Proceedings of the 1994 AAAI Spring Symposium on Goal-Driven Learning, pp. 72-79.

Issues in Goal-Driven Explanation

David B. Leake
Computer Science Department
Indiana University
Bloomington, IN 47405, U.S.A.
leake@cs.indiana.edu

Abstract

When a reasoner explains surprising events for its internal use, a key motivation for explaining is to perform learning that will facilitate the achievement of its goals. Human explainers use a range of strategies to build explanations, including both internal reasoning and external information search, and goal-based considerations have a profound effect on their choices of when and how to pursue explanations. However, standard AI models of explanation rely on goal-neutral use of a single fixed strategy—generally backwards chaining—to build their explanations. This paper argues that explanation should be modeled as a goal-driven learning process for gathering and transforming information, and discusses the issues involved in developing an active multi-strategy process for goal-driven explanation.

Introduction

Many AI systems construct explanations for internal use to further their processing. The resulting explanations, which often have the form of deductive proofs (e.g., [Mitchell *et al.*, 1986]) or plausible reasoning chains (e.g., [Leake, 1992; Pearl, 1988; Schank, 1986]) tracing dependency relationships between beliefs, are an important resource for determining the goal-relevant features of a situation. For example, medical diagnosis involves explaining how observed symptoms arise from an underlying disease, and the resultant explanations provide the information needed to accomplish the goal of treating the illness.

Internal explanations have been used to further many different classes of goals in AI systems, such as establishing coherence and formulating accurate predictions during story understanding [Leake, 1992; Mooney, 1990; Ram, 1993; Schank, 1986; Wilensky, 1983]; identifying relevant features for forming useful generalizations [DeJong, 1986; Mitchell *et al.*, 1986]; focusing plan repair [Birnbaum *et al.*, 1990; Hammond, 1989]; guiding selection of learning methods in multistrategy learning [Hunter, 1990b; Hunter, 1990a; Ram and Cox, 1994]; and indexing learned information in useful ways [Barletta and Mark, 1988; Hammond, 1989; Leake and Owens, 1986; Ram, 1993; Schank, 1982]. Unfortunately, effectively building the explanations required by these models is a difficult problem. In order to generate any explanations at all, it is necessary to overcome the problems of intractable domain theories and incomplete knowledge [Mitchell *et al.*, 1986; Leake, 1992]; in order to generate *useful* explanations, it is necessary in addition to assure that the resultant explanations actually benefit the explainer [Keller, 1988; Leake, 1991; Leake, 1992; Minton, 1988].

Given an event, situation, or concept to explain, current AI explanation systems generally use a fixed goal-neutral explanation-generation strategy to generate their explanations: backwards chaining from the information that must be explained [Hobbs *et al.*, 1990; Kautz and Allen, 1986; Mitchell *et al.*, 1986]. As the explanation process proceeds or after it is completed, candidate explanations are evaluated for usefulness [Keller, 1988; Minton, 1988; Mitchell *et al.*, 1986; Thagard, 1989], but goal-based considerations do not directly guide incremental decisions of how to proceed in explanation construction. As a result, considerable effort may be expended pursuing one course of explanation, only to have it revealed in the evaluation phase that the results are irrelevant or unhelpful to achieving system goals [Leake, 1992; Minton, 1988; O'Rorke, 1989].

This paper proposes an alternative framework for constructing explanations. In that framework, the explanation construction process is viewed as a form of *goal-driven learning* [desJardins, 1992; Leake and Ram, 1993; Ram and Leake, In press]. Goals provide direction for all phases of the process: The explainer decides whether to explain and how to pursue explanations based on reasoning about its information needs, its strategies for gathering the information needed to build explanations, and current circumstances.

