Evaluating Explanations*

David B. Leake
Department of Computer Science, Yale University
P.O. Box 2158 Yale Station, New Haven, CT 06520

Abstract

Explanation-based learning (EBL) is a powerful method for category formation. However, EBL systems are only effective if they start with good explanations. The problem of evaluating candidate explanations has received little attention: Current research usually assumes that a single explanation will be available for any situation, and that this explanation will be appropriate. In the real world many explanations can be generated for a given anomaly, only some of which are reasonable. Thus it is crucial to be able to distinguish between good and bad explanations.

In people, the criteria for evaluating explanations are dynamic: they reflect context, the explainer's current knowledge, and his needs for specific information. I present a theory of how these factors affect evaluation of explanations, and describe its implementation in ACCEPTER, a program to evaluate explanations for anomalies detected during story understanding.

1 Introduction

Any system that deals with real-world situations will sometimes encounter novel events. Explanation-based learning (EBL) is a powerful method for learning from such situations, often on the basis of a single example. EBL has been the subject of much research; for example, see [DeJong and Mooney, 86] or [Mitchell et al., 86].

Explanation-based systems are only as good as the explanations on which they base their processing, but EBL research concentrates on using an explanation that is assumed to be appropriate, and gives little attention to the problem of finding a good explanation. Researchers often view explanations as deductive proofs, for which validity is guaranteed. But in the real-world situations that people explain, we cannot assume that any candidate explanation is correct, or that only one candidate explanation will be available. People faced with an anomaly often generate and reject a number of hypotheses before finding one they accept. Thus a vital part of understanding novel situations is deciding when an explanation is acceptable.

In psychology, the choice of explanations is considered in *attribution theory* [Heider, 58]. However, since attribution theory considers the choice at a very abstract level, it provides little guidance for finding the specific factors needed to understand an event. More recent work has argued for a knowledge structure approach to attribution, which provides more useful information [Lalljee and Abelson, 83]. In what follows, I first discuss the contributions and difficulties of these approaches. I then present a theory of evaluation and its implementation in ACCEPTER, a story understanding program that that detects anomalies and evaluates candidate explanations for them, taking into account the goals underlying the explanation effort.

*This work is supported in part by the Air Force Office of Scientific Research, under grant 85-0343, and by the Defense Advanced Research Projects Agency, monitored by the Office of Naval Research under contract N00014-82-K-0149. I thank Chris Riesbeck for helpful comments on a draft of this paper.

2 Attribution Theory

Attribution theory [Heider, 58] considers how people decide whether an action should be explained by features of the actor, or of the environment. (Most work on attribution theory assumes that either personal or situational factors will apply, but not both.) Kelley's *covariation principle* [Kelley, 67] hypothesizes that people look at covariation across different people, time, and entities in order to decide which type of factor applies. For example, if John enjoys a movie, but most other people do not, the covariation principle suggests that John's enjoyment should be explained by aspects of John, rather than of the movie. But attribution theory does not go beyond saying that a good explanation involves *some* aspect of John: deciding *which* is beyond its scope, even though people would usually seek that information.

Attribution theory also assumes that explanations are judged by the same criteria regardless of context. However, [Lalljee et al., 82] shows that the explanations people seek, rather than being determined by abstract criteria, vary with circumstances: unexpected behavior requires more complex explanations than expected behavior, and is likely to require more of *both* situational and personal elements.

2.1 A knowledge structure approach

[Lalljee and Abelson, 83] responds to problems in attribution theory by suggesting a knowledge structure approach to attribution. They identify two types of explanation: constructive and contrastive explanation. In constructive explanation, people explain events by accounting for them in terms of knowledge structures such as scripts and plans [Schank and Abelson, 77]. Constructive explanation is useful because it provides expectations for the future. For example, if we hypothesize an actor's goal, we can predict plans he will use to achieve it. Contrastive explanation explains surprising events by showing why they deviated from expectations given by knowledge structures. For example, "John left his bicycle unlocked" might be explained

in terms of *goal reversal*: perhaps rather than having the normal goal of wanting to protect it, he actually wanted to get rid of it.

