar.

In order for case-based explanation to be able to explain flexibly, it must be able to search for near-miss explanations when no precisely matching explanations are found. In such cases, the characterization structure can suggest the significant features for a partial match: some of the slots can be forced to match, and others allowed to vary. For example, the structure in table 1 suggests that if no explanations can be found for why a particular actor chooses a particular surprising plan, we might look for why other actors choose the plan, even if they are not abstractions of the actor currently under consideration—for example, why a fireman would use a torch rather than a crowbar to force entry into a locked room.

Using the characterization to suggest abstract explanation strategies: No explainer can hope to retrieve an appropriate prior explanation for every new anomaly. However, the previously described anomaly vocabulary can also facilitate explanation from scratch by organizing abstract *explanation strategies* (Hammond, 1987) in addition to specific prior explanations.[4] Explanation strategies provide information about types of factors that are particularly likely to be relevant when explaining a particular type of anomaly; that information is then used to guide the search for situation-specific information needed to complete the explanation.

For example, a possible explanation for any instance of SURPRISING-PLAN-CHOICE is "the usual plan is blocked." This suggests an explanation strategy: Given an unexpected plan choice that cannot be accounted for in terms of similar previous episodes, look for factors that make the expected plan impractical. If such factors are found, and if the new plan is reasonable as a fallback, that explains the change in plan. By suggesting the type of factor to look for, the explanation strategy makes it possible to use world knowledge to constrain search even when no specifically applicable explanation is available. For example, if we wonder why John used a torch rather than a crowbar to break into the ATM, the explanation strategy suggests looking for reasons that a crowbar could not be used, suggesting explanations such as lack of maneuvering room in the ATM enclosure. Although explanation strategies provide less guidance than explanation patterns, they still narrow the search space.

In addition, just as the subcategories of our anomaly vocabulary group explanations with a specific focus, the subcategories can group more focused explanation strategies. For example, if what was anomalous about John breaking into the ATM was that it took longer than expected, the anomaly would be characterized as the PLAN-DELAY subcategory of PLAN-EXECUTION-FAILURE, for which possible explanation strategies include:

- **Link to task scale increase.** Increasing the scale of a task can make it more time

[4] Explanation strategies have not been implemented in ACCEPTER.

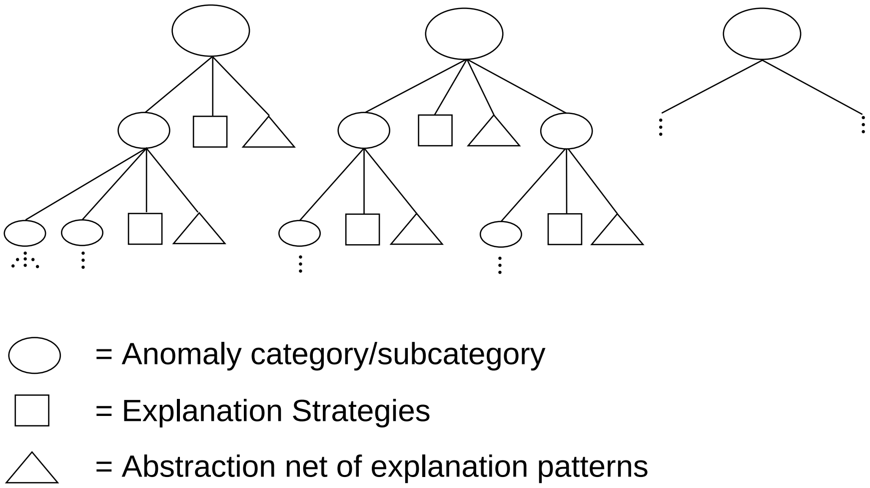

Figure 3: Organization of explanation strategies and explanation patterns under anomaly categories and subcategories.

consuming. For example, a delay in melting through an ATM wall may result from the machine having a thicker wall than expected.

- **Link to shared resource problem.** Another reason for delay is competition for the resources the plan needs, forcing a wait until other actors relinquish them. For example, a delay in John's breaking into the ATM might be caused by waiting to use tools shared with a burglar breaking into an adjacent machine.

- **Link to actor skills.** The delay may be caused by lack of skill: Perhaps John is out of practice.

Figure 3 sketches how explanatory information—explanation strategies and explanation patterns—is organized by the anomaly categories.

4.3 Judging the Vocabulary

The previous sections sketch a categorization scheme and its intended function; here we consider how to judge how well it fulfills that function. In order for the anomaly vocabulary to facilitate explanation, it must be able to guide search towards the most relevant prior explanations in memory—towards precisely matching explanations when possible and towards near-miss explanations when no precisely-matching explanation is available. The need to retrieve near-miss explanations makes the specific anomaly vocabulary crucial: Retrieving appropriate near-miss explanations is only possible if similarities in anomaly characterizations correspond to similarities in the explanations that apply to those characterizations.

