Task-Based Criteria for Judging Explanations[1]

David B. Leake
Center for Research on Concepts and Cognition, Indiana University
leake@cogsci.indiana.edu

Abstract

AI research on explanation has not seriously addressed the influence of explainer goals on explanation construction. Likewise, psychological research has tended to assume that people's choice between explanations can be understood without considering the explainer's task. We take the opposite view: that the influence of task is central to judging explanations. Explanations serve a wide range of tasks, each placing distinct requirements on what is needed in an explanation.

We identify eight main classes of reasons for explaining novel events, and show how each one imposes requirements on the information needed from an explanation. These requirements form the basis of dynamic, goal-based explanation evaluation implemented in the program ACCEPTER. We argue that goal-based evaluation of explanations offers three important advantages over static criteria: First, it gives a way for an explainer to know what to elaborate if an explanation is inadequate. Second, it allows cross-contextual use of explanations, by deciding when an explanation built in one context can be applied to another. Finally, it allows explainers to make a principled decision of when to accept incomplete explanations without further elaboration, allowing explainers to conserve processing resources, and also to deal with situations they can only partially explain.

Introduction

It seems obvious that people who explain usually do so for a reason— that explanation is done to serve an overarching task. Explanations allow people to understand unexpected situations, in order to deal with them more effectively. However, the effect of overarching tasks on explanation has received little attention in psychology and artificial intelligence. Attribution theory [Heider, 1958], the central current in psychological study of people's judgement of explanations, tries to account for their judgements in a context-independent way. AI work in explanation-based learning (EBL) also fails to address the influence of changing goals on explanation (see [DeJong, 1988] for an overview of EBL research).

Context-independent theories fail in two main ways. First, they simply cannot account for the choices that people make. Second, in an AI system, they limit system performance by sometimes failing to accept explanations that are useful, sometimes accepting explanations that are not. Context-independent theories often require complete explanations, showing

[1]This work was conducted in part at the Center for Research on Concepts and Cognition at Indiana University, supported by a grant from Indiana University, and at Yale University, supported in part by the Defense Advanced Research Projects Agency, monitored by the Office of Naval Research under contract N0014-85-K-0108 and by the Air Force Office of Scientific Research under contract F49620-88-C-0058.

necessary and sufficient conditions for occurence of the outcome being explained. However, a robbery victim who wants to keep from being robbed again does not need to form a complete picture of the robbery to benefit in the future: as long as he realizes that leaving a window open made the robber choose to rob his house, he can prevent future robberies. Nor is a complete explanation guaranteed to be useful: An account of the robber's motivation, no matter how complete, would not help the victim make his house more secure.

To develop useful criteria for evaluating explanations, we must consider how explanations are used. The following sections identify eight purposes for explanation, and illustrate how each purpose places different requirements on the information that an explanation needs to provide. These requirements in turn determine the goodness of an explanation. We illustrate the effect of purpose on explanation with output from evaluation of an explanation for two different purposes by ACCEPTER, a program that examines the information provided in explanations, according to the information requirements for a range of user-specified tasks.[2] We conclude with a discussion of how task-based evaluation allows explanation-based approaches to be applied despite the lack of a a complete explanation.

Attribution theory

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

The covariation principle makes no reference to how the explanation might be used. More recent work has noted one area in which attributions are influenced by overarching goals: *excuse theory* shows how people displace blame by focusing on external reasons when explaining their own bad performance (see, for example, [Mehlman and Snyder, 1983]). Excuse theory is built on Kelley's model, and shows how the desire to form excuses makes an explainer manipulate the balancing that would otherwise be determined by covariation. However, attribution theory has not addressed the role of other goals in explanation. We will show both that other goals exert strong influences, and that they require characterizing information along dimensions beyond the person–situation distinction that is central to attribution theory.

Explanation-based learning

AI research on explanation-based learning (EBL) has shown that explanations provide valuable guidance for feature selection in new situations. However, EBL research concentrates on how to learn from an explanation presented to the system, giving less attention to

[2] ACCEPTER is a system that detects gaps in its knowledge that require explanation, and that evaluates explanations' plausibility as well as the type of information they provide. Those phases of the system are beyond the scope of this paper, but are described further in [Leake, 1988].

explanation selection. To the extent that it has addressed intended use, it has concentrated on explaining for a single purpose: efficient object recognition (*e.g.*, [Mitchell *et al.*, 1986] and [Kedar-Cabelli, 1987]). Keller [1988] points out that studying EBL only in the context of object recognition has had a strong, and falsely limiting, effect on analysis of explanations: an explanation that permits effective object recognition may not be useful for other tasks.