Explanation-based learners like GENESIS [Mooney and DeJong, 85] do constructive explanation: they try to link observed facts to motivating goals and the plans that achieve them. These plans are then learned for future use. However, such systems do not address the problem of focusing explanation: motivations are not always the most useful aspect of a situation to explain. Nor do they address the problem of judging an explanation's acceptability.

3 ACCEPTER

ACCEPTER implements a theory of what should be explained and what constitutes a good explanation. It is a story understanding program that detects anomalies and evaluates candidate explanations for them.

ACCEPTER's domain is incidents of unexpected death; its primary example is the death of the racehorse Swale. Swale, a star racehorse, was found dead a week after winning an important race. People generated many explanations of his death, but the actual cause was never found. Many of the explanations ACCEPTER evaluates were suggested by deaths that Yale students were reminded of when told about Swale. One student was reminded of the death of the runner Jim Fixx, who died when jogging over-taxed a hereditary heart defect. Although that explanation does not apply directly (since Swale was not jogging before his death), it suggests that Swale might have had a heart-attack because his racing overtaxed a heart defect. Another student was reminded of the death of Janis Joplin, who died of a drug overdose. This suggested the fanciful explanation that Swale took drugs to escape the pressure of stardom, and died of an overdose. It also led to the less frivolous possibility that Swale died from an accidental overdose of performance-enhancing drugs.

ACCEPTER grew out of the evaluation module of the SWALE system [Kass et al., 86]. SWALE uses a case-based approach to generate explanations, and addresses issues of retrieval from memory, revision, and evaluation of explanations. ACCEPTER concentrates on evaluation, performing a wider range of tests and using finer-grained criteria than used for evaluation in SWALE. ACCEPTER maintains a library of explanations, and uses a problem characterization as the index to retrieve possibly-relevant explanations of anomalies. However, the final selection of explanations to evaluate is done by the user of the program, as is revision of problematic explanations.

Expectations from pre-stored schemas guide ACCEPTER's routine processing. When it detects conflicts with these expectations, it retrieves candidate explanations. The user can select one of these or interactively define a new explanation. The resultant explanation is then evaluated, and problems are identified. For example, the explanation *Swale died from jogging + heart defect* is rejected because horses don't jog. The user then has the option of choosing a new explanation or interactively revising to fix the problem. (*E.g.*, replacing jogging by horse-racing as the source of exertion.) ACCEPTER repeats the evaluation and revision cycle until it accepts an explanation. (See figure 1.) Beliefs in the accepted explanation are added to the system's beliefs, and the explanation is stored in memory to be available for explaining future anomalies.

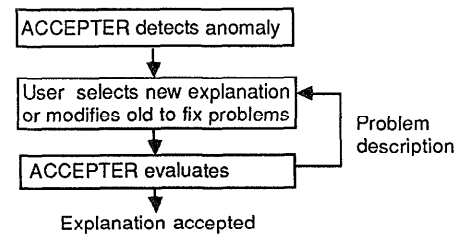

Figure 1: ACCEPTER's evaluation cycle

3.1 ACCEPTER's Evaluation Criteria

3.1.1 Relevance to an anomaly

EBL systems for story understanding explain in order to generate new schemas. In most systems, the explanations generated are completely determined by the event being explained: the reason for explaining does not influence the explainer's focus. But when people explain, they focus on filling gaps in their knowledge: rather than simply asking why an event happened, they try to explain the aspects of the situation that they found anomalous.

For example, people hearing for the first time about a recall of cars would explain different things depending on the circumstances. If the recall is mentioned during a conversation about greedy companies' refusal to accept responsibility for problems after sale, the admission of a defect would be surprising. A useful explanation would reconcile it with old beliefs: perhaps the company thought lawsuits would cost more than the repairs. If the recall is mentioned during a discussion of the excellent quality control of the company, an explanation might address how the defect slipped through the company's checks. In this context, explaining motivation for the recall would not be relevant.