Our characterizations are designed to capture important anomaly classes in everyday events. At this point, it is impossible to generate a representative set of real-world anomalies and explanations on which to test the categories' ability to associate new situations with appropriate stored explanations: Retrieval in both case-based reasoning systems and in people is based on idiosyncratic experience. However, we *can* support our vocabulary of anomaly types on other grounds.

Our argument is based on the existence of explanation strategies associated with the anomaly categories and subcategories. To establish that our categories actually group problems in such a way as to facilitate their solution, it suffices to show that we can describe sets of general explanation strategies, associated with elements in our vocabulary, that can be applied to *any* anomaly of that type, and that the strategies are disjoint from each other. The reason this suffices is that if the elements in the vocabulary correspond to sets of abstract strategies, and those sets are disjoint, the vocabulary partitions the set of anomalies according to abstract causal properties relevant to their explanations. Such a partition suggests that the vocabulary corresponds to differences that are likely to apply to a wide class of problems, rather than coincidentally working on the specific anomalies and explanations tested.

Because we have argued that explanations of anomalies must address both the underlying reasoning that failed and how it failed, and that information is represented in our anomaly subcategories, the subcategories become the main testing ground for of the vocabulary's generality: The subcategories should correspond to groups of explanation strategies. In fact, (Leake, 1992) shows that for each subcategory, there does exist a distinct set of relevant explanation strategies. (Some of the main anomaly categories also organize their own strategies, disjoint from the strategies for other main categories, and applicable to their subcategories as well.) The existence of these sets of explanation strategies suggests that our vocabulary captures distinctions needed to guide explanation search.

5 Choosing the Explanation to Accept

Once candidate explanations have been generated, a second fundamental problem arises: how to judge those candidates. Traditional methods judge likelihood by structural comparisons intended to reflect validity, but as shown previously, that approach fails to address important problems for the open-ended situations in everyday explanation. We begin the section by describing an alternative knowledge-based approach. We then argue that despite the importance of validity, the goodness of explanations also depends crucially on another factor that has been generally neglected in abductive understanding systems: whether the explanation provides the information that the explainer needs.

5.1 Judging the Validity of Explanations

Our approach to validity evaluation is based primarily on the content of explanations, rather than their structure. It focuses on the relationship between the explanation's antecedents and prior knowledge, judging how well the explanations reconcile the anomalous event with prior understanding. To evaluate the plausibility of the antecedents, it checks whether they are supported by prior system knowledge (specific prior beliefs, expectations, or patterns), or whether they conflict with that knowledge. When pattern conflicts are found, additional checks are used to find the underlying cause of the problem. The underlying cause is used to characterize the anomaly as specifically as possible, in order to facilitate search for explanations of the problem and for adaptation strategies to resolve it.

After verifying the reasonableness of the explanation's antecedents, the process traces the explanatory chain from its antecedents to the event being explained, checking the applicability of the explanation's rules and the reasonableness of its assumptions and intermediate beliefs, given the context of prior knowledge and the expectations built up from previously-processed beliefs in the explanation. This verification of the explanation's derivation is needed because the rules used in everyday explanations are unavoidably approximate and imperfect; ACCEPTER's explanatory chains reflect plausible reasoning.

In ACCEPTER's model, minimality criteria are used only to choose between explanations whose plausibility cannot be distinguished by system knowledge. In its emphasis on the content of explanations, rather than their form, our view is in the spirit of probabilistic approaches (Pearl, 1988; Charniak & Goldman, 1991). However, as the next section describes, our primary focus is different from that work: Our focus is on developing a theory of the types of factors an everyday explainer should consider when judging an explanation's likelihood.

5.1.1 The problem of controlling inference

Relating an explanation's assumptions to prior knowledge is complicated by a classic problem: how to determine the relevance of particular prior beliefs. It is well known that arbitrary amounts of inference may be needed to derive the relationship between new information and any given piece of background knowledge (Rieger, 1975). Identifying *all* relevant relationships between beliefs requires checking for matches and conflicts between all ramifications of new information and prior beliefs. To make verification practical in rich domains, ways are needed to achieve a reasonable level of verification without incurring this overwhelming cost.