Some AI systems do apply their explanations to other tasks (*e.g.*, plan repair in [Hammond, 1986]), but rely on being presented with good explanations. This cannot be assumed when using explanations from external sources, or re-using explanations built in other contexts. For example, if we want to find out why our brand X car will not start, in order to repair the problem, we might ask a friend. If he had previously urged us to buy a different brand, he might reply "because brand X is junk." A person would not be satisfied with this answer; nor should an EBL system.

Tasks that drive explanation

In order to understand human explanation, and to build systems that can explain effectively, we must first look at *why* explanation is done. Once we know the purposes for explanation, we can investigate what makes explanations good for those purposes— what information the explanations must provide.

We have identified eight primary tasks served by explanation, each of which imposes different requirements on what constitutes a good explanation. We sketch these tasks below, and describe the requirements they impose. As we describe the requirements, we build up a vocabulary of *evaluation dimensions*, which categorize the aspects of causes that are important to deciding an explanation's usefulness for a given task. We show at the end of this paper how those dimensions are used to implement task-based evaluation in the program ACCEPTER.

We consider that explanations have the form of a belief-support chain [Schank and Leake, 1989]. Belief-support chains are belief dependency networks, tracing inferences that lead from a set of premises to the outcome being explained. The inferences are plausible connections, not deductions: a belief-support chain increases the tendency to believe the outcome, but does not prove that it must occur.

The sections below sketch eight tasks that drive explanation, and examine the types of information they require from an explanation. We concentrate on tasks that arise when people explain surprising events for their own benefit, going beyond simply trying to make sense of those events.

Learning when to predict the event

When an event is surprising, it may be important to learn how *not* to be surprised by it in the future— how to predict it before it happens next time. For example, a college admissions officer might try to explain why a student who had seemed promising had dropped out, to better predict problems when next looking at applications.

Explanations useful for future predictions must have four properties. First, the links from the explanation's premises to the outcome must be strong enough to make the explainer expect the same outcome, the next time the premises recur. The degree to which an explanation licenses future predictions is its *predictive force*.

Second, the premises must be factors that the predictor is likely to know about in the future. For example, suppose a gambler explains a team's surprising loss by their lack of

concentration. Since he is unlikely to know the team's concentration level in advance of future games, the explanation probably will not help him predict future losses in time to profit. However, a coach might be able to tell in advance, from observing the team in the locker room, so he might be able to use the explanation to predict before future games. How easy it is for an actor to recognize that a premise holds is its *knowability*. But the usefulness of knowable causes depends on a both knowability, and a third property, their *timeliness*: they need to happen far enough in advance for the explainer to benefit from revising his predictions. For example, it would not help the gambler to know that the team would lose after he saw the first play, if he had to bet before the game started.

Finally, the explanation can only be used predictively if it shows an unusual factor of the situation. Even if a team's lack of concentration is a contributing factor to its loss, accounting for the loss by bad concentration will not be helpful if the team never concentrates— what the explanation needs to focus on is the unusual aspect of the situation (*e.g.*, that their superstar was injured). Thus an explanation must trace a surprising event to at least one cause with *distinctiveness*.

Controlling future occurrence of the surprising event

Obviously, preventing future occurrence of an event involves finding premises with *causal force*— that cause the outcome— and that the explainer can block (*controlability*). However, this is not sufficient. For example, if someone burns a cake he is baking, he knows the basic cause of the burning— that the cake became too hot— and how to prevent it— turning down the oven. However, he still needs to know *when* to turn down the oven in the future. If he uses a lower temperature for everything, the things that used to come out perfectly will be underdone. Thus for an explanation to help in preventing an outcome, it must show when to apply the preventative steps: it must allow prediction of the bad outcome, early enough to take steps to block it.

To learn how to achieve the outcome that was surprising, an explainer needs to find a set of causes that are all either controllable, or that have *routineness*, so that they are likely to hold in the future, even without the actor taking action to achieve them himself.

Assigning responsibility and blame

If an actor could have prevented an event's occurrence, or contributed voluntarily to its causes, he bears some responsibility. Depending on the *desirability* of the outcome, the actor may be blameable. He can also be blamed, even if he could not predict or control an outcome, if he contributed to it through an undesirable act. For example, we might blame a drug dealer for an addict's death by overdose, even if overdose deaths are relatively unlikely.