What needs to be explained depends on the understander's *expectation failure* [Schank, 82]; the same event can cause different expectation failures in different contexts. ACCEPTER requires that candidate explanations focus on the features of the situation that conflicted with its expectations. An explanation is relevant to the expectation failure if it identifies the faulty beliefs on which the expectation was based, and shows how revision of those beliefs accounts for the surprising aspects of the situation. By identifying the faulty beliefs underlying the bad expectation, it can correct them and form more accurate expectations in similar future situations. By accounting for the aspects of the situation that were surprising, it can better understand the current case. (For discussion of the need to explain *both* anomalous features of a situation and why the bad expectation was generated, see [Leake, 88].)

3.1.2 Believability of an explanation

ACCEPTER's explanations are instantiated *explanation patterns* (XPs) [Schank, 86], dependency networks tracing how a belief can be inferred from a set of hypotheses. To verify an explanation, the system checks both the plausibility of the hypotheses it involves, and the inferences connecting them. Links between beliefs are checked against

inference rules in ACCEPTER's rule library. When an explanation uses a link that is unknown to the system, the program asks the user to supply a chain of known rules supporting the connection. For links involving known rules, it verifies that stored restrictions on rules' role-fillers are satisfied by the rule's antecedents and consequents.

Although AI systems have often used probabilistic approaches to judge the plausibility of hypotheses (*e.g.*, [Shortliffe, 76]), knowledge of relevant probabilities is unlikely to be available in many real-world situations. [Kahneman et al., 82] demonstrates that rather than using probabilities, people judge plausibility by seeing how well a hypothesis matches common patterns. ACCEPTER uses a similar approach: when a hypothesis matches no existing belief, it is checked against stereotyped knowledge. To control inferencing done during verification, ACCEPTER's consistency checking is highly constrained. Rather than attempting to check all ramifications of a fact, it checks only for discrepancies between the fact and the closest matching structures in memory. Thus verification is strongly memory based: the verification process is the same process used for integration of new facts into memory.

Because the basic understanding process is used to test hypothesized facts, the checks used to fit a fact into a schema must be finer-grained than in most understanders. ACCEPTER uses the algorithm below to integrate facts and hypotheses into memory:

Check whether input fact is already in memory: If the input refers to a state, object, or event that is already known, its features are compared with the features in memory. Any conflicting features are judged anomalous.

If fact is not already known, check whether it satisfies an expectation: ACCEPTER's process for understanding routine facts is modeled on [Cullingford, 78]. Events are understood by fitting them into *Memory Organization Packets* (MOPs) [Schank, 82], which are schemas providing stereotyped expectations to guide understanding. For example, the stereotyped events involved in eating in a restaurant might include first waiting for a table, then sitting down, ordering, receiving food, etc. If an input fact satisfies the expectations provided by an active MOP, it is stored in memory under that MOP ([Schank, 82], [Kolodner, 84], [Lebowitz, 80]), and expectations for the MOP's next scenes are activated. For example, the fact that Swale raced at Belmont places Swale in the racing phase of the MOP M-racehorse-life, and generates the expectation that he will race for a few years, live at the stud farm for a few years, and then die.

When an input only partially matches an expectation, the conflicts are detected as anomalous. For example, when ACCEPTER installs the event of Swale's death in Swale's M-racehorse-life, the death is earlier than predicted by M-racehorse-life, which expects racehorses to die a few years after the end of their racing careers. Consequently, the death is considered anomalous.

If fact was not expected, instantiate a knowledge structure that would have predicted it: When an input fact is irrelevant to active expectations, ACCEPTER attempts to instantiate a new MOP to accept it. For example, when the system begins to process the story of Swale, it places Swale in memory by instantiating the MOP M-racehorse-life with Swale as its actor.

ACCEPTER also accounts for facts in terms of *role themes* [Schank and Abelson, 77]. Role themes represent stereotyped knowledge about the plans and goals associated with actors in certain societal roles. For example, we expect that a policeman will direct traffic, investigate crimes, etc. If a hypothesized action is part of its actor's role theme, the role theme provides confirmation for the action's likelihood. Conflicts are noted as anomalies.

Check whether fact's role-fillers are consistent with normal stereotypes and restrictions: ACCEPTER's MOPs include stereotyped information on common types of role-fillers, and particular role-fillers are checked against those stereotypes. For example, ACCEPTER represents that the filler of the jogger role in M-jogging is usually human. When the system tries to apply the Jim Fixx XP to Swale's death, it detects a problem because horses do not fit the stereotype for joggers. These checks detect problems, but do not give confirmation: although joggers are usually human, the fact that a hypothesized jogger is human does not make his jogging more likely.