When people have no specific knowledge about whether a fact is true, they estimate its likelihood by comparing the hypothesis to standard stereotypes. For example, (Kahneman, Slovic, & Tversky, 1982) shows that people use stereotypes to decide if it is reasonable for someone to have a given profession. If people in the profession normally fit a stereotype, people expect those who fit that stereotype to have the profession—no matter how uncommon the profession may be. This heuristic can cause errors, but is often useful, and can be applied with low cost since it relies on information likely to be accessible.

Our approach to controlling inference cost when deciding plausibility is pattern-based; we replace inference by comparing assumptions only with explicit prior beliefs, active expectations, and a limited set of standard patterns for events in the world. In particular, ACCEPTER's process for judging the reasonableness of assumptions is as follows:

- **Attempt to match assumption with schema-based expectations and prior beliefs.** If they match, no further checks are needed. Conflicts are reported as plausibility problems.
- **Attempt to match assumption with schema-based prohibitions.** If the assumption is explicitly prohibited by a currently active schema (e.g., an athlete breaking training), it is a plausibility problem.
- **Check whether features of assumed event or state match a predefined basic set of causal restrictions on slot-fillers.** If restrictions are not satisfied (e.g., a non-flammable object is burning), the conflict is a plausibility problem.
- **Compare assumption to standard patterns.** ACCEPTER has a library of stereotyped patterns describing standard expectations such as the normative types of actors involved in events (e.g., that surgery should be performed by surgeons), the types of actions particular actors favor and avoid (e.g., that an athlete in training avoids partying), and the features that predispose actors to fill certain roles in actions (e.g., that being high-strung contributes to a predisposition for heart attacks). Comparisons with patterns can identify plausibility problems in an explanation, if its assumptions contradict prior patterns, or can support the assumptions' plausibility.

After evaluating the plausibility of the assumption itself, ACCEPTER tries to relate it to memory through the system's routine understanding process, attempting to activate schemas to package it and recursively evaluating their plausibility.

As an example of pattern-based plausibility evaluation, consider ACCEPTER's judgement of the plausibility of explaining Swale's death by the explanation for Jim Fixx's death (recreational jogging overtaxed a hereditary heart defect). ACCEPTER judges recreational jogging implausible for Swale because ACCEPTER's knowledge includes the pattern that joggers are usually human (even though a non-human jogger is possible: we can imagine a monkey being taught to jog). However, ACCEPTER's patterns do provide support for two aspects of the explanation: The assumption that Swale died from a heart-attack is supported by the pattern that being high-strung is a predisposing feature for heart-attacks, because racehorses tend to be high-strung, and the assumption that Swale was involved in exertion is supported by exertion being a generalization of participating in horse racing, which is a pattern for race horses.

In the example of attempting to apply the Jim Fixx explanation, pattern-based checks provide two useful pieces of information. First, they show that the explanation is implausible as it stands. Second, they show that the central core of the explanation *is* plausible. This suggests that the explanation is worth pursuing but that the assumption that Swale was jogging should be revised. Simply deleting that assumption results in a plausible explanation—*exertion + heart defect causes fatal heart attack.* That explanation can also be supported more specifically by explicitly providing a link to background knowledge, resulting in *racing-induced exertion + heart defect causes fatal heart attack.*

ACCEPTER's patterns are organized hierarchically in a series of tables in memory. ACCEPTER judges plausibility by searching the tables for confirming or disconfirming information; it accepts or rejects inputs on the basis of this partial verification, avoiding costly inference. The potential drawback of this method is an obvious tradeoff between efficiency and completeness: Is it possible to define a family of patterns with sufficient coverage? The goal of the pattern-based approach is not to detect all plausibility problems or confirmations of new assumptions, but to achieve a reasonable tradeoff between the efficiency and coverage—to detect a reasonable proportion of anomalies that people notice in everyday situations. In (Leake, 1992) we survey the types of anomalies that this method can and cannot detect and substantiate its coverage by presenting a family of patterns that covers a wide range of everyday anomalies.

5.1.2 Accepting explanations based on plausibility

Based on the plausibility problems it finds, ACCEPTER does a simple scoring of the plausibility of available explanations. This scoring is based primarily on the believability of each explanation's weakest assumption or rule: An explanation is considered as strong as its weakest part. A coarse-grained scheme is used; the system's likelihood classes are:

1. **Confirmed** by prior beliefs or active expectations
2. **Supported** by patterns or predisposing circumstances, or by known actor goal
3. **Unsupported**, but without conflicts
4. **Conflicting** with patterns, beliefs, etc.