Focusing repair of an undesirable current situation

In order to fix a problem, we need to find problems that both have causal force, and *repairability*. We could explain any automobile breakdown by "there's something wrong with the engine," but the only repair possible at that level of detail is to replace the entire engine, which is not within the financial constraints of most drivers. In addition, we need to find a cause with *independence* from prior causes: if a burned-out transformer in a television is caused by a short circuit, replacing only the transformer will not be an effective repair: unless the short circuit is also fixed, a new transformer would immediately burn out also.

Focusing repair is an instance of a more general category of explanation task: clarifying the current situation to choose an appropriate response. The explainer needs to find an explanation that allows selection of a feasible plan, or the choice between competing alternatives. The competing plans determine which distinctions an explanation must make. For example, an insurance company might try to determine whether a death was suicide or not, in order to determine whether to pay the claim, or refuse. If an explanation for the death shows that the victim were killed in the crash of an airliner, the company would not need further information. However, a lawyer for the family might still want to explain the crash further, to determine whether negligence was involved, and sue those who were responsible.

Sketch of additional purposes

We briefly sketch some additional purposes for explanation, that also affect the type of information a good explanation must provide:

- **Learning a new plan, or refining plan selection**: We can sometimes learn new plans by explaining surprising actions. For example, if we ride home with someone who takes an unusual route, we might explain that the route avoids rush hour traffic, and start using it ourselves. This task for explanation is investigated in [Mooney and DeJong, 1985], which argues that explanations for learning new plans must account for actions in terms of known plan schemas.

- **Changing others' view of an outcome**: An explainer might try to focus on causes that absolve an actor of blame (which is closely related to the task studied in excuse theory), or causes that associate an effect with something desirable or undesirable, to make people see the causes in a new light. For example, if someone took a wrong turn, he might try to make his passengers take a more positive view of the incident, by attributing the turn to distraction because conversation in the car was so captivating.

- **Testing or extending a special-purpose theory**: In order to test a theory, we might require that an explanation use a particular class of rules. For example, an economist might explain layoffs to substantiate his economic theory, by showing it would have predicted them.

ACCEPTER

ACCEPTER is a story understanding program that requests explanations when it encounters anomalies, and judges user-selected explanations both in terms of whether they resolve the anomaly, and in terms of whether they provide the information needed for user-selected goals (see [Leake, 1988] for an overview of the system). Thus ACCEPTER makes the judgement needed for an EBL system to assure that it starts from an explanation relevant to its goals. ACCEPTER was developed as part of SWALE (*e.g.*, [Kass *et al.*, 1986]), a system that uses ACCEPTER's judgements to determine which explanations to accept and generalize for future use.

ACCEPTER implements simple heuristics for judging explanations' premises along the dimensions identified above: predictive force, knowability, timeliness, distinctiveness/routineness, desirability, repairability, and independence. These heuristics allow it to evaluate explanations for four of the purposes above: learning to predict the outcome in the future, repairing device defects, preventing recurrence of the outcome, and assigning blame.

The example below shows ACCEPTER's evaluation of two plausible explanations for a hypothetical Audi recall:

1. The mechanical problems resulted from the car being manufactured by Acme Car Company, under contract as a supplier, due to Acme's bad quality control.

2. The defect resulted from a flaw in the transmissions, which aren't checked by Audi's quality control department.

It would be possible for both the explanations to reflect the facts of the recall. However, the following output, in which ACCEPTER evaluates their usefulness for repairing the defect, shows that they are not equally useful to a mechanic. A mechanic needs to find a state that he can repair— causes that happened in the past, and no longer affect the situation, are unimportant to him, because past events can no longer be repaired. Although the first explanation gives information on factors that led to the defect, it doesn't show a continuing cause that can be repaired in the current situation, so that explanation is useless for him:

```
Checking detail for repair.
  To aid in repair, explanation must show a cause that:
     1. Is repairable.
     2. Is predictive of the problem occurring.
     3. Will not be restored by another state if repaired.

      Checking whether some antecedent satisfies the following tests:
      CAUSAL FORCE TEST (does fact cause consequent?),
      REPAIRABILITY, PREDICTIVENESS, and INDEPENDENT CAUSE.

         Applying test for REPAIRABILITY to AUDI'S PRODUCTION-CONTRACT to ACME.

            Searching up abstraction net for pointers to standard repair plans.
            ... test failed.

         Applying test for REPAIRABILITY to ACME'S NOT M-QUALITY-CONTROL.

            Searching up abstraction net for pointers to standard repair plans.
            ... test failed.

         ... Detail is unacceptable.
```

The second explanation involves two factors that continue to contribute to the car's bad condition: the transmission's defect, and the fact that it is part of the car. ACCEPTER finds that a plan exists for correcting one of them, so that a repair can be done:

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

   Checking repairability of features of TRANSMISSION-743'S
   PART-OF-RELATIONSHIP to AUDI'S ENGINE.
```

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

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

... Detail is acceptable.
```

Thus even if both explanations are accurate, an explanation-based system doing repair needs to reject one, and use the other.