Check for predisposing circumstances: Predisposing circumstances can provide partial confirmation of a fact. ACCEPTER's MOPs include information about the circumstances that make them more likely to occur: For example, its MOP M-heart-attack includes the information that high-strung people are likely to have heart attacks. When ACCEPTER knows of features that predispose an actor to fill a particular role, it checks whether the hypothesized role-filler is known to have those features, or if they can be assumed from property inheritance. (To avoid excessive inferencing, it does not try to derive the features from other information.)

Try to connect actions to actor goals: ACCEPTER's approach to ascribing motivations is modeled on PAM. [Wilensky, 78]. Since plan recognition is much more costly than doing the preceding checks, it is only used when they cannot account for the input.

3.1.3 Information given by an explanation

Believable explanations are still unsatisfying if they fail to provide sufficient information. Needs for information depend on the explainer's goals and the plans available to achieve them. For example, when someone without mechanical skills wants to explain a car not starting, he only needs to determine whether the car actually has mechanical problems (*e.g.*, the problem might only be extremely cold weather). If the problem is mechanical, he can pass the problem to a mechanic. But the mechanic needs a more detailed explanation than "mechanical problems," since he needs to identify which part to change or adjust.

ACCEPTER evaluates explanations in light of actors' needs to respond to new situations.[1] The system can now evaluate explanations in terms of standard information required by the veterinarian's or detective's role themes. When a vet explains an animal's death, he looks for a medical cause acceptable for an autopsy report. A detective,

[1] For discussion of evaluation for other purposes, see [Keller, 87] and [Kedar-Cabelli, 87].

whose role theme requires identifying foul play, investigates until the problem is either traced to a criminal plan or to innocent causes. For each role theme, ACCEPTER stores the following characterization of theme-related needs for information:

- A list of types of anomalies whose resolution is important to the theme.
- For each important anomaly, criteria for deciding if an explanation provides adequate information for a standard theme-based response.

Examples of anomalies important to a vet are animals' unexpected changes of health, physical changes (*e.g.*, weight loss) and behavioral changes (such as loss of appetite); they might be signs of a health problem that needs treatment. Anomalies important to a detective include surprising deaths and violent acts; he would trace the cause of a surprising deterioration of health to find whether it was due to natural causes or foul play. For a violent act, he would investigate the actor's motivation to see if the act was unacceptable or justified (*e.g.*, self-defense).

If an anomaly is important to its active theme, ACCEPTER tests the most believable explanation to see if it provides adequate information for a theme-based response. It checks by matching the explanation to a stored template for the needed type of information. This template is an abstract form of XP: its belief-support chain can specify *classes* of nodes and links rather than specific nodes and links. For example, the template for the vet's explanation of changes in health specifies *the explanation must connect a negative health change, via a sequence of any number of* physical-result *links, to a medical cause* (which is restricted to being an instance of disease, trauma, organ-failure, or administering medication).

Matching against the template serves two purposes: it verifies the structure of the explanation's belief-support section, confirming that the XP has the needed causal structure, and binds variables in the template to specific aspects of the explanation that a theme-driven actor needs to know. For example, matching the vet's explanation template to an XP can bind the template's variable cause-of-health-change to a specific disease. Given identification of the disease, the vet could decide on a treatment.

While ACCEPTER's knowledge of theme-related needs for information is pre-compiled, a future goal is to supplement this knowledge with the ability to judge dynamically on the basis of active goals.

3.2 Finding an acceptable explanation

ACCEPTER evaluates candidate explanations until it finds a relevant one with confirmable hypotheses. If it exhausts the candidate explanations before finding one, it accepts the best candidate from the explanations it has tried (provided its hypotheses do not conflict with system beliefs). Ranking of explanations is based on the believability of their weakest hypotheses: an explanation is favored if the likelihood of its weakest hypotheses is greater than that of the weakest hypotheses of competing explanations.[2] If two explanations' weakest hypotheses have equal strength, ACCEPTER favors the explanation with the fewest hypotheses of that strength. If both have the same number, the next-weakest hypotheses of each explanation are compared, until a difference is found at some level of belief strength. (If the comparison reaches previously-believed facts, the program considers the explanations equally likely.) The best explanation is then checked to see if it gives adequate information. If not, ACCEPTER prompts the user for elaboration.