For classifying the likelihood of rules, only two categories are used:

1. **Confirmed** by the system's causal knowledge (the rule is known, and its restrictions are satisfied)

2. **Conflicting** with the system's causal knowledge (no matching system rule is found, or the restrictions are not satisfied)

Although ACCEPTER compares the plausibility of alternative explanations, its decision of which explanation to accept is not only comparative: it can recognize situations in which no explanation is sufficiently plausible, allowing it to request that other candidates be generated. It rejects explanations that include rules or beliefs conflicting with prior knowledge, even if no other alternatives are available.[5] Explanations that do not conflict with system knowledge are considered possible candidates for being accepted; ACCEPTER can rank a set of candidates by the strengths of their confirmations and problems. When two explanations have the same strength, ACCEPTER attempts to break the tie by simple structural minimality considerations. If no differences are found, ACCEPTER favors the explanation best supported by experience: It favors the explanation whose anomaly characterization most specifically matches the characterization of the current anomaly, assuming that the similarity between old and new situations makes it more likely to apply.

To illustrate, we consider the result of ACCEPTER's plausibility ranking for the following five explanations of Swale's death:

`FIXX-XP`: Swale, like Jim Fixx, died because recreational jogging overtaxed a hereditary heart defect.

`HEART-DEFECT+HORSE-RACING-HEART-ATTACK`: Swale died because running in a horse race overtaxed a hereditary heart defect.

`INBRED-HORSE+HORSE-RACING-HEART-ATTACK`: Swale died because inbreeding caused him to have a heart defect, which was overtaxed by running in a race.

`PERFORMANCE-DRUG-OVERDOSE-BY-TRAINER`: Swale died because his trainer administered a fatal dose of performance-enhancing drugs.

`POISONING-BY-OWNER-FOR-INSURANCE`: Swale's owner poisoned him in order to collect property insurance.

By ACCEPTER's evaluation, `INBRED-HORSE+HORSE-RACING-HEART-ATTACK` is the most acceptable explanation. Its belief-support chain is reasonable, and both its antecedents are accounted for by prior information: ACCEPTER's background knowledge includes the pattern that racehorses tend to suffer from inbreeding, and Swale was already known to be involved in racing. The second most acceptable is the explanation `HEART-DEFECT+HORSE-RACING-HEART-ATTA` It is possible, but does not substantiate why Swale would have had a heart defect, and ACCEPTER has no previous pattern or expectation supporting the belief that Swale had a heart defect.

[5]Note that adaptation may be able to generate an acceptable revised explanation that includes the conflicting rules or beliefs, provided that the new explanation resolves the conflicts.

ACCEPTER detects plausibility problems in each of the other candidate explanations. PERFORMANCE-DRUG-OVERDOSE-BY-TRAINER is supported by ACCEPTER's belief that trainers often administer performance-enhancing drugs, but conflicts with ACCEPTER's knowledge that dosage of performance enhancing drugs is carefully regulated; it is ranked third. The explanation FIXX-XP is unlikely because the hypothesis that Swale was jogging conflicts with the pattern that joggers are human (it also fails to substantiate the heart defect). POISONING-BY-OWNER-FOR-INSURANCE conflicts with the pattern that Swale's owners are law-abiding, although it provides a reasonable motivation (getting money) if they are not. The explanation also depends on Swale being insured, and the program has no reason to believe that he was.

Note that ACCEPTER's plausibility rankings reflect primarily how well each explanation links the anomalous event to factors that are already expected or believed. This shows how well the explanations reconcile the anomalous event with prior understanding, but it may not parallel the explanations' objective probabilities. For example, the explanation INBRED-HORSE+HORSE-RACING-HEART-ATTACK has a lower probability than the explanation HEART-DEFECT+HORSE-RACING-HEART-ATTACK, because the former explanation includes the additional assumption that the heart defect was caused by inbreeding. However, it receives a higher ranking because it better connects the death to ACCEPTER's prior beliefs: it shows how the death is supported by the belief that inbreeding is common in racehorses.

Although we have argued against purely comparative evaluation schemes, it is clear that some comparative validity considerations must be addressed to give a complete account of explanation evaluation. ACCEPTER's method uses minimality considerations to break ties between explanations that are equally plausible given system knowledge, but the system should include other comparative criteria as well. Our scheme does not consider the strength of competing candidates when deciding if an explanation is sufficiently plausible, but the certainty with which people accept a given explanation depends not only on the plausibility of the explanation in isolation but on the plausibility of competing alternatives and on reasoning about the completeness of the set of alternatives considered (Josephson, 1991).

5.2 Beyond Validity: Judging Explanations' Usefulness

5.2.1 Judging Relevance to Anomalies

Regardless of an explanation's validity, it will serve no purpose unless it provides useful information. Consequently, explanation evaluation must reflect explainer needs.

