Abduction, Experience, and Goals: A Model of Everyday Abductive Explanation*

David B. Leake
Computer Science Department
Lindley Hall 215
Indiana University
Bloomington, IN 47405
812–855–9756
leake@cs.indiana.edu
February 5, 1995

Accepted by *The Journal of Experimental and Theoretical Artificial Intelligence*
Running head: Abduction, Experience, and Goals

*This research was supported in part by the National Science Foundation under Grant No. IRI-9409348.

Abduction, Experience, and Goals: A Model of Everyday Abductive Explanation

Abstract

Many abductive understanding systems generate explanations by a backwards chaining process that is neutral both to the explainer's previous experience in similar situations and to why the explainer is attempting to explain. This article examines the relationship of such models to an approach that uses case-based reasoning to generate explanations. In this case-based model, the generation of abductive explanations is focused by prior experience and by goal-based criteria reflecting current information needs. The article analyzes the commitments and contributions of this case-based model as applied to the task of building good explanations of anomalous events in everyday understanding. The article identifies six central issues for abductive explanation, compares how these issues are addressed in traditional and case-based explanation models, and discusses benefits of the case-based approach for facilitating generation of plausible and useful explanations in domains that are complex and imperfectly understood.

1 Introduction

Abductive inference is the pattern of reasoning involved in forming and accepting explanatory hypotheses (Peirce, 1948). In general, many competing hypotheses may be generated for any phenomenon, requiring the explainer to choose between competing alternatives; as a result, abduction is often characterized as "inference to the best explanation" (Harman, 1965). Modeling the abductive explanation process requires addressing fundamental questions of what constitutes an explanation, how candidate explanations are generated, and what constitutes the "best" explanation.

Many models of abduction address these questions as follows. They view abductive explanations as deductive proofs based on correct domain theories, but whose premises include additional "abductive assumptions;" They model explanation generation as a backwards chaining process that generates each new explanation by starting from scratch, without being influenced by prior experience; and they model the selection of the "best" explanation as being based solely on likelihood or plausibility, rather than reflecting the changing needs for information that motivate explanation (Charniak & Goldman, 1991; Charniak & Shomony, 1994; Hobbs, Stickel, Appelt, & Martin, 1993; Josephson & Josephson, 1994; Kautz & Allen, 1986; Konolige, 1990; Levesque, 1989; O'Rorke, 1994; Poole, 1989; Zadrozny, 1994). Although such approaches have proven useful in a number of contexts, problems arise when trying to apply these methods to the rich domain of everyday abductive explanation. Modeling everyday abductive reasoning requires confronting the ramifications of imperfect knowledge, limited reasoning resources, and pragmatic motivations for explaining.

An alternative model explanation generation relying on case-based reasoning was developed largely to address the problems of everyday abductive explanation (Kass, 1986; Leake, 1992; Leake & Owens, 1986; Schank, 1986; Schank, Riesbeck, & Kass, 1994). The case-based model generates new explanations by retrieving explanations of relevant prior episodes and adapting them to fit the new situation in light of the explainer's needs for information. In this model, the prior experiences of the explainer are fundamental to focusing search for candidate explanations, and the motivations for explaining are reflected in both the explanation generation and selection processes.

Detailed descriptions of the mechanisms involved in the case-based explanation generation process are available elsewhere (Kass, 1990, 1992; Leake, 1992; Schank et al., 1994). Consequently, this article will not discuss specifics of those mechanisms. Instead, its goal is to analyze the commitments and contributions of the case-based approach to explanation generation. It will accomplish this goal by identifying a set of fundamental issues affecting abductive explanation and using those issues as dimensions along which to compare different approaches to abductive explanation.[1] In these comparisons, the article's primary focus will be on comparing the case-based model to other models of abductive understanding.

The article identifies six central issues in abductive reasoning to use as points of comparison for different approaches: whether explanations are viewed as deductive proofs or as plausible reasoning, when to decide to explain a situation, what to explain about a given situation, how to generate explanations, how to determine the "best" explanation, and the appropriate level of integration between the processes that generate candidate explanations and that select the "best" candidate. The comparison illuminates the need of an everyday explainer to deal with imperfect domain theories, to focus explanation generation according to goal-based information needs, and to closely couple pragmatic considerations with explanation generation. This discussion illuminates

[1]The paper Leake (1993) contains an initial sketch of a number of these points of comparision.

the commitments of case-based explanation and its benefits as a way to facilitate explanation in the complex and imperfectly understood domains of everyday explanation.

2 Case-based explanation generation

To provide a background for the following comparison, we begin with a brief overview of the application of case-based reasoning to abductive explanation. Case-based reasoning solves new problems by re-applying the lessons learned from specific prior reasoning episodes (Kolodner, 1993; Riesbeck & Schank, 1989). A functional motivation for case-based reasoning is the principle that in a regular world, similar problems have similar solutions. When this principle holds, starting from similar previous solutions can be more effective than reasoning from scratch.

In the context of explanation generation, this means generating new explanations by retrieving prior explanations for similar situations and adapting those explanations to fit the new circumstances. Case-based explanation for understanding has been investigated in an ongoing series of projects beginning with the SWALE system (Kass, 1986; Leake & Owens, 1986; Schank & Leake, 1989; Schank et al., 1994) and continuing with SWALE's descendents ABE (Kass, 1990, 1992, 1994) and ACCEPTER (Leake, 1992, 1994a). The approach is now being extended to explanation in the context of diagnosis of device failures (Leake, 1994b; Sooriamurthi & Leake, 1994).

In SWALE, ABE, and ACCEPTER, abductive explanation is used to explain anomalous events in news stories. The domain involves incidents of death, defects, and destruction; specific examples processed include the explosion of the space shuttle Challenger, the accidental shootdown of an Iranian airliner by the American warship Vincennes, a car recall, and the premature death of Swale, a star racehorse who was in peak form when he suddenly died. The aim is to generate plausible explanations in a complex domain despite imperfect domain knowledge and incomplete information.

The example of explaining Swale's death illustrates the task and the difficulties involved in explaining everyday events. Because the information provided by initial news stories about Swale's death was simply that Swale was found dead soon after winning an important race, almost any cause of death was a potential candidate. However, human explainers appeared to have little difficulty generating a few particularly plausible hypotheses. For example, after the death the *New York Times* published an article discussing a number of common hypotheses about Swale's death.

People trying to explain the death often attributed their hypotheses to being reminded of prior episodes. For example, one person was reminded of the death of the runner Jim Fixx, who died when the exertion of recreational jogging overtaxed a hereditary heart defect. The explanation for Fixx's death does not apply directly to Swale—Swale was unlikely to do recreational jogging. However, that explanation can easily be adapted to the Swale episode by substituting horse racing for jogging. The result is a plausible explanation: that the stress of running in a race overtaxed a hereditary heart defect. This example and similar informally-collected accounts helped to suggest that explanation generation could be facilitated by applying case-based reasoning. Later psychological experiments have supported the psychological validity of this reminding-based explanation process and the tendency of people to favor explanations that are based on prior explanations of similar episodes (Read & Cesa, 1991).

2.1 The case-based explanation algorithm

The algorithm used in the SWALE system and its descendents can be summarized as follows:

- **Problem characterization:** Generate a description of what must be explained, i.e., the information that a good explanation must provide.
- **Explanation retrieval:** Use the results of the problem characterization step as an index for retrieving relevant explanations of prior episodes from memory.
- **Explanation evaluation:** Evaluate the retrieved explanations' plausibility and usefulness. Generate problem characterizations for any problems that are found.
- **Explanation adaptation:** If problems were found, use the evaluator's problem characterization to select adaptation strategies for modifying the explanation to repair the problems. Apply the strategies and return to the explanation evaluation phase to evaluate the new explanation.