In order to accurately reflect explainer goals, the pursuit of explanations must be dynamically shaped by on-going interaction between circumstances in the external world and the explainer's internal state (current explainer knowledge, information needs, and active goals). Incremental information obtained by the explanation process must be able to affect explainer goals, and changing knowledge and goals must in turn be able to change the strategies (including internal reasoning and external action) used to obtain the information needed when building explanations. Changes in goals to perform tasks in the world may also change explainer *knowledge goals* [Leake and Ram, 1993; Ram, 1987; Ram, 1990; Ram and Hunter, 1992], redirecting the focus of explanation, prompting attempts to explain additional aspects of the situation, or causing current explanation efforts to be abandoned. To emphasize the interaction between explanatory reasoning, other goal-driven internal processing, and action in the external world, in the remainder the remainder of this paper our model will be called *goal-driven interactive explanation* (GDIE).

As an illustration of the need to reason about goals when explaining everyday events, consider how a human explainer might attempt to explain an automobile break-down. This explanation process is likely to be motivated by a set of goals such as reaching a particular destination on time and preserving a valuable piece of property—the explainer's car. These two goals may prompt another goal: to find how to repair the car.

The goal to repair the car prompts new knowledge goals to determine the repair that is needed. Those goals my prompt either reasoning (e.g., using reasoning from a theory of engine function to hypothesize an explanation based on symptoms observed before the break-down), or action to gather information (e.g., looking under the hood for anomalous conditions). Which strategies to apply will depend on the goals and circumstances surrounding explanation—on factors such as the expected utility of pursuing particular types of information [desJardins, 1992; Pryor and Collins, 1992], given current goals and the state of current hypotheses, and on environmental factors such as the availability of help, how unpleasant the weather is for getting out of the car to look under the hood, whether the explainer is dressed in formal clothes, and so forth.

The human behavior described in the previous paragraph differs from AI models in which explanation is done by pure reasoning using a single strategy, but it is consistent with a top-down model of goal-driven explanation in which a fixed goal determines what to explain [Krulwich *et al.*, 1990; Leake, 1992]. However, purely top-down models fail to capture another important aspect of human explanation. During the explanation generation effort, incremental information may change a human explainer's plans and goals, redirecting explanation as new learning goals arise. If someone rushing to the airport begins to suspect, based on a partial explanation, that the car problem is serious (even if it has not been completely diagnosed), the attempt to generate a complete explanation of the breakdown may be abandoned in favor of taking a taxi instead of repairing the car and driving. This change gives rise to a new goal—getting a taxi. Achieving that goal may depend on additional explanations (e.g., explaining why no taxis have been seen on the street where the explainer is) that may require action in the world to generate (e.g., going to a policeman and asking). Thus the human explanation process involves strategic use of multiple information-seeking strategies, using reasoning and action that interacts with changing circumstances and goals.

This paper argues on both cognitive and functional grounds that explanation cannot be modeled as an isolated reasoning process or as being exclusively a tool to serve fixed higher-level goals. Instead, GDIE models explanation as shaped by the interaction of top-down and bottom-up influences from the current environment, system task goals in the world, and system learning goals. In viewing the explanation process as interacting with its environment, GDIE is a model of "situated explanation," in that it adheres to the general precept of theories of situated activity (e.g., [Suchman, 1987]) that interactions with the environment shape action. However, rather than being behavior-based, as is generally the case in models of situated activity, the GDIE framework is strongly goal-based and strategic.

This paper begins by discussing and motivating the GDIE framework as it applies to four forms of learning: gathering new information in the world, transforming prior knowledge through inference, re-organizing knowledge by establishing new relationships between beliefs in memory, and augmenting and re-organizing a memory of cases including explanations, plans for verifying explanations, and plans for repairing problems identified by explanations. The paper closes with the a discussion of ramifications of GDIE in a broader context of goal-driven learning and a sketch of research in progress on developing the GDIE framework in context of a computer model.

The Need for GDIE

The key questions for guiding explanation are when to explain, what to explain, and how to focus explanation generation [Leake, 1992]. In most artificial intelligence views of explanation, explanations are treated as dependency chains that derive the conclusion to be explained from known information. Although the interpretation of the chains varies, with explanations sometimes considered to be deductive proofs [Mitchell *et al.*, 1986; Mooney, 1990] and sometimes plausible reasoning chains [Leake, 1992; Pearl, 1988; Schank, 1986], all these approaches must face the problem of how the needed chains are generated. Current explanation systems use a single fixed strategy for explanation; the most prevalent approach is to use theorem-proving techniques (e.g., [Hobbs *et al.*, 1990; Kautz and Allen, 1986; Mitchell *et al.*, 1986]), but case-based approaches have also been proposed (see [Schank *et al.*, 1994] for an overview).