Anomalies show that an understander's world model is flawed. In order to resolve anomalies, explanations must provide two types of information. First, they must provide an account of why the observed event makes sense, which is the standard requirement for explanations in abductive understanding systems. Second, in order to avoid similarly flawed reasoning in the future, they must identify the belief problems leading to flawed beliefs or expectations. For example, if the ATM robbery is anomalous because John was expected to use a crowbar, the explanation must show not only why John used a torch, but why he used it *contrary to the understander's expectation.*

Flawed beliefs and expectations result from flaws in the reasoning used in generating those beliefs and expectations—from the inference rules used, from the beliefs or assumptions to which they were applied, or from failure to consider the influence of relevant information. ACCEPTER's evaluation of relevance to an anomaly tests two of these factors: whether the expectation was based on false beliefs or whether the explanation failed to take relevant factors (factors that could have changed expectations) into account.

In principle, identifying the sources of an understander's failed explanations requires maintaining a justification network accounting for system beliefs (O'Rorke, 1983; Collins & Birnbaum, 1988). ACCEPTER's ability to identify the relevant beliefs is limited because it does not maintain reasoning traces for all of its expectations, but it does maintain a record of how each scheme-based expectation was generated. When a schema-based expectation fails, it compares beliefs in the explanation to standard conditions in the schemas. If the explanation shows unusual factors leading to an event, those factors are a possible reason that the schema failed to apply: The presence of the unusual feature is what makes the new reasoning chain take precedence. For example, Swale's death was anomalous because his death occurred prematurely compared to expectations from the standard schema for a racehorse's life. The explanation *exertion + heart defect causes fatal heart attack* explains why the schema did not apply, by showing that the death was caused by a condition that is unusual and consequently was not accounted for when forming the schema's expectations: the hereditary heart defect. Consequently, the explanation accounts for the anomaly and allows the system's world model to be repaired.

5.2.2 The Range of Goals Affecting Explanation

The previous section illustrates that explanations for anomalies must be sufficient to guide repair of the understander's world model. However, when an understander has goals beyond simple understanding, anomalies affect more than just the understander's world model. Whenever its world model needs to be revised, current goals and plans may need to be revised as well. For example, consider again the problem of explaining John's ATM break-in. Candidate explanations included:

- John needed money to pay back a loan shark.
- The bank's security camera had been removed for repairs.
- New high-temperature torches can quickly melt the alloys used in ATMs.

There are important differences in the usefulness of these explanations to different explainers, depending on their intentions for using the explanations to further their goals. For example, John's parents might want to find a way to absolve John of some of the responsibility for the robbery, by blaming circumstances. "John needed money to pay back a loan shark" provides information useful for that purpose—that John is under duress—while the other explanations do not. A bank officer attempting to avoid future break-ins might want to find how to block similar chains of events contributing to a robbery, leading to the desire to find an explanation that would identify factors that contributed to the robbery and that it would be possible for the bank to alter. For this purpose, an explanation such as "The bank's security camera had been removed for repairs" is a good explanation because it identifies a cause that the bank can remedy to discourage future robberies. Choosing the better of these two explanations is impossible without knowing the purpose of the explanation effort.

Information needs can be triggered by goals in the world and by *knowledge goals* (Ram, 1990) to change an understander's internal knowledge state. Information needs determine the types of antecedents and rules that must be included in a good explanation, and a goal-driven explainer must find explanations including these factors. Consequently, building an explanation that links the explained event to useful factors becomes the *explanation purpose* that drives the explanation effort.

We have identified ten explanation purposes triggered by goals that arise when explaining anomalous events.[6] The first goal is simply to establish that it is reasonable for the anomalous event to have occurred, in order to verify that it should be believed. The explainer's desire to satisfy this goal prompts the explanation purpose reflected by traditional abductive understanding models:

1. Connect situation to expected/believed conditions

For example, an insurer might be skeptical of an ATM robbery claim and want to confirm that the break-in really took place before paying the claim. This would lead to the explanation purpose of showing that robbery's occurrence could be supported by other evidence (e.g., by prior knowledge of John's intentions).

The next purpose reflects the goal of repairing the flawed knowledge that led to a faulty prediction or belief (e.g., to avoid making similar flawed predictions in the future). In order to repair flawed knowledge, an explainer must find which factors leading to the anomalous the situation were previously overlooked or misjudged, prompting the explanation purpose:

2. Connect situation to previously unexpected conditions

For example, a parole officer who trusted John might try to form a more accurate picture of John's character by finding what made him less trustworthy than expected. The parole

[6]Although we concentrate on explanation of anomalous events, many of these purposes can apply to explanations of nonanomalous situations as well.

officer might also wish to learn how to anticipate crimes by parolees in the future, leading to the purpose:

3. Connect situation to factors from which it can be predicted

One reason to learn how to predict an event is as a step towards its prevention. In order to prevent an event, a system can predict it when it is imminent and then take steps to block its occurrence. The goal to prevent an action gives rise to the explanation purpose:

4. Connect situation to factors that a given actor can control in order to block its occurrence.

For the ATM manufacturer, an explanation for this purpose might focus on the aspects of the ATM that physically enabled the break-in to take place, such as the alloy used; for a bank owner, the explanation might include the absence of a security camera.