The viability of this process depends on having effective methods to perform each of these steps. Substantiation of the methods used is beyond the scope of this article but is provided elsewhere: problem characterization and retrieval issues are addressed in (Leake, 1991b, 1992, 1994a), explanation evaluation issues are addressed in (Leake, 1991a, 1992, 1994a), and adaptation issues are addressed in (Kass, 1990, 1992, 1994).[2]

3 Six fundamental issues for abductive reasoning

Comparing models of abductive reasoning depends on first identifying fundamental issues to serve as points for comparison. The discussion in this article is centered on six major issues that arise in the many models of abduction that treat explanations as reasoning chains deriving or supporting belief in a state or event to be explained (Charniak, 1986; Charniak & Goldman, 1991; Charniak & Shomony, 1994; Hobbs et al., 1993; Josephson & Josephson, 1994; Kautz & Allen, 1986; Konolige, 1990; Leake, 1992; Levesque, 1989; O'Rorke, 1994; Mooney, 1990; Poole, 1989; Wilensky, 1983).[3]

The first issue we will address is the *nature of explanatory reasoning chains.* In many models, these chains are viewed as deductive proofs that may depend on additional abductive assumptions; the case-based model instead maintains that models of everyday abductive reasoning must explicitly treat explanatory chains as reflecting plausible reasoning. We will show that this has important ramifications for the explanation generation process.

[2]Being able to apply case-based reasoning requires that the explanation system have access to a library of candidate explanations to use as its starting point. The cited research on case-based explanation construction endows the explanation system with that initial library and augments the library with the new explanations that are generated as those explanations are applied. In general, a case-based explanation system's initial explanation library could be provided by external sources (e.g., by reading about explained episodes) or built up by chaining methods (Koton, 1988).

[3]An alternative view of the nature of explanations is presented by set-covering models (e.g., Peng & Reggia, 1990), in which explanations are sets of factors that provide a covering for a set of findings, according to pre-defined associational links, rather than derivational chains built up from more primitive rules. Discussion of that approach is beyond the scope of this article.

The second issue is *when to explain,* that is, what prompts explanation. In many models of explanation, the decision of when to explain is not addressed; it is assumed that the explainer is provided with queries to explain (e.g., Josephson & Josephson, 1994; Mitchell et al., 1986; O'Rorke, 1994; Zadrozny, 1994). The case-based model proposes a method for automatically generating appropriate targets to fill in gaps in understanding, based on detection of understanding failures revealed by *anomalies* that arise during the understanding process.

The third issue is *what to explain* about a situation selected for explanation. Often, the answer to this question is that an explanation must provide a proof that the fact to be explained follows from prior knowledge and plausible abductive hypotheses. Although the case-based explanation viewpoint agrees that an important role of explanation is to show why new facts follow from prior knowledge, the understanding task that it addresses—resolving anomalies by reconciling surprising information with conflicting prior beliefs or expectations—places a new requirement on explanations as well. Instead of simply connecting an assertion to prior beliefs and assumptions, explanations of anomalies must also identify the flaws in the understander's prior beliefs that led it to generate flawed expectations. Thus instead of assuming that prior beliefs are correct, the case-based model is aimed at identifying and repairing flaws in prior beliefs.

The fourth issue is *how explanations should be generated.* In many models of abduction, explanation generation is done by backwards chaining, starting from scratch from a query to explain. We contrast this method to case-based reasoning, and also contrast the tenets of case-based reasoning to those of another method for learning from explanations, explanation-based schema acquisition.

The fifth issue is *how to evaluate* explanations in order to select the "best" explanation. In abduction systems this evaluation is normally based solely on plausibility estimates, often using structural criteria based on Occam's razor; we contrast this method to the similarity- and experience-based methods of case-based explanation. We also demonstrate that task-based information needs can play a crucial role in deciding the goodness of explanations.

The final issue is *the relationship between explanation evaluation and generation.* In many models, these processes are assumed to take place as two independent sequential steps: a complete set of candidate explanations is generated and then the candidates are compared to select the best explanation. The case-based explanation model presents a way to integrate explanation construction and evaluation, using evaluation of partial explanations to provide very specific guidance about how to proceed when generating new candidate explanations. This integration helps to focus the explanation generation process. The following sections examine each of the six issues in more detail.

4 The nature of explanatory reasoning: deductive proofs vs. plausible reasoning

A fundamental question for any theory of abduction is what constitutes an explanation. Peirce (1948)[p. 151] characterized abductive inference with the following framework:

> The surprising fact, C, is observed;
> But if A were true, C would be a matter of course,
> Hence, there is reason to suspect that A is true.

This process is often treated as a form of "deduction applied in reverse" (Charniak & McDermott, 1987).[4] The result of this process is a deductive proof whose premises, rather than being definitively known, may include "abductive assumptions" that are generated during the explanation process. If the abductive assumptions are in fact true, the explanations will entail the state or event being explained (e.g., Charniak, 1986; Hobbs et al., 1993; Kautz & Allen, 1986; Konolige, 1990; O'Rorke, 1994; Poole, 1989).

An assumption of such approaches is that the explainer has access to a perfect domain theory from which to generate explanations. Unfortunately, as is widely known, perfect domain theories are unattainable in the everyday world (e.g., Leake, 1992; Mitchell et al., 1986). This classic "imperfect theory problem" is often addressed by methods aimed at repairing known flaws in a domain theory (e.g., Dietterich & Flann, 1988; Ourston & Mooney, 1994; Pazzani, 1990; Rajamoney, 1993), which may yield valuable refinements of a domain theory. Unfortunately, however, rules describing the everyday world can never be perfect. For example, even if a reasoner has perfect knowledge of the world, rules can never include all the factors that may potentially be relevant in a situation. This problem, known as the qualification problem (McCarthy, 1980), is illustrated by McCarthy's famous observation that any rule stating that a car will start when the ignition is turned on must in principle depend on an infinite set of conditions such as "there isn't a potato in the tail pipe." Consequently, it is useful to guide the choice of domain rules to apply in order to favor those rules and rule combinations that have proven appropriate in similar previous situations.

The focus of the case-based approach is to allow generation of reasonable explanations, given the likelihood that flaws will exist in the explainer's domain theory. Consequently, the case-based model explicitly treats explanations as plausible inference chains rather than deductive proofs. In the case-based explanation model, explanations are represented as *explanation patterns* (XPs) (Schank, 1986). Explanation patterns encode chains of belief dependencies showing how belief in a conclusion follows from belief in a set of premises. Syntactically these are like deductive explanations, but the derivation in an explanation pattern is not considered to entail the chain's consequent. Instead, it simply provides support for why the consequent might be expected to hold. Treating explanations as involving plausible reasoning, rather than as deductive proofs, has ramifications for how to generate explanations, as will be discussed in section 7, and for what should be learned from explanations, as is the focus of section 7.2.

The view that explanations are based on plausible reasoning is shared by models that use probabilities or assumption costs as the basis for selecting the "best" explanation (Charniak & Shomony, 1994; Hobbs et al., 1993; Pearl, 1988). Those methods differ from case-based explanation, however, in assuming that information on the probability or "cost" of each assumption and rule is available to the explainer. That assumption requires extensive knowledge of relationships between beliefs that may not be available for explanation of everyday events. The case-based model instead assumes that only coarse-grained likelihood information is available, and, as will be described in section 8.1, augments that information with a more holistic measure focusing on similarity to prior explanations.

Another difference between explanation patterns and deductive proofs is that explanation patterns are not treated as if they enumerate a sufficient set of conditions to entail the event that they explain. Just as everyday human explanations are often incomplete, the information included in an explanation pattern may omit some causally relevant factors. Explanation patterns simply

[4] As Zadrozny (1994) points out, such a description is primarily suggestive; he presents logical theory of abduction intended as a step towards a calculus of abduction.

highlight a limited set of factors involved in a situation that are relevant to the explainer's needs for information. For example, an XP might involve an explanation stating that cold weather causes a car not to start. That rule obviously omits many relevant factors (e.g., the state of the battery and the grade of oil in the engine), but it can nevertheless be useful for directing action. This partial explanation focusing on the cold is all that is needed, for example, to avoid problems by keeping the car in the garage on cold nights. Section 8.2 discusses the role of explainer needs for information in selecting appropriate factors to include in an explanation.