In standard approaches, the explanation construction process does not interact with either the external world or with the process that requests an explanation. The external world and the knowledge and goals of the explainer are treated as fixed from the time that explanation construction starts, so that the task of explanation construction is to use existing knowledge to build the needed explanation according to fixed goals. Such models reflect an idealized conception of explanation as a purely rational act—one that follows Sherlock Holmes' directive that "the art of the reasoner should be used rather for the sifting of details than for the acquiring of fresh evidence" [Doyle, 1976, p. 1]. Accordingly, although the explanation component may be called upon by an overarching system (e.g., a planner) to build particular types of explanations in order to advance the overarching system's goals, the explanation construction component is isolated from system goals in two respects. First, goals that are furnished to the explanation component are used to evaluate candidate explanations, rather than to guide ongoing construction of those candidates. Second, information obtained during the on-going explanation process does not influence explainer goals until explanation construction has been successfully completed. In particular, explanation construction cannot call upon the overarching system to perform action to gather needed information; as a result, the explainer is limited to the information provided to it at the start of the explanation process. Thus the explanation construction process is isolated from its environment. This traditional model of explanation will be referred to as the *isolated explanation* model.

Goal-driven interactive explanation differs from isolated explanation in making explanation construction part of a unified framework allowing interactions between goals, goal-driven actions, and the explainer's environment. Reasoning and action are integrated by guiding explanation according to knowledge goals that can be planned for not only by internal inferential actions [Michalski, 1993], but also by physical actions in the world. As the following sections show, this increases the ability to generate explanations and facilitates effective use of system effort.

The need for goal-driven action supporting explanation: When real-world explainers start their explanation effort, they often lack the information that they need to complete it successfully. In those cases, explaining purely by chaining from prior knowledge—the standard method in AI explanation systems—is impossible. Consequently, regardless of dreams of building explanations by the pure force of reasoning as proposed by Sherlock Holmes, the robustness of an explanation mechanism depends on it being able to use multiple strategies to gather the information it needs, including strategies that pursue information by taking action in the external world. In fact, the previous edict by Sherlock Holmes in favor of pure reasoning occurs in a context that emphasizes the need for action and belies Holmes' edict: the statement is made as Sherlock Holmes is traveling to the scene of the crime to investigate for himself.

In addition, even if an explainer *could* eventually reach the needed conclusion by reasoning in isolation, explaining in isolation might still not be worthwhile. The most effective means for drawing a conclusion may be to verify it in the world, rather than to infer it from what is already known. For example, an engineer designing an airplane wing may have all information needed to calculate its expected performance, but a wind tunnel test may still be the most practical means for determining its characteristics—or for identifying unexpected aspects of its behavior on which to focus reasoning.

The need for interaction between knowledge goals, new information, and task goals: In traditional explanation systems, incremental information gathered during explanation is not available to overarching processes until a complete explanation has been constructed. However, an explainer's effort can be directed more effectively if the overarching system monitors the incremental results of explanation construction and revises goals accordingly (in turn redirecting the explanation process as needed). For example, attempts to generate one explanation may result in opportunities that should be addressed concurrently with the initial explanation effort. If a car owner is disassembling the car's engine in order to explain a strange noise, the engine disassembly process may give rise to opportunities to easily explain other problems that were not worthwhile to explain before (e.g., why the car is burning oil).

Likewise, the information available from a partial explanation may cause a prior goal to be abandoned, causing the attempt to explain to be abandoned as well. If a car owner attempts to explain why a car burns oil in order to repair it, and partial explanation reveals other severe problems before the oil problem is diagnosed, the owner may abandon the repair goal and generate a new goal—selling the car—before completing the diagnosis of why the car is burning oil. That new goal may make it unnecessary to complete the current explanation (e.g., if the car burning oil is not expected to affect the sale price), and may make it necessary to explain other problems (e.g., problems that did not bother the original owner but that the owner believes would dissuade potential buyers), in order to repair those problems.