If the anomalous state being explained is an undesirable state that is still in effect, the explainer might have the goal of changing the current state. Changing the current state can be facilitated by knowing about continuing factors leading to the state that a given actor can repair, leading to the explanation purpose:

5. Connect undesirable state to possible repair points.

For example, if the anomaly was that the ATM alarm was broken, the manufacturer might try to connect the alarm's state to a factor such as a disconnected wire.

An additional goal is to use the anomalous situation as a sample case for testing a theory, or to make theory-specific predictions. For example, a metallurgist might be interested in establishing whether the quick melting of the ATM door could be accounted for by a new theory of the properties of alloys. This goal leads to the explanation purpose:

6. Connect situation to factors within a given theory.

The types of goals that prompt the remaining purposes are clear, so we list those purposes without additional elaboration:

7. Connect situation to factors that suggest praise or blame for an actor.

8. Connect an action to the actor's motivations.

9. Connect situation to factors that discriminate between alternative responses.

10. Connect situation to types of factors that, when their influence is communicated to another agent, will cause a desired change in that agent's beliefs.

5.2.3 Evaluation for the Explanation Purposes

Each explanation purpose requires that an explanation include different types of links and beliefs, which in ACCEPTER are described in terms of basic *evaluation dimensions.*[7] ACCEPTER has individual heuristics for evaluating along each of ten evaluation dimensions, and it combines the heuristics as needed to build tests reflecting the multiple dimensions of an explanation that may be important. Thus in our model, goal-based evaluation decisions are determined by the following chain:

Goals and plans	⇒	Information needs	⇒	Explanation purposes	⇒	Evaluation criteria	⇒	Evaluation decision

As an illustration of the evaluation criteria for an explanation purpose, consider the explanation purpose *connect situation to events from which it can be predicted.* This purpose might arise from the goal for an explainer to take appropriate action before the next occurrence of a similar situation (e.g., for John's parents to dissuade him before he does his next robbery).

In order for an explanation to be useful for prediction, a subset of its antecedents (assumptions and confirmed premises) must satisfy the following evaluation criteria:

1. When those assertions hold, the anomalous event's occurrence is sufficiently likely to be a reasonable prediction. This tests the evaluation dimension *predictive power.*

2. The system is likely to be aware of the future occurrence of those factors. This tests the evaluation dimension *knowability.*

3. Those factors happen long enough in advance of the anomalous event for it to still be useful to form the prediction when they occur. This tests the evaluation dimension *timeliness.*

4. Some of the factors distinguish the current situation from the situation in which prior expectations would have held (this allows the system to refine its predictions by distinguishing between circumstances in which it should form the prior prediction, and in which it should form the new one). This tests the evaluation dimension *distinctiveness.*

To complete the evaluation process, heuristics are needed for judging along each of the evaluation dimensions (e.g., for judging how likely the system is to be aware of future occurrence of similar factors). Evaluation dimensions and heuristics for judging them are discussed further in (Leake, 1992). Each of the other purposes is associated with its own requirements, described in terms of these and other evaluation dimensions.

[7] Although different evaluation criteria are needed for each explanation purpose, a particular explanation will sometimes be sufficient for multiple purposes.

ACCEPTER implements requirements for four of the previous purposes: predicting an outcome, repairing an undesirable state, finding how to control recurrence of the surprising state or event in the future, and finding whom to blame for an outcome. When ACCEPTER evaluates an explanation for a given purpose, it notes the useful information that is present and summarizes any missing information, in order to guide adaptation of the explanation to include the proper types of factors.

6 Integrating Explanation Construction and Evaluation

For clarity, our separate discussion of explanation generation and evaluation reflects the custom of treating explanation construction and evaluation as two successive processing steps: first a complete set of candidate explanations is generated and then explanations in the set are compared to select the best alternative. However, this separation is poorly suited to the everyday explanation process. For any real-world event, arbitrarily large numbers of explanations can be built, making it impossible to generate all candidates. Even constructing a small set of candidates may be prohibitively expensive: A detective explaining a break-in may need to conduct days of questioning to identify a few suspects for a crime. Consequently, in real-world explanation it is impractical to blindly build all possible explanations and then pick the best one.