5 When to explain

In order to apply the explanation process, an explainer must decide when explanation is merited. This decision is beyond the scope of the many theories of abductive explanation that start by assuming that there is a query to explain (e.g., Charniak & Shimoney, 1994; Josephson & Josephson, 1994; Konolige, 1990; Levesque, 1989; O'Rorke, 1994; Poole, 1989; Zadrozny, 1994; an exception is Mooney, 1990). Likewise, theories of deductive explanation often start by assuming that their target concepts are provided as input (e.g., Mitchell et al., 1986; an exception is Kedar-Cabelli, 1987).

Abductive understanding systems commonly attempt to build explanatory chains to account for every input (Charniak, 1986; Hobbs et al., 1993; Rieger, 1975; Wilensky, 1983). However, because of the expense of generating explanations, and because of the enormous number of events and features of events that could potentially be explained, attempting to explain all facets of a situation is an overwhelming task. One way to address this is to try to make explanation generation more efficient, as will be discussed in section 7. Another way to address it, which has received less attention, is increased selectivity about when to explain.

In order to be selective about when to explain, a reasoner needs criteria for when the effort of generating an explanation is merited. This depends in turn on the task to be served by explanation. The primary task served by explanation in the case-based explanation systems described here is explaining novel events that cannot be accounted for by pre-existing schemas; the background processing of the systems, schema-based story understanding (Charniak, 1978; Cullingford, 1978; DeJong, 1979; Lebowitz, 1980; Minsky, 1975; Mooney, 1990; Schank & Abelson, 1977), is augmented by generating new explanations as needed to understand novel events. Consequently, explanation is motivated when existing schemas prove insufficient. However, this raises the question of when to consider the existing set of schemas to be insufficient.

DeJong and Mooney (1986) propose one answer: that a schema-based understanding system should generate a new explanation whenever a new fact fails to fit into any existing schema. Although that approach is more constrained than explaining every input, it still can involve explaining large numbers of new events, and it provides no guidance as to specific features to focus on when explaining. (In practice, their systems explain with a fixed focus on the motivations of the actors involved.)

The case-based approach proposes a different answer. It assumes that the system begins with a fairly extensive set of schemas—a sufficient set to capture the classes of events that are of interest—and explains only in response to evidence of flaws in those schemas or in previous beliefs. Its explanation process is triggered when *anomalies* arise, i.e., when new information conflicts with prior beliefs and expectations. The goal is only to account for the surprising aspects of those

- John needed money to pay back a loan shark for gambling debts.
- John believes that robberies of ATMs are more likely to succeed than bank robberies.
- Mark was sick, forcing John to replace him at the last minute.
- The bank's security camera had been removed for repair.

Table 1: Four explanations that could be generated for "John broke into an automatic teller machine."

situations, because only those aspects reveal problems in current understanding. Explaining only anomalies provides much stronger focus than explaining each event that is not accounted for by a pre-stored schema.

The value of anomalies in providing motivation and focus for remedial explanation has been observed both in the artificial intelligence literature (Hammond, 1989; Leake, 1988, 1992; Schank, 1986)[5] and in psychological studies showing that anomalies are an important trigger for prompting learning in human understanders (Chinn & Brewer, 1993; Garner, 1981; Otero & Campanario, 1990). As is described in the following section, anomalies not only provide guidance about when to explain, but of what to explain as well.

6 What to explain

Models of abductive explanation often attempt only to generate *some* plausible explanation for the fact they are explaining. However, in everyday explanation more focus is often needed: there are many ways to explain a given fact, each providing different information. For example, table 1 presents four possible explanations for the event "John broke into an automatic teller machine (ATM)." All these explanations are possible ways to account for the event, and they may all be valid simultaneously. However, their appropriateness as responses to a query depends strongly on the motivation for the query.

When the query is motivated by an *anomaly* arising during understanding, the explanation that is appropriate depends on why the event was found to be was anomalous. For example, if what was anomalous was that John performed the robbery instead of Mark, who was previously expected to perform the crime, the explanation that Mark was sick would be relevant but the other explanations would not. Likewise, if what was anomalous was that the break-in succeeded despite bank precautions, the missing security camera would be relevant and Mark's illness would be irrelevant.

It might appear possible to generate a "complete" explanation that included all of the factors relevant in every context. However, even if the qualification problem could be set aside, attempting to generate as complete an explanation as possible severely aggravates the problem of the cost involved in generating explanations: in everyday explanation there are simply too many possible facets of a situation on which to focus to attempt to explain them all.

[5]According to report of a spoken communication with Harry Pople, anomalies are also used as focus for explanation in Pople's EAGOL system for power-plant diagnosis, but no description of that system has yet been published (Josephson & Josephson, 1994, p. 265).

In general, unless the explainer focuses on resolving the anomaly, rather than simply accepting any derivation of the anomalous event, there is no guarantee that the explanation will be useful for repairing flawed understanding. In limited domains it may be possible to restrict the rules available to the explainer to assure that only task-relevant explanations may be generated, but for everyday explanation tasks, that restriction of rules may not be possible. Consequently, it is necessary to examine the requirements for resolving anomalies and to target explanations appropriately.

In the view of explanation developed in the case-based model, an explanation to resolve an anomaly must answer two questions. The first question, "why did the surprising event (or features of the event) happen?", must be answered to make sense of the surprising fact. Answering it involves generating a reasoning chain supporting the surprising event or its surprising features, as done by traditional abduction systems when they do backwards chaining to find a derivation of an input fact. The second question to answer is "why was the wrong belief or expectation formed?" The answer to that question can be used to guide revision of flawed prior knowledge.

The task of identifying flaws in prior reasoning is radically different from the task of deriving why a surprising event happened in a neutral context. Abductive systems that only derive why a surprising event happened assume that their prior beliefs are correct; they attempt to derive the new fact from those beliefs and from new assumptions that are consistent with prior beliefs. The task of identifying reasoning flaws, however, requires an explanation to include some information that actually conflicts with or supersedes prior reasoning. The reason is that in order to explain why an inaccurate expectation was generated, it is necessary to show where the reasoning leading to the expectation went wrong.

There are multiple reasons that prior reasoning may have gone wrong. It may be that relevant information was missing or overlooked, leading to the wrong conclusions, or that some belief on which the explanation was based was simply false. An explanation of an anomaly provides the information needed to resolve the reasoning failure provided it highlights information that, had the information been considered previously, would have prevented the flawed expectation from being generated. For example, if an ATM robbery is anomalous because the understander expected that the robbery would have been prevented by constant remote monitoring of the ATM, explaining the robbery by "The bank's security camera had been removed for repair" is sufficient: it shows that the previous belief of monitoring was invalid, invalidating the reasoning that led to the expectation that a robbery could not take place. That information, combined with the routine information that robberies are likely if valuables are left unguarded, answers both questions involved in resolving the anomaly. This example illustrates that the target for an explanation of an anomaly is not only a surprising state or event, but a failure of prior reasoning to generate proper expectations (Leake, 1991b, 1992; Schank, 1986; Schank et al., 1994). As discussed in the next section, this has important ramifications for the problem of generating candidate explanations.

7 How to generate candidate explanations

Once the target for explanation has been chosen, the issue is how to generate an explanation relevant to that target. In this section we first discuss problems involved in generating explanations by backwards chaining from scratch. We then discuss case-based reasoning methods, contrasting their re-use of specific explanations to the re-use of generalized explanations done by explanation-based schema acquisition systems.