ACCEPTER's emphasis on using patterns to suggest likely hypotheses differs from the approach to choosing between explanations in [Pazzani, 88]. Pazzani's strategies include avoiding explanations that predict events that were not observed, and preferring explanations that account for more of the observed features of the situation. Applying these strategies may require considerable inference, and such strategies also require both that relevant effects be observeable, and that observed features be restricted to relevant effects. Real-world situations often require explaining when effects cannot be verified, and where the set of features to account for is uncertain. For example, if a guest is late, and radio news has reported some drug-related arrests, the delay could be explained by the guest's being arrested or by heavy traffic. Although the arrest accounts for both the news report and the delay, for most guests we would still favor the later explanation.

4 Sample ACCEPTER Output

ACCEPTER starts with a library of nine XPs. It runs on two stories, the death of Swale and the death of basketball star Len Bias. For Swale's death, input is a conceptual representation of:

> Swale was a successful racehorse. Swale won the Belmont Stakes. Swale died a week later.

The early death contradicts expectations for horses' lifespans, so ACCEPTER attempts to explain the death. In the output below, ACCEPTER evaluates the explanation *Swale died because the exertion of racing over-taxed a heart defect* from two perspectives.

A vet's view

```
Checking whether the explanation is relevant to
  [PREMATURE-EVENT
       EXPECTATION-SOURCE - SWALE's RACEHORSE-LIFE
       EARLY-EVENT - SWALE's DEATH]

Confirmed:  It would account for the surprising
aspect of the event.

Checking believability of the explanation.

SWALE'S HORSE-RACE matches previous beliefs.

Although the explanation assumes HEART-ATTACK,
which is unconfirmed, the fact that SWALE has
HIGH EXCITABILITY is a predisposing feature
that supports the assumption.

The explanation assumes the HEART-OF SWALE's role
in HEREDITARY-DEFECTIVE-HEART.
```

[2]Hypotheses' likelihood rating depends on the type of confirmation or problem found when integrating them into memory. ACCEPTER's confirmation classes follow (in order of decreasing confirmation): confirmed by prior beliefs or active expectations; supported by predisposing circumstances; unsupported, but without problems; conflicting with patterns, beliefs, etc. A future goal is to determine a finer-grained ranking.

```
This hypothesis is unsubstantiated but possible.
Believability is ACCEPTABLE.
SWALE'S DEAD HEALTH is important to a vet.
Checking whether the explanation traces
SWALE's DEAD HEALTH to the disease, organ failure or
physical cause responsible.
Explanation hypothesizes the ORGAN-FAILURE:
HEART-ATTACK. It also shows a physical-result
chain between the cause and SWALE's DEAD HEALTH.
Conclusion: explanation is ACCEPTABLE.
```

A detective's view

Since the anomaly is an unexpected death, the explanation is important to a detective. His tests for relevance and believability are the same as the vet's, but he needs different information, as shown by the output below:

```
SWALE'S DEAD HEALTH is important to a detective.
Checking whether the explanation traces
SWALE's DEAD HEALTH to natural causes, to an
accident, or to a crime and suspect.
Explanation hypothesizes the NATURAL-CAUSE:
HEART-ATTACK. It also shows a physical-result
chain between the cause and SWALE's DEAD HEALTH.
Conclusion: explanation is ACCEPTABLE.
```

5 Conclusion

Explanation-based systems rely on having a good explanation of each novel situation they deal with. In most real-world situations, an entire range of explanations can be built for any phenomenon; it is important to know whether a satisfactory explanation has been generated.

This evaluation cannot be done in the abstract: it must be influenced by what the explainer knows and needs to learn. When expectation failures reveal gaps in its knowledge, ACCEPTER augments its knowledge by explaining. It judges relevance of candidate explanations by checking if they address the surprising aspect of the situation. It checks believability based on whether an explanation's hypotheses account for the event in terms of prior beliefs and known patterns. Finally, it evaluates detail in terms of the system's needs for information to deal with the new situation in accordance with particular goals.