Issues in Building a Model of Goal-Driven Interactive Explanation

A model of goal-driven interactive explanation must be able to reflect the effects of learning goals on explanation construction and the effects of ongoing explanation construction on learning goals. In particular, an architecture for a GDIE system must enable the following behaviors:

1. **Goal-based focusing for explanation:** The cost and value of pursuing a particular explanation may change markedly as a result of changes in the explainer's knowledge, capabilities and environment. Consequently, the explanation construction process must be able to switch its attention between candidate explanations as circumstances change. In addition, the process determining which explanations to focus on must consider not only the plausibility of competing candidates, but also the utility of pursuing a particular candidate given factors such as current knowledge goals, their relative priorities, and environmental circumstances affecting the difficulty of pursuing the candidate explanation.

   Maintaining a set of alternative explanations and reasoning about the relative utility of pursuing the competing alternatives changes the explanation process from focusing on single explanations to a process more akin to differential diagnosis [Harvey and Bordley, 1970], in which competing hypotheses are simultaneously considered and weighed against one another. Although differential diagnosis is a common approach in medical expert systems (e.g., [Clancey and Shortliffe, 1984]), it is rarely applied to other applications of abductive reasoning such as abductive understanding.

   Using a differential diagnosis process makes it important to be able to determine the utility of learning information relevant to the current *set* of candidate explanations, rather than to a single candidate. For example, tests that simultaneously provide information about multiple explanations may be more worthwhile than less expensive tests that are only relevant to a single explanation.

2. **Interleaving explanation with action:** Rather than waiting for generation of a complete explanation, as in the isolated explanation model, GDIE allows any system component to begin to use the information from partial explanations and tentative hypotheses as soon as that information is useful. For example, when responding to car problems it is often appropriate to act based on a tentative hypothesis that can later be revised. If a car fails to start on a cold day, it is reasonable to attribute the problem to the cold as an initial hypothesis and to respond to it by keeping the car in the garage on cold nights, but to store other alternative hypotheses (e.g., a weak battery) to be considered if information relevant to the choice between them becomes available opportunistically [Ram, 1993].

3. **Revising system goals based on incremental information:** New information obtained during explanation—either from the explanation effort itself, or from other sources—should be able to change system goals. For example, a driver whose car broke down while the driver was rushing to the airport may initially seek an explanation that will enable him or her to repair the problem. If the explainer only diagnoses the problem as occurring somewhere within the transmission, the explanation is insufficient for the original purpose—unless the explainer happens to have a spare transmission, it is too vague to direct a repair. However, the information is sufficient to revise system goals: Because transmission repairs are generally lengthy, executing any transmission repair is incompatible with the higher-level goal of reaching the airport in time. Based on the new information, the repair goal and explanation towards that goal should simply be abandoned.

   Likewise, if the driver is in the midst of attempting to explain a breakdown while stranded on a deserted road, and a taxi miraculously passes, the repair goal and explanation effort should be abandoned: flagging down the taxi is more important. As described previously, as goals change, the changes should in turn change the course of explanation.

In order to build a model that supports the dynamic focusing of explanation effort, the interleaving of explanation with action, and revision of system goals as hypotheses are formed or new information is opportunistically obtained during explanation, content theories—theories of the *content* of the knowledge required for particular tasks—are needed for the following four types of knowledge:

1. **Types of goal-based needs for information:** In current systems that learn in response to information needs, the primary motivation for learning is the detection of processing failures or expectation failures (e.g., [Birnbaum *et al.*, 1990; Hammond, 1989; Leake, 1992; Ram, 1993; Ram and Cox, 1994; Riesbeck, 1981; Schank, 1986]). However, information needs may arise without explicit failures. Many tasks can drive explanation, and performing those tasks may require gathering information through explanation, without being prompted by a failure *per se*. For example, the policy to improve performance at a task may cause attempts at learning even if current performance matches expectations and is satisfactory.