7.1 Explaining by backwards chaining

The computational complexity of generating explanations is a significant problem for abductive reasoning. Formal complexity analyses show that in general it is NP-hard (Bylander, Allemang, Tanner, & Josephson, 1991; Selman & Levesque, 1990); practical experiments demonstrate problems arising from computational cost for real-world abductive reasoning tasks (Tuhrim, Reggia, & Goodall, 1991). In abductive understanding systems, standard theorem-proving chaining techniques are often the mechanism for generating candidate explanations (e.g., Hobbs et al., 1993; Kautz & Allen, 1986). These methods take an event as the target for explanation and generate a proof of why that event follows from prior knowledge and abductive assumptions, reasoning from scratch using backwards chaining through a space of rules or operators from the event. For these methods, the problem of computational cost arises due to the "combinatorial explosion" of alternatives to pursue when generating an explanatory chain. In response to this problem, many ideas have been proposed for reducing chaining cost during the search for explanations, such as combining of top-down and bottom-up processing (Wilensky, 1983), limiting the maximum chain length (Mooney, 1990), using heuristics to limit the branching factor of search (Hobbs et al., 1993), using marker-passing to propose candidate paths (Charniak, 1986; Norvig, 1989), making simplifying assumptions about the explanations (Chien, 1989; Tadepalli, 1989), and using plausibility estimates to guide the choice of which explanations to pursue (de Kleer & Williams, 1989; Ng & Mooney, 1990). Nevertheless, the practical problem remains.

The task of explaining anomalies—accounting for reasoning flaws as well as accounting for surprising events—aggravates the problem. To resolve an anomaly, not only must the explanation account for the event, but it must also provide information showing the error in the understander's prior reasoning that made the event surprising.

At first glance it appears models of abductive explanation within the backwards chaining framework already can provide the needed focus through selection of the right queries to prove. In fact, it is easy to tailor a query to focus on a particular surprising aspect of an event (e.g., if the actor's decision-making was surprising, that decision-making could be the object of the query). However, this provides only part of the needed focus for actually resolving an anomaly.

For example, suppose that the anomalous aspect of the event "John broke into an ATM" is that John was not seen. If the query chosen to explain is "John was not seen during the break-in," the explanation is guaranteed to focus on the anomalous aspect of the situation. Unfortunately, however, an explanation for that query may still fail to address the reasoning failure underlying the anomaly. Many different explanations can be generated for the query "John was not seen during the break-in," and their relevance will depend on the particular reasoning chain that led to the expectation that John would be seen. An explanation such as "the lights were broken" is relevant if the understander thought that John would be seen because of a reasoning chain including that the ATM was well-lighted; it is not relevant if the understander already knew that the ATM was dark but also knew that the guard had night-vision goggles that would enable him to see John regardless of the darkness. The need to address this the background of prior reasoning cannot be directly reflected in the type of query provided to explanation systems based on backwards chaining. Their queries specify a state or event to explain (i.e., an assertion to prove), but the information that an explanation of an anomaly must provide depends not only on the fact being explained but on the *relationship* between the contents of the yet-to-be-generated explanation and the flawed background knowledge.

Thus relevance of an explanation depends not only on the aspect of the event being explained (which is easily captured by selection of the state or event to take as the starting point for backwards chaining), but on whether the derivation of that explanation provides information showing flaws in the understander's prior reasoning such as erroneous or overlooked beliefs. As previously discussed in section 6, this problem of accounting for erroneous beliefs is beyond the scope of traditional models of abductive explanation: they assume that all prior beliefs are correct and reject explanations conflicting with those beliefs. The additional constraints it imposes will also increase the difficulty of explanation. For example, in another context Selman and Levesque (1990) have proven that requiring explanations to include assumptions from a particular set of assumptions dramatically increases the complexity of finding a non-trivial explanation with the ATMS explanation procedure, making it NP-hard. However, everyday explanation must often resolve anomalies, which requires generating explanations that include a particular subset of beliefs conflicting with prior reasoning.

Addressing the focus problem in the case-based framework: The backwards chaining process provides a means to focus on one important part of an explanation—accounting for why an event makes sense—but not for the other component, showing why the reasoning leading to prior beliefs was erroneous. Case-based explanation construction, however, can focus search for explanations on candidates that simultaneously address both needs. The explanations stored in the explanation library of the program ACCEPTER are organized and retrieved using an indexing scheme in which the indexing vocabulary reflects both which features of a situation were anomalous and why they were anomalous, so that the process for retrieving stored explanations can focus on explanations relevant to both points (Leake, 1991b, 1992).

To illustrate ACCEPTER's indexing vocabulary, we describe some of the categories that could apply to the ATM robbery example. If the anomaly were the conflict between the ATM break-in and the prior belief that John was too cautious to perform risky actions, the anomaly would be an instance of the category SURPRISING-PLAN-CHOICE. That category indexes explanations applying to uncharacteristic plan choices, such as being driven to select an unusual plan because of desperation (e.g., "John needed money to pay back a loan shark for gambling debts"). If the anomaly were that the robbery were successful despite expectations for inviolable security, the anomaly would be an instance of BLOCKAGE-VIOLATION, used to organize explanations for events occurring despite preventative steps (e.g., "The bank's security camera had been removed for repair").

In ACCEPTER's memory, explanations are first organized by the anomaly categories and then by additional features of the anomalies such as the specific actor or event involved. The process for retrieval of prior explanations focuses on prior explanations whose anomalies match the current anomaly as specifically as possible. For example, a new instance of a BLOCKAGE-VIOLATION anomaly concerning a bank robbery will prompt retrieval of explanations for other bank robberies when possible, and otherwise will consider more abstract matches as needed (e.g., involving other crimes that occurred despite preventative measures). More complete descriptions of the categories themselves and how they are used are available in Leake (1991b, 1992).

7.2 The role of experience

Because of the computational cost of generating explanations from scratch, many understanding systems rely on pre-stored generalized schemas to provide a form of explanation for stereotyped events. These schemas, in the form of knowledge structures such as "frames", "scripts," or "MOPs" (Charniak, 1978; Cullingford, 1978; DeJong, 1979; Lebowitz, 1980; Minsky, 1975; Schank & Abelson, 1977; Schank, 1982), encapsulate information about the standard features of stereotyped situations (e.g., eating in a restaurant). When new information is placed in such a knowledge structure (e.g., a person is recognized as filling the "waiter" role in a restaurant), the knowledge structure makes sense of the information by relating it to the pre-existing connections to stereotyped events provided by the knowledge structure (e.g., that the waiter is expected to take a customer's order). In addition to the efficiency benefits for understanding systems, schema-based approaches to diagnosis (Stern & Luger, 1992; Turner, 1994) have been shown to be very useful for controlling the search for candidate explanations.

Despite the benefits for efficiency, however, schema-based models are limited in their flexibility to address novelty: Because their understanding process depends entirely on applying their pre-existing knowledge structures, purely schema-based systems have no way to deal with novel situations. Consequently, a question that arises in any model depending on pre-stored knowledge is how to augment that knowledge in response to new experiences. One method is to use case-based reasoning; another is explanation-based schema acquisition, in which explanations are used to highlight aspects of a new situation to include in new schemas for future use (DeJong & Mooney, 1986; Mooney, 1990). Although the two methods share similar goals in combining flexibility and efficiency, there are important differences in the stances of the two methods, with the case-based approach particularly well-suited for explaining from the imperfect domain knowledge that must often be used in everyday explanation. A discussion of the differences follows.

Case-based explanation vs. explanation-based schema acquisition: A key difference between case-based explanation and explanation-based schema acquisition concerns the preferred level of generalization. Explanation-based schema acquisition assumes a sufficiently high-quality domain theory to allow immediate generalization of new episodes whenever licensed by the rules of the domain theory. Whenever explanation-based schema acquisition systems encounter new situations that do not fit previous generalizations, they first explain the situation by doing backwards chaining, using their library of previous rules and schemas. After completing an explanation, they immediately perform explanation-based generalization of the explanatory chain to form a new generalized schema for future use. Case-based explanation instead takes a very conservative approach to generalization. At the time an explanation is generated, case-based explanation simply stores *that specific explanation.* If that explanation must be generalized to apply to another situation, the generalization is done only at the time that the explanation must be re-applied, and only to the extent required to explain the new situation.