A natural alternative is to use intermediate estimates of explanations' goodness to focus explanation effort (DeKleer & Williams, 1989; Ng & Mooney, 1990; Leake, 1992). In the case-based explanation model, evaluation identifies specific problems within current explanations—both plausibility problems and problems of insufficient information—in order to suggest whether additional explanations need to be generated, which of the current candidates it is worthwhile to pursue, and how to pursue them through adaptation.

For example, promising explanations with implausible assumptions can be supported either by recursively explaining the implausible assumptions and adding the new explanatory chain onto the explanation, or by modifying the explanation to replace the problematic portions with more reasonable ones (e.g., replacing the assumption that Swale was doing recreational jogging in the explanation of Jim Fixx's death with the more plausible one that Swale was running in a race). Likewise, if the information provided by an explanation is plausible but insufficient for current goals, adaptation can be focused on providing the needed information. Kass (1990) describes how a set of adaptation strategies can be organized by an evaluator's problem descriptions to expedite repair of particular types of problems in explanations. Integrating evaluation into explanation construction avoids expending effort on unpromising alternatives, and may even make it possible to stop explanation generation after generating a single candidate.

7 Conclusions

In rich domains, explainers must search through a vast space of potential explanations. As is well known, performing this search by standard chaining techniques can be extremely expensive for abductive reasoning systems, but human explainers have comparatively little difficulty controlling the search for candidate explanations. The preceding sections present an alternative theory of explanation construction and evaluation inspired by observations of human explanation. In that theory, explanation is guided by prior experience and current goals. We argue that this method offers two-fold advantages over standard methods: increased efficiency of explanation construction and increased reliability in generating and selecting good explanations.

Our basic framework is case-based explanation. Motivations for the case-based approach include reducing explanation construction cost, by allowing an explanation system to re-use prior effort from similar situations; improving the quality of results, by guiding search towards explanations supported by experience; and generating more useful explanations, by allowing an explainer to concentrate on candidate explanations likely to provide the information it needs. Realizing the desired benefits of case-based explanation depends on the ability to retrieve appropriate stored explanations, and effective retrieval requires a theory of how explanation memory is organized—which in turn depends on a theory of everyday anomaly and explanation. The previous sections sketch such a theory and use it to develop an indexing vocabulary and characterization scheme for organizing explanations of anomalies. Because similarities in its characterizations correspond to similarities in explanations, the vocabulary can suggest relevant near-miss explanations when no precise matches are available, making the process flexible enough to guide explanation search even in novel situations.

In addition to using experience to suggest candidate explanations, our model uses it to judge their plausibility. Evaluation of explanations is based on the reasonableness of their assumptions and derivations, in light of prior knowledge, rather than on the purely structural grounds used by most abductive understanders. Our pattern-based method can efficiently decide whether a given candidate explanation is sufficiently plausible, rather than only producing a comparative ranking as done by structural methods; this makes it possible for an explainer to decide whether to accept an explanation without having to generate an exhaustive set of candidates. In addition, the information it provides gives precise guidance for adaptation of flawed candidates, further streamlining explanation generation.

Even if a plausible explanation is generated, however, it may not be sufficient. Explanation is often done in service of goals beyond routine understanding, requiring that good explanations provide the specific information needed to achieve those goals. Consequently, any evaluation scheme for explanations must be based on a theory of the goals that drive explanation. Our model includes a theory of the standard purposes guiding explanation and the requirements they impose on good explanations.

Thus in our model, explanation is driven by explainer knowledge and needs for information. By combining experience with a theory of those needs and how they can be satisfied, case-based explanation provides guidance for explanation construction that makes it practical to generate good explanations in complex everyday domains.

References

Bareiss, R. (Ed.)., Bareiss (1991). *Proceedings of the Case-Based Reasoning Workshop*, Palo Alto. DARPA, Morgan Kaufmann, Inc.

Charniak, E. (1978). On the use of framed knowledge in language comprehension. *Artificial Intelligence*, *11*(3), 225–265.

Charniak, E. (1986). A neat theory of marker passing. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pp. 584–588 Philadelphia, PA. AAAI.

Charniak, E. & Goldman, R. (1991). A probabilistic model of plan recognition. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pp. 160–165 Anaheim, CA. AAAI.

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

Cress, D. (1984). Clot suspected in swale's death. *The Washington Post*, E1. June 19.

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

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

DeKleer, J. & Williams, B. (1989). Diganosis with behavioral modes. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 1324–1330 Detroit, MI. IJCAI.