References

[Cullingford, 78] Cullingford, R., *Script Application: Computer Understanding of Newspaper Stories,* Ph.D. Thesis, Yale University, 1978. Technical Report 116.

[DeJong and Mooney, 86] DeJong, G., and Mooney, R., *Explanation-Based Learning: An Alternative View,* Machine Learning, 1/1 (1986), pp. 145–176.

[Heider, 58] Heider, F., *Current Theory and Research in Motivation,* Volume XV: *The Psychology of Interpersonal Relations,* John Wiley and Sons, New York, 1958.

[Kahneman et al., 82] Kahneman, D., Slovic, P. and Tversky, A., *Judgement under uncertainty: Heuristics and biases,* Cambridge University Press, 1982.

[Kass et al., 86] Kass, A. M. and Leake, D. B. and Owens, C. C., *SWALE: A Program that Explains,* 1986. In [Schank 86].

[Kedar-Cabelli, 87] Kedar-Cabelli, S.T., Formulating Concepts According to Purpose, *Proceedings of the Sixth Annual National Conference on Artificial Intelligence,* AAAI, Seattle, WA, July 1987, pp. 477–481.

[Keller, 87] Keller, R. M., Defining Operationality for Explanation-Based Learning, *Proceedings of the Sixth Annual National Conference on Artificial Intelligence,* AAAI, Seattle, WA, July 1987, pp. 482–487.

[Kelley, 67] Kelley, H. H., Attribution Theory in Social Psychology, Levine, D. ed., *Nebraska Symposium on Motivation,* University of Nebraska Press, Lincoln, 1967, pages 192–238.

[Kolodner, 84] Kolodner, J.L., *Retrieval and Organizational Strategies in Conceptual Memory,* Lawrence Erlbaum Associates, Hillsdale, N.J., 1984.

[Lalljee and Abelson, 83] Lalljee, M. and Abelson, R., The Organization of Explanations, Hewstone, M. ed., *Attribution Theory: Social and Functional Extensions,* Blackwell, Oxford, 1983.

[Lalljee et al., 82] Lalljee, M., Watson, M. and White, P., *Explanations, Attributions, and the Social Context of Unexpected Behavior,* European Journal of Social Psychology, 12 (1982), pp. 17–29.

[Leake, 88] Leake, D. B., Using Explainer Needs to Judge Operationality, *1988 Spring Symposium Series: Explanation-Based Learning,* AAAI, Stanford, 1988, pp. 148–152.

[Lebowitz, 80] Lebowitz, M., *Generalization and Memory in an Integrated Understanding System,* Ph.D. Thesis, Yale University, October 1980. Technical Report 186.

[Mitchell et al., 86] Mitchell, T.M., Keller, R.M. and Kedar-Cabelli, S.T., *Explanation-Based Generalization: A Unifying View,* Machine Learning, 1/1 (1986), pp. 47–80.

[Mooney and DeJong, 85] Mooney, R. and DeJong, G., Learning Schemata for Natural Language Processing, *Proceedings of the Ninth International Joint Conference on Artificial Intelligence,* IJCAI, Los Angeles, CA, August 1985, pp. 681–687.

[Pazzani, 88] Pazzani, M. J., Selecting the Best Explanation for Explanation-Based Learning, *1988 Spring Symposium Series: Explanation-Based Learning,* AAAI, Stanford, 1988, pp. 165–169.

[Schank and Abelson, 77] Schank, R. C. and Abelson, R., *Scripts, Plans, Goals and Understanding,* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.

[Schank, 82] Schank, R.C., *Dynamic Memory: A Theory of Learning in Computers and People,* Cambridge University Press, 1982.

[Schank, 86] Schank, R.C., *Explanation Patterns: Understanding Mechanically and Creatively,* Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[Shortliffe, 76] Shortliffe, E.H., *Computer-based medical consultations: MYCIN,* American Elsevier, New York, 1976.

[Wilensky, 78] Wilensky, R., *Understanding Goal-Based Stories,* Ph.D. Thesis, Yale University, 1978. Technical Report 140.