The case-based approach's choice of need-based generalization reflects the fundamental difference between viewing explanations as deductive proofs, as in explanation-based learning, and viewing explanations as plausible reasoning chains, as in case-based reasoning. When explanations are considered deductive proofs based on a perfect domain theory, it follows that explanation-based generalization is guaranteed to give valid results. However, if explanations are simply plausible reasoning, as in the case-based explanation view, there is no assurance that an explanatory chain will include all factors relevant to determining whether the chain is valid in other contexts. Consequently, even if an explanation is sufficiently complete to apply in the context in which it was built, generalizing as far as is licensed by the rules that it uses may lead to erroneous overgeneralization because of the failure to include contextual factors omitted from the explanation.

The case-based method is more conservative in that explanations are stored in their specific form and re-applied in with as little generalization as possible, rather than generalized as far as is licensed by the rules they involve. In this way, the re-use of explanations can implicitly reflect useful information about specifics that are not captured by the domain theory. For example, if a car breaks down and has previously broken down with similar symptoms, a case-based explanation system would favor the prior explanation for the most similar situation encountered previously. Consequently, even if the domain theory and explanation did not explicitly include all the factors that might be important for generalizing the explanation (e.g., that the explanation only applied to cars with a certain type of engine layout), it would be safe to re-apply it to the similar cars under similar circumstances. In that way, the use of cases may implicitly capture the influence of features outside of the understander's domain theory.[6]

It might appear that the conservatism of case-based explanation would itself lead to problems: Refusing to allow explanations to be used in novel contexts would of course severely limit the usefulness of prior experience. Some method to apply explanations to new circumstances is needed. To this end, case-based explanation replaces immediate generalization with case adaptation done when a previous explanation must be applied to a new situation. This allows explanations to be re-used in a manner that is both more conservative about the reliability of domain rules and more flexible for explaining novel situations.

The case-based method is more flexible than explanation-based generalization in the following sense. If it is necessary to adapt an explanation to a new situation, the strategies allowed to revise the explanation are more powerful than explanation-based generalization alone. During adaptation, parts of the explanation can not only be generalized but also added or substituted (Kass, 1990, 1992, 1994). This allows case-based explanation to apply prior explanations to a wider range of circumstances than is possible for explanation-based generalization.

Because of the flexibility of case adaptation, it is not practical to perform all possible adaptations in advance—an enormous number of variant explanations would be generated unnecessarily. Consequently, in contrast to the "eager" generalization strategy of explanation-based learning, which does all possible generalization when an explanation is generated, case-based explanation waits to adapt an explanation until adaptation is needed to apply it to a specific problem. When adaptation is done, the amount of adaptation is limited to what is needed in order to apply the explanation to the particular anomaly being explained. This has the benefit of saving effort by avoiding forming generalizations until it is known that they will actually be needed. Doing adaptation "to order" for a specific problem also has the benefit of making it possible to evaluate the reasonableness of the results of adaptation for that specific situation, allowing adaptations that result in implausible explanations in that context to be detected and abandoned, even if those adaptations are licensed by general system rules.

As an example of the contrast between case adaptation and explanation-based generalization, consider the namesake example of the SWALE system, explaining the premature death of the race-horse Swale. When the SWALE system explains Swale's death, one of the explanations it retrieves concerns the death of another young superstar, Janis Joplin. Joplin was driven to recreational drug use by the stress of being a star and the availability of recreational drugs, and she died from an accidental drug overdose. An explanation-based schema acquisition program that had previously explained Joplin's death would have generalized the explanation for Joplin's death at the time it was built, to form a general schema such as "stress and access to drugs can lead to death from

[6]Obviously, a central issue is how to determine similarity; see Leake (1992) for a discussion of the similarity assessment scheme that is used by the case-based explanation systems being described.

accidental overdose." However, such a generalization does not apply to Swale. Consequently, an explanation-based schema acquisition system could not apply that stored generalization to Swale's death. As a result, it would chain together a new explanation without guidance from the Joplin explanation.

In the case-based approach, however, the way to apply the Janis Joplin explanation is not determined by a precomputed generalization. Instead, after retrieving the specific explanation of Joplin's death SWALE attempts to decide how to adapt that explanation in light of the particulars of the episode—Swale's death—to which it will be applied. During adaptation, SWALE abandons the parts of the explanation that do not apply and retains the kernel of the explanation that is potentially applicable to a racehorse: the hypothesis that a drug overdose caused the death. That hypothesis is unsupported, so adaptation takes drug overdose as a starting point (considerably narrowing the field of options to consider compared to simply attempting to explain Swale's death from scratch) and attempts to adapt the explanation to include additional factors supporting the occurrence of a drug overdose. SWALE's knowledge includes that racehorses are sometimes given performance-enhancing drugs. Using that knowledge, it generates the explanation that Swale might have died from an accidental overdose of performance-enhancing drugs.

Thus the case-based approach uses experience to suggest alternatives even in situations that are not straightforwardly subsumed by generalizations of prior explanations, allowing more flexible reuse of the results of prior explanation construction. This process depends on having effective strategies for guiding adaptation, and such strategies are described in (Kass, 1990, 1992, 1994).

8 Criteria for the "best" explanation

In order to choose the explanation to accept, an abductive explanation system requires criteria for what constitutes the "best" explanation. The literature on abduction almost universally bases this decision solely on criteria that in some sense measure the "plausibility" of explanations (Charniak, 1986; Charniak & Goldman, 1991; Charniak & Shomony, 1994; Hobbs et al., 1993; Josephson & Josephson, 1994; Kautz & Allen, 1986; Kass, 1990; Konolige, 1990; Levesque, 1989; O'Rorke, 1994; Mooney, 1990; Poole, 1989; Zadrozny, 1994). While we agree that plausibility considerations are often crucial in selecting explanations, we also believe that goal-based criteria reflecting the explainer's needs for information are crucial in deciding the goodness of explanations. Thus in our view, explanation is more properly viewed as a means for obtaining useful information than as a goal-neutral process. In this section we begin by considering the role of plausibility and then examine the effects of goals on what constitutes the "best" explanation.

8.1 Plausibility criteria

Models of how to select the best candidate explanation are often based on minimality criteria using Occam's razor to compare sets of explanations according to structural criteria such as number of assumptions or structural coherence; other methods perform theorem proving to determine the consistency of new information with old, or use probabilities or "assumption costs" to judge which explanations are most likely. This section contrasts those and other methods to the criteria used to decide plausibility in case-based explanation, which are strongly influenced by *similarity* to previously-explained episodes and to stereotyped patterns.

Structural minimality criteria: Many models of plausibility evaluation judge plausibility according to Occam's razor, with explanations ranked by "simplicity" or "minimality" according to some syntactic minimality criterion (Charniak, 1986; Kautz & Allen, 1986; Konolige, 1990; Levesque, 1989; Poole, 1989; Wilensky, 1983). Structural comparisons are neutral to the content of the explanation, focusing instead on factors such as the lengths of the chains involved in the explanation (favoring the shortest chains) (e.g., Wilensky, 1983) or the number of abductive assumptions they require (e.g., Charniak, 1986). Structural comparisons can also be aimed at measuring the "coherence" of explanations (Ng & Mooney, 1990; Thagard, 1989).

When structural methods are used as the sole criteria for evaluating candidate explanations, two problems result. First, structural properties alone are simply inadequate to rank the likelihood of explanations reliably. For example, judging plausibility by the number of assumptions involved in an explanation may be misleading because two commonplace assumptions may be more likely than one unusual one.

The second problem with structural criteria is that the only information that structural criteria can provide is a comparative ranking. Even if this ranking is correct, it is not sufficient to determine whether an explanation is acceptable unless the set of explanations being compared is known to be complete. If only some of the possible explanations are compared, the explanation ranked as best in the comparison may still be one that would not be ranked as best overall.[7]

Unfortunately, in everyday explanation, incomplete sets of candidate explanations are unavoidable. It would be an overwhelming task to try to generate *all* possible explanations for an everyday event. Rather than being presented with a complete set of candidate explanations to compare, everyday explainers must generate a stream of explanations and decide when to stop generating explanations, even if more candidates could be generated. That decision depends not on which available explanation is most plausible, but on whether *any* of those candidates is sufficiently plausible.