Hammond, K. (1987). Learning and reusing explanations. In *Proceedings of the Fourth International Workshop on Machine Learning*, pp. 141–147 Irvine, CA. Machine Learning, Morgan Kaufmann.

Hammond, K. (Ed.). (1989). *Proceedings of the Case-Based Reasoning Workshop.* Morgan Kaufmann, Inc., San Mateo.

Harman, G. (1965). The inference to the best explanation. *Philosophical Review*, *74*, 88–95.

Hobbs, J., Stickel, M., Appelt, D., & Martin, P. (1990). Interpretation as abduction. Tech. rep. 499, SRI International.

Josephson, J. (1991). Abduction: conceptual analysis of a fundamental pattern of inference. Technical Research Report 91-JJ-DRAFT, Laboratory for Artificial Intelligence Research, The Ohio State University.

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

Kass, A. (1986). Modifying explanations to understand stories. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* Amherst, MA. Cognitive Science Society.

Kass, A. (1990). *Developing Creative Hypotheses by Adapating Explanations.* Ph.D. thesis, Yale University. Northwestern University Institute for the Learning Sciences, Technical Report 6.

Kass, A. & Leake, D. (1988). Case-based reasoning applied to constructing explanations. In Kolodner, J. (Ed.), *Proceedings of the Case-Based Reasoning Workshop*, pp. 190–208 Palo Alto. DARPA, Morgan Kaufmann, Inc.

Kautz, H. & Allen, J. (1986). Generalized plan recognition. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pp. 32–37 Philadelphia, PA. AAAI.

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

Keller, R. (1987). *The Role of Explicit Contextual Knowledge in Learning Concepts to Improve Performance.* Ph.D. thesis, Rutgers University. Computer Science Department Technical Report ML-TR-7.

Keller, R. (1988). Defining operationality for explanation-based learning. *Artificial Intelligence*, *35*(2), 227–241.

Kolodner, J. (Ed.). (1988). *Proceedings of the Case-Based Reasoning Workshop.* Morgan Kaufmann, Inc., Palo Alto.

Krulwich, B., Birnbaum, L., & Collins, G. (1990). Goal-directed diagnosis of expectation failures. In O'Rorke, P. (Ed.), *Working Notes of the 1990 Spring Symposium on Automated Abduction*, pp. 116–119. AAAI. Technical Report 90-32, Department of Information and Computer Science, University of California, Irvine.

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

Leake, D. (1988). Evaluating explanations. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pp. 251–255 Minneapolis, MN. AAAI, Morgan Kaufmann Publishers, Inc.

Leake, D. (1990). Task-based criteria for judging explanations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 325–332 Cambridge, MA. Cognitive Science Society.

Leake, D. (1992). *Evaluating Explanations: A Content Theory.* Lawrence Erlbaum Associates, Hillsdale, NJ.

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

Mackie, J. (1965). Causes and conditions. *American Philosophical Quarterly*, *2*(4), 245–264.

McDermott, D. (1982). A temporal logic for reasoning about processes and plans. *Cognitive Science*, *6*, 101–155.

Minsky, M. (1975). A framework for representing knowledge. In Winston, P. (Ed.), *The Psychology of Computer Vision*, chap. 6, pp. 211–277. McGraw-Hill, New York.

Minton, S. (1988). *Learning Search Control Knowledge: An Explanation-Based Approach.* Kluwer Academic Publishers, Boston.

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.* Morgan Kaufmann Publishers, Inc., San Mateo.

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

Norvig, P. (1989). Marker passing as a weak method for text inferencing. *Cognitive Science*, *13*(4), 569–620.

O'Rorke, P. (1983). Reasons for beliefs in understanding: applications of non-monotonic dependencies to story processing. In *Proceedings of the National Conference on Artificial Intelligence* Washington, DC.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo.

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

Ram, A. (1990). Goal-based explanation. In O'Rorke, P. (Ed.), *Working Notes of the 1990 Spring Symposium on Automated Abduction*, pp. 26–29. AAAI. Technical Report 90-32, Department of Information and Computer Science, University of California, Irvine.

Ram, A. & Leake, D. (1991). Evaluation of explanatory hypotheses. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pp. 867–871 Chicago, IL. Cognitive Science Society.

Read, S. & Cesa, I. (1991). This reminds me of the time when ...: expectation failures in reminding and explanation. *Journal of Experimental Social Psychology*, *27*, 1–25.

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

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

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

Schank, R. & Leake, D. (1989). Creativity and learning in a case-based explainer. *Artificial Intelligence*, *40*(1-3), 353–385. Also in Carbonell, J., editor, *Machine Learning: Paradigms and Methods*, MIT Press, Cambridge, MA, 1990.

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

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

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

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