Conclusion

Judging explanations according to explainer task provides a way of deciding when to accept one of the many explanations that can be constructed for an event, and finding out what information an incomplete explanation lacks. For example, after a robbery, different tasks make different information important to find out: the victim might focus on what made him a target, to prevent it next time; a policeman might focus on the lack of patrols enabling the crime, to blame his superiors; a social worker rehabilitating the robber might focus on the robber's motivations, to decide how to proceed. The range of purposes and explanations possible for a single event has been discussed in philosophical works such as [Hanson, 1961], but has not been addressed in research on EBL or attribution.[3]

In addition to helping to choose between complete explanations, task-based criteria allow an explainer to use partial explanations in a principled way. Even if the explanation does not completely account for an outcome, it can be useful. For example, if we know just one of the factors that contributed to an event, we may be able to block its occurrence: people who know that high-fat diets contribute to heart attacks can lower their risk, even if they do not know all the other factors necessary to predict heart-attacks. However, traditional approaches to EBL require that learning start from a complete explanation, giving necessary and sufficient conditions for an event, which will be impossible to generate in many situations.

Goal-based focusing is needed because of the complexity of real-world situations. No real-world explanation can include *all* the causally-relevant factors in an event, so that explanations necessarily highlight a few causes out of many. If those causes are irrelevant to the explainer's goals, the explanation will be useless. In order for explanation-based processing to be effective in complex situations, AI systems need to be able to identify which of the many causes of an event are important to their goals, and require that those causes be highlighted in the explanations they use.

[3][Souther *et al.*, 1989] presents an argument close in spirit to ours— that it is essential to be able to generate explanations from a given viewpoint— and identifies classes of explanations that students might seek when studying college-level botony. However, since they discuss only tutoring applications, they do not connect their classes to over-arching goals that make them important, beyond simply doing well in a course.

References

[DeJong, 1988] G. DeJong. An introduction to explanation-based learning. In H.E. Shrobe, editor, *Exploring Artificial Intelligence: Survey Talks from the National Conferences on Artificial Intelligence*. Morgan Kaufmann, Palo Alto, 1988.

[Hammond, 1986] K.J. Hammond. *Case-based Planning: An Integrated Theory of Planning, Learning and Memory*. PhD thesis, Yale University, 1986. Technical Report 488.

[Hanson, 1961] N. Hanson. *Patterns of Discovery*. Cambridge University Press, Cambridge, 1961.

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

[Kass *et al.*, 1986] A. M. Kass, D. B. Leake, and C. C. Owens. Swale: A program that explains. In *Explanation Patterns: Understanding Mechanically and Creatively*, pages 232–254. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[Kedar-Cabelli, 1987] S.T. Kedar-Cabelli. Formulating concepts according to purpose. In *Proceedings of the Sixth Annual National Conference on Artificial Intelligence*, pages 477–481, Seattle, WA, July 1987. AAAI.

[Keller, 1988] R. Keller. Operationality and generality in explanation-based learning: Separate dimensions or opposite endpoints? In *Proceedings of the 1988 AAAI Spring Symposium on Explanation-based Learning*. AAAI, 1988.

[Kelley, 1967] H. H. Kelley. Attribution theory in social psychology. In D. Levine, editor, *Nebraska Symposium on Motivation*, pages 192–238. University of Nebraska Press, Lincoln, 1967.

[Leake, 1988] D.B. Leake. Evaluating explanations. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 251–255, Minneapolis, MN, August 1988. AAAI, Morgan Kaufman Publishers, Inc.

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

[Mitchell *et al.*, 1986] T.M. Mitchell, R.M. Keller, and S.T. Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine Learning*, 1(1):47–80, 1986.

[Mooney and DeJong, 1985] R. Mooney and G. DeJong. Learning schemata for natural language processing. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 681–687, Los Angeles, CA, August 1985. IJCAI.

[Schank and Leake, 1989] R.C. Schank and D.B. Leake. Creativity and learning in a case-based explainer. *Artificial Intelligence*, (40), 1989.

[Souther *et al.*, 1989] A. Souther, L. Acker, J. Lester, and B. Porter. Using view types to generate explanations in intelligent tutoring systems. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, pages 123–130, Ann Arbor, MI, August 1989. Cognitive Science Society.