Proof-based approaches: In models of abduction that treat explanations as deductive proofs, the validity of explanations depends entirely on the validity of their abductive assumptions. One way to verify abductive assumptions is to use theorem-proving to show that the assumptions are actually entailed by prior beliefs. Likewise, it is possible to check whether an assumption is inconsistent with prior beliefs by attempting to prove its negation from prior beliefs. For example, Charniak (1986) describes a method that judges abductive assumptions in these two ways. A drawback of this approach is its computational cost: deciding the plausibility of an explanation may require generating multiple proofs, each of which again raises the problem of the combinatorial explosion.

Probabilistic and cost-based criteria: Probabilistic and related cost-based approaches have also been proposed to judge the reasonableness of both assumptions and explanatory chains (Charniak & Shomony, 1994; Hobbs et al., 1993; Lin, 1992; Pearl, 1988). However, fine-grain probability or cost information will not necessarily be available for everyday explanation problems. Another

[7]Tuhrim et al. (1991) provide an empirical examination of the effectiveness of a number of minimality criteria for expert diagnosis tasks. Their analysis provides support for abductive inference methods but also identifies a number of problems arising from minimality-based approaches to selecting explanations. They attribute some of those problems to lack of knowledge in their knowledge base, which again underlines the importance of using abductive explanation methods that are robust with respect to incomplete domain knowledge.

alternative is to simply use coarser-grained likelihood estimates rather than probabilities. However, that method will often be insufficient to distinguish between the many possible candidate explanations for an everyday event.

Because of the problems with structural criteria and coarse-grained likelihood estimates, the case-based model implemented in ACCEPTER relies primarily on a novel approach compared to other abductive understanding systems: favoring explanations that are analogous to prior explanations for similar situations.

ACCEPTER's plausibility criteria: In case-based explanation, the most important criterion for judging plausibility is similarity-based: Explanations of new anomalies are favored if they are similar to explanations that applied to similar prior anomalies. This similarity judgment is done implicitly through the case retrieval process; retrieval of stored explanations is aimed at retrieving explanations from similar prior situations, based on the similarity criteria reflected by ACCEPTER's indexing vocabulary for anomalies. This emphasis on similarity is in the same spirit as the analogical model of explanation generation developed by Falkenhainer (1990).

After an explanation has been retrieved and adapted, its plausibility is judged in two additional ways. First, the likelihoods of the abductive assumptions *and* of the inference links and intermediate conclusions of the explanatory chain are evaluated. Although approaches that assume a perfect domain theory need only to verify the reasonableness of abductive assumptions, with an imperfect domain theory the entire reasoning chain must be verified.

The method used to estimate the reasonableness of the assumptions and rules in an explanatory chain is to compare them to standard stereotypes. When conflicts occur, those conflicts are flagged as plausibility problems (Leake, 1992, 1994a). No inference is done to evaluate how their ramifications interact. The motivation for estimating likelihood by similarity to stereotyped patterns, rather than using formal probability calculations, is the need to decide plausibility even when probabilities are unavailable. The motivation for the stereotype-based method, as opposed to methods that attempt to prove the consistency of assumptions with prior beliefs, is to control verification cost. Judging plausiblity according to limited verification is consistent with a number of psychological studies that suggest that human inferencing and verification during routine reading comprehension are fairly limited, tending towards establishing local rather than global consistency (Baker & Anderson, 1982; McKoon & Ratcliff, 1992; Sanford, 1990; Vonk & Noordman, 1990).

When similarity and coarse-grained likelihood criteria are insufficient to distinguish between available candidate explanations, the evaluation process falls back on a simple structural minimality criterion: given two explanations that explain equally similar prior situations, and whose beliefs are considered equally likely, the explanation with the fewest assumptions is favored.

One important evaluation criterion omitted from this model is the closeness of competing candidates; if the two best candidates are equally plausible, their closeness may suggest that neither one should be accepted until they can be better distinguished (e.g., Josephson & Josephson, 1994; Miller, Pople, & Myers, 1982).

8.2 Goal-based criteria

As observed in the introduction, in standard models of abduction, the "best" explanation is simply the "most plausible." Although it has been pointed out in the abduction literature that pragmatic factors determine the level of certainty to require in an explanation, as when a doctor requires a diagnosis with high certainty before attempting a risky operation (Josephson & Josephson, 1994), pragmatic factors usually do not enter elsewhere into the explanation evaluation process. A number of researchers have begun to point out that in fact the information provided by an explanation also plays a central role in abduction (Krulwich, Birnbaum, & Collins, 1990; Leake, 1988, 1992, 1994a; Norvig & Wilensky, 1990; Ram & Leake, 1991). In the view of explanation developed in research on case-based explanation, pragmatic factors play a key role: what constitutes the ultimate goodness of an explanation is its ability to satisfy the needs for information that motivated generating it.

To illustrate that usefulness considerations go beyond plausibility judgments, recall the four alternative explanations for an ATM break-in in table 1. *All* of the listed explanations could be valid simultaneously, but even if they are, their appropriateness will differ depending on the tasks for which they will be used. For example, a childhood friend of the robber who was surprised by the robbery might wonder what drove the robber to commit the crime, making the need to pay back a loan shark the best explanation; a bank officer might wonder how the robbery had been able to succeed, making the absence of the camera the best explanation. 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 (e.g., Hale & Barsalou, in press; Lalljee & Abelson, 1983; Snyder, Higgens, & Stucky, 1983) and philosophical works (e.g., Mackie, 1965; Van Fraassen, 1980) argue that different explanations are needed to reflect different explainer goals. For example, subjects attempting to absolve themselves of blame will focus on different features of a situation from those stressed by subjects without that goal (Snyder et al., 1983). In general, in any multi-task system the only way to assure useful explanations is to explicitly evaluate their goodness according to current system goals (Leake, 1991a, 1992). However, as observed previously, models of abductive explanation seldom consider the effect of different possible uses of explanations on which explanation to favor.

The influence of intended uses for explanations on explanation generation has begun to be investigated in abductive diagnosis for tasks such as integrating medical diagnosis and response (Rymon, 1993), and performing medical diagnosis within a planning framework (Turner, 1994). The largest body of artificial intelligence research considering the role of goals in explanation, however, is in explanation-based learning (EBL) (Mitchell, Keller, & Kedar-Cabelli, 1986; DeJong & Mooney, 1986).[8] Yet although EBL has been applied to many tasks, such as learning rules for object recognition, problem solving, and search control, their role in all these tasks can be viewed as falling into the same broad class: forming rules for concept recognition. For example, in the formulation developed by Mitchell et al. (1986), the basic process starts from a training example, a target concept, and operationality conditions that determine when the premises of an explanation involve predicates that are easily evaluated (e.g., for visual object recognition, the measure of whether a predicate is easily evaluated might be whether it describes a property that is recognizable by the vision system). Given an instance of an object (e.g., a cup), and given a target concept describing a desired function for the object (e.g., to hold liquids), an explanation can be generated to account for how operational features contribute to membership in the target concept (e.g., that recognizable features such as being "concave up" entail being able to hold liquids). Based on the features identified by the explanation, a new recognition rule can be generated. Keller (1988)

[8]Although that research focuses on deductive explanation rather than abductive explanation, its usefulness criteria are applicable within either framework.

discusses how the apparently diverse tasks of many EBL systems can be viewed as types of concept recognition.

EBL systems addressing the concept recognition problem implicitly reflect the needs of that task by making two basic assumptions about the form of explanations. First, they require explanations to be complete proofs showing sufficient conditions for concept membership. Second, because the types of rules used *within* the derivations (as opposed to the antecedents) are irrelevant to concept recognition—only the antecedents are important to that task, because they are the features that need to be recognized—EBL treats all explanations with the same antecedents as equivalent.

For the tasks that motivate everyday explanation of anomalies, neither of these assumptions may be valid. First, complete explanations may not be necessary; partial explanations are sufficient for many tasks. For example, an explanation for preventing an undesirable event needs only to identify a single necessary condition for the event that is preventable in the future. (E.g., a driver who knows that his car sometimes fails to start when it has been parked in the cold can prevent the problem on cold days by putting it in the garage, even if he does not know all the other factors that may also be necessary for the problem to arise.) In fact, everyday explanations are *necessarily* partial explanations; it is impossible to provide a complete account of the factors that are sufficient for an event to occur. In everyday explanation, it is vital for the explainer to be able to make a principled decision about which partial explanations to accept and to be able to use partial explanations whenever they provide the needed information.

Second, the goodness of everyday explanations often depends not only on the premises they involve, but on how those premises are linked to the conclusion and the types of rules that they use. For example, we might imagine a doctor explaining the symptoms of a disease by using associational rules (e.g., that everyone with a certain symptom has a given disease), or by using causal rules that account for the process by which the symptoms arise from the disease. Both explanations might be equally useful for prescribing a treatment, but the second would be much more useful for determining how to alleviate the symptoms before the cure took effect. Chandrasekaran (1994) discusses how even identical processes must sometimes be described in different ways depending on the needs of the explainer.

Leake (1991a, 1992) discusses the disparate requirements for explanations that arise from ten different tasks, each corresponding to a different way that explanations will be used. Each of these uses places different requirements on explanations and affects which types of antecedents and derivations will provide useful information. For example, explaining the breakdown of a car by "brand X frequently breaks down" provides sufficient information to help a prospective buyer avoid breakdowns (by predicting future problems with brand X and consequently buying a different brand of car), but it does not provide sufficient information for a mechanic to effect a repair. ACCEPTER implements goal-based evaluation criteria for six tasks: resolving anomalies, allowing future prediction, preventing undesirable events, enabling repair, assigning blame, and assigning responsibility. Each of these tasks is associated with particular needs for information that can be used to evaluate the adequacy of explanations and to identify how an inadequate explanation should be adapted to provide additional information.

By considering not only the plausibility of explanations but also how well they satisfy the explainer's needs for information, the explanation evaluation process developed in this research provides a pragmatic and context-sensitive answer to the question of what constitutes the "best" explanation.

9 Interaction between explanation generation and evaluation

Previous sections have discussed the generation and evaluation of explanations as independent issues. A final question, however, is the level of interaction needed between explanation generation and evaluation. Many models of abductive explanation treat explanation generation and evaluation as two sequential steps: an abductive explanation component generates a set of candidate explanations that are then compared to select the "best" explanation from that set (Charniak, 1986; Kautz & Allen, 1986; Peng & Reggia, 1990; Thagard, 1989; Wilensky, 1983). Unfortunately, for any real-world event, arbitrarily large numbers of explanations can be generated. Consequently, an everyday explanation system that attempts to generate all candidate explanations without further focus will be swamped with candidate explanations (e.g., Leake, 1992; O'Rorke, 1989; Poole, 1993). As a result, it is desirable to integrate explanation generation and evaluation of candidate explanations in order to use knowledge of what constitutes a good explanation to focus the search for explanations on worthwhile candidates.

The case-based explanation process integrates explanation generation and evaluation in two ways. The first way is by using its anomaly-based indexing vocabulary to guide retrieval of candidate explanations towards previous explanations addressing similar anomalies. In this way, knowledge of what the explanation must account for is used immediately to guide search for candidate explanations.

The second way is to guide adaptation of retrieved explanations according to incremental evaluation of their plausibility and of the adequacy of the information they provide for the explainer's task. As sketched in section 2, case-based explanation adapts explanations to fit new situations in a cycle driven by the evaluation process. When a new explanation is retrieved, the evaluation process detects any problems and passes their characterizations to the adaptation component. The adaptation component then attempts to revise them in response to the specific problems that were identified. The evaluation/adaptation process can be applied as needed until an acceptable explanation is generated, resource limits for adaptation are exceeded, or no adaptation rules are available for repairing the problems. An important aspect of the process is that explanation evaluation identifies *why* specific parts of the candidate explanation are problematic—which parts of the explanation are implausible, or where needed information is missing—allowing the adaptation system to strongly focus its choice of repairs. In this way, the case-based model uses on-going evaluation not only to choose which explanation to pursue but to provide very specific guidance of how to proceed when augmenting or modifying candidate explanations.

For example, if the explainer's task is to predict an event, a useful explanation must trace the causes of the event back to causes that occur early enough for prediction to be useful. (E.g., if we want to explain the increase in the price of a stock in order to buy low and sell high, it is not enough to attribute the increase to strong profits: We need to be able to predict future profits before others do.) If ACCEPTER's evaluation process determines that the explanation does not include sufficiently early factors, it guides adaptation towards elaborating the explanation to include earlier causes.

The integration between explanation generation and evaluation in our model of case-based explanation differs markedly from the relationship between evaluation and generation of explanations in chaining-based approaches such as explanation-based learning. In explanation-based generalization, goal-based criteria are used to test whether an explanatory chain is sufficient (i.e., whether the leaves of an explanation are operational) but not to guide the choice of which alternatives to pursue when adding to an explanatory chain (Mitchell et al., 1986). In this way, the role of goal-based considerations in EBL is limited to deciding when to stop a goal-neutral chaining process or when to accept a complete explanation (Keller, 1988). Case-based explanation uses goals to decide *which paths to follow and how to follow them* while making decisions about how to repair a flawed explanation.

The integrated approach in our model also differs from models that use plausibility estimates to determine the most plausible candidate explanation so far, in order to preferentially devote further attention to expanding those candidates (de Kleer & Williams, 1989; Hobbs et al., 1993; Ng & Mooney, 1990). Those methods use their evaluation of plausibility to assign a single number to an entire explanation as a whole, in order to decide which explanation to pursue, but not to give any indication of how to proceed in elaborating that explanation. Case-based explanation uses its evaluation of goodness both to identify promising partial explanations and to pinpoint particular aspects of those explanations that need to be fixed. Based on the description of the problem, an adaptation rule tailored to fixing the problem is selected and applied to repair that problem (Kass, 1990, 1992, 1994). This gives more precise guidance.

Thus case-based explanation generation differs from other models not only in its commitments on individual issues such as how to generate explanations and how to evaluate them, as discussed in the five previous sections, but also in its strong integration of explanation generation and evaluation.

10 Conclusion

Abductive explanation is often examined in isolation from the motivations for explaining and the explainer's prior experience. In the resultant models, goals and experience have little effect on explanation. The selection of what to explain is determined outside the explanation process; the goal of explanation is to generate the most plausible explanation, neutral to the particular needs for information prompting explanation; and each new explanation is generated starting from scratch.

This article contrasts such models with an alternative account, developed in research on applying case-based reasoning to explanation, in which explanation is guided by experience and aimed at generating useful explanations. The comparison is organized around six central issues for explanation: the nature of explanatory chains, when to explain, what to explain about a given situation, how explanations are generated, the criteria to use for selecting the "best" explanation, and the level of interaction between explanation generation and evaluation. These points illuminate central issues of everyday explanation and how they are addressed by case-based explanation.

When relevant prior explanations are available as starting points for explanation generation, the potential benefits of retrieving and adapting prior explanations instead of chaining from scratch are threefold: generation of better candidate explanations, by favoring explanations supported by prior experience; increased efficiency of explanation generation over reasoning from scratch; and, because retrieval and adaptation are focused according to system needs, integrating explanation generation and evaluation, providing more precise focus on explanations that are likely to be useful.

Thus in the case-based model of explanation, goals and experience play a key role in explanation generation. By using goals and experience to guide processing, this method provides a way to generate plausible and useful explanations in domains that are complex and imperfectly understood.

References

Baker, L. & Anderson, R. (1982). Effects of inconsistent information on text processing: evidence for comprehension monitoring. *Reading Research Quarterly*, *17*, 281–294.

Bylander, T., Allemang, D., Tanner, C., & Josephson, J. (1991). The computational complexity of abduction. *Artificial Intelligence*, *49*, 25–60.

Chandrasekaran, B. (1994). Functional representation and causal processes. In Yovits, M. (Ed.), *Advances in Computers*. Academic Press, New York.

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.

Charniak, E. & McDermott, D. (1987). *Introduction to Artificial Intelligence*. Addison-Wesley, Reading, MA.

Charniak, E. & Shomony, S. (1994). Cost-based abduction and MAP explanation. *Artificial Intelligence*, *66*, 345–374.

Chien, S. (1989). Using and refining simplifications: explanation-based learning of plans in intractable domains. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 590–595 Detroit, MI. IJCAI.

Chinn, C. & Brewer, W. (1993). Factors that influence how people respond to anomalous data. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pp. 318–323 Boulder, CO. Cognitive Science Society.

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

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

DeJong, G. (1979). *Skimming Stories in Real Time: An Experiment in Integrated Understanding*. Ph.D. thesis, Yale University. Computer Science Department Technical Report 158.

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

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

Falkenhainer, B. (1990). Abduction as similarity-driven explanation. In O'Rorke, P. (Ed.), *Working Notes of the 1990 Spring Symposium on Automated Abduction*, pp. 135–139. AAAI. Technical Report 90-32, Department of Information and Computer Science, University of California, Irvine.

Garner, R. (1981). *Metacognition and Reading Comprehension*, chap. 3. Ablex, Norwood, NJ.

Hale, C. & Barsalou, L. (In press). Explanation content and construction during system learning and troubleshooting. *The Journal of the Learning Sciences.*

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

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

Hobbs, J., Stickel, M., Appelt, D., & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, *63*(1-2), 69–142.

Josephson, J. & Josephson, S. (1994). *Abductive Inference: Computation, Philosophy, Technology.* Cambridge University Press, Cambridge, England.

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

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

Kass, A. (1992). Question asking, artificial intelligence, and human creativity. In Lauer, T., Peacock, E., & Graesser, A. (Eds.), *Questions and Information Processing.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Kass, A. (1994). Tweaker: adapting old explanations to new situations. In Schank, R., Riesbeck, C., & Kass, A. (Eds.), *Inside Case-Based Explanation*, chap. 8, pp. 263–295. Lawrence Erlbaum Associates.

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

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

Kolodner, J. (1993). *Case-Based Reasoning.* Morgan Kaufmann, San Mateo, CA.

Konolige, K. (1990). A general theory of abduction. In O'Rorke, P. (Ed.), *Working Notes of the 1990 Spring Symposium on Automated Abduction.* AAAI. Technical Report 90-32, Department of Information and Computer Science, University of California, Irvine.

Koton, P. (1988). Reasoning about evidence in causal explanations. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pp. 256–261 Minneapolis, MN. AAAI, Morgan Kaufmann Publishers, Inc.

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. (1991a). Goal-based explanation evaluation. *Cognitive Science*, *15*(4), 509–545.

Leake, D. (1991b). An indexing vocabulary for case-based explanation. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pp. 10–15 Anaheim, CA. AAAI.

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

Leake, D. (1993). Focusing construction and selection of abductive hypotheses. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 24–29 Chambery, France. IJCAI.

Leake, D. (1994a). ACCEPTER: evaluating explanations. In Schank, R., Riesbeck, C., & Kass, A. (Eds.), *Inside Case-Based Explanation*, chap. 6, pp. 167–206. Lawrence Erlbaum Associates.

Leake, D. (1994b). Issues in goal-driven explanation. In Ram, A. & desJardins, M. (Eds.), *Proceedings of the 1994 AAAI Spring Symposium on Goal-Driven Learning*, pp. 72–79 Stanford, CA. AAAI.

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.

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

Levesque, H. (1989). A knowledge-level account of abduction. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 1061–1067 Detroit, MI. IJCAI.

Lin, D. (1992). *Obvious Abduction.* Ph.D. thesis, University of Alberta.

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

McCarthy, J. (1980). Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence*, *13*(1 & 2), 27–39.

McKoon, G. & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, *99*(3), 440–466.

Miller, R., Pople, H., & Meyers, J. (1982). Internist-i, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, *307*(8), 468–476.

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.

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.

Norvig, P. & Wilensky, R. (1990). Problems with abductive language understanding models. In O'Rorke, P. (Ed.), *Working Notes of the 1990 Spring Symposium on Automated Abduction*, pp. 18–22. AAAI. Technical Report 90-32, Department of Information and Computer Science, University of California, Irvine.

O'Rorke (1994). Abduction and explanation-based learning: case studies in diverse domains. *Computational Intelligence*, *10*(3), 295–330.

O'Rorke, P. (1989). Coherence and abduction. *The Behavioral and Brain Sciences*, *12*(3), 484.

Otero & Campanario (1990). Comprehension evaluation and regulation in learning from science texts. *Journal of Research in Science Teaching*, *27*, 447–460.

Ourston, D. & Mooney, R. J. (1994). Theory refinement combining analytical and empirical methods. *Artificial Intelligence*, *66*, 273–309.

Pazzani, M. (1990). *Creating a Memory of Causal Relationships: An Integration of Empirical and Explanation-Based Methods.* Lawrence Erlbaum Associates, Hillsdale, NJ.

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

Peirce, C. (1948). Abduction and induction. In Buchler, J. (Ed.), *The Philosophy of Peirce: Selected Writings*, chap. 11. Harcourt, Brace and Company, New York.

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

Poole, D. (1989). Normality and faults in logic-based diagnosis. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 1304–1310 Detroit, MI. IJCAI.

Poule, D. (1993). Probabilistic horn abduction and bayesian networks. *Artificial Intelligence*, *64*, 81–129.

Rajamoney, S. (1993). Designing experiments to extend the domain theory. In DeJong, G. (Ed.), *Investigating Explanation-Based Learning*, chap. 5, pp. 166–189. Kluwer.

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.

Riesbeck, C. & Schank, R. (1989). *Inside Case-Based Reasoning*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Rymon, R. (1993). *Diagnostic Reasoning and Planning in Exploratory-Corrective Domains*. Ph.D. thesis, The University of Pennsylvania.

Sanford, A. (1990). On the nature of text-driven inference. In Balota, D., d'Arcais, G. F., & Rayner, K. (Eds.), *Comprehension processes in reading*, chap. 24. Lawrence Erlbaum, Hillsdale, NJ.

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

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.

Schank, R., Riesbeck, C., & Kass, A. (Eds.). (1994). *Inside Case-Based Explanation*. Lawrence Erlbaum Associates, Hillsdale New Jersey.

Selman, B. & Levesque, H. J. (1990). Abductive and default reasoning: a computational core. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pp. 343–348 Boston, MA. AAAI.

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

Sooriamurthi, R. & Leake, D. (1994). Towards situated explanation. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, p. 1492 Seattle, WA.

Stern, C. & Luger, G. (1992). A model for abductive problem-solving based on explanations templates and lazy evaluation. *International Journal of Expert Systems*, *5*(3), 249–265.

Tadepalli, P. (1989). Lazy explanation-based learning: a solution to the intractable theory problem. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 694–700 Detroit, MI. IJCAI.

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

Tuhrim, S., Reggia, J., & Goodall, S. (1991). An experimental study of criteria for hypothesis plausibility. *Journal of Experimental and Theoretical Artificial Intelligence*, *3*, 129–144.

Turner, R. M. (1994). *Adaptive Reasoning for Real-World Problems: A Schema-Based Approach.* Lawrence Erlbaum Associates, Hillsdale, NJ.

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

Vonk, W. & Noordman, L. (1990). On the control of inferences in text understanding. In Balota, D., d'Arcais, G. F., & Rayner, K. (Eds.), *Comprehension processes in reading*, chap. 21. Lawrence Erlbaum, Hillsdale, NJ.

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

Zadrozny, W. (1994). Is there a prototypical rule of abduction? (Yes, e.g. in proximity based explanations). *Journal of Experimental and Theoretical Artificial Intelligence*, *6*, 147–162.