   In order to direct explanation towards fulfilling the information needs of a planning system, a model of goal-driven interactive explanation must be based on a theory of the information that a planner needs to carry out its planning and pursue its plans. This in turn depends on categorizing the classes of information needs that are generated by particular tasks (already done for a number of tasks in [Leake, 1991]), and developing mechanisms for identifying each type of need when it arises.

2. **Strategies for determining how far to explain:** It is possible to explain any event at an arbitrarily high level of detail and to trace the chain of its causes arbitrarily far back from the event itself. For example, a car breakdown could be explained by "a mechanical failure", "a transmission failure", "a failure of part X within the transmission," etc.; the failure of particular part X might be explained by the soft metal from which it was made, which might in turn be explained by the transmission designer's decision to economize in order to keep the cost of the transmission below that of a competitor's transmission, which might be explained by management decisions about competition strategies, and so on. Consequently, principles must be formulated to decide how far to trace an explanation, and the level of detail to use in the explanation, given current knowledge goals and other circumstances. For example, a reasoner might have rules that consider current tasks in order to decide whether an explanation is necessary, such as *when time is limited, explain only as much as is necessary to resume execution of the plan whose failure triggered the explanation effort.* Other rules might give guidance based on partial explanations, as in *if a candidate explanation involves a device failure, and the device is used in a current plan or pending plan, explain until the failure is confirmed or ruled out.*

3. **Classes of explanation methods:** Once a reasoner has decided the information to search for, it must decide how to search. Because a GDIE system can generate goals to be satisfied by any of the components of the overarching system, it can either conduct internal reasoning, as done in the traditional isolated explanation model, or can take action in the world to gather needed information. Choosing which learning strategies to apply requires reasoning about the types of knowledge goals driving the system, and the functionality of different types of strategies for performing the needed learning [Ram and Cox, 1994]; it also requires reasoning about resources required to perform particular information-seeking actions, the circumstances under which they apply, and their tradeoffs [Hunter, 1990b].

   As a concrete example of the type of reasoning that must be performed, an automobile mechanic must be able to reason about the tradeoffs between different means of determining whether a particular part is defective, such as attempting to remember and compare the symptoms of previous failures, disassembling the part, swapping the suspect part with a replacement and observing the effects of the swap, etc. In general, which tests are appropriate will depend strongly on both the current external environment and system capabilities and resources (e.g., the time and tools available to the mechanic and the mechanic's competing obligations). Thus in combination with knowledge of the functionality of different types of learning strategies, utility-based considerations should play a role in deciding what should be learned and which learning methods to apply [desJardins, 1992; Gratch and DeJong, 1993; Pryor and Collins, 1992].

4. **Ways new information can impinge on current plans and goals:** In order for a GDIE system to monitor new information and re-adjust its goals accordingly, its processing must be based on a theory of the ways in which new information can relate to prior plans and goals and of how new information is analyzed to reveal threats and opportunities that prompt the generation of new goals.

Relationship to Previous Work

The GDIE process shares elements with a number of other approaches in addition to those referenced in the previous sections. The need to request additional information during explanation has long been recognized in diagnostic expert systems, as have the benefits of a differential diagnosis process for choosing between of competing alternatives (see [Clancey and Shortliffe, 1984] for a sampling of some of these approaches). However, such models do not include the introspective reasoning about needs for information that is fundamental to goal-driven learning.

The interaction of task goals and explanation has been stressed by [Leake, 1991; Leake, 1992; Rymon, 1993], and the role of interactive dialog in learning has been investigated by Redmond (1992). Research on expert system explanation has also considered the role of dialog in generating useful explanations for human questioners. For example, Moore and Swartout (1989) argue that expert system explanation should be *reactive* in the sense of being considered as an on-going dialogue with the user, with previous user questions determining how later questions are interpreted. However, a GDIE system corresponds both to the explainer and the *user* of a dialogue system in that it generates its own requests for explanation, generates its own responses, and evaluates the suitability of its answers in order to refine their goals and to generate new knowledge goals as needed.

Krulwich et al. (1990) suggest a model of goal-driven focusing of hierarchical explanation construction, but with internal reasoning as the only means of building explanations. AQUA [Ram, 1991; Ram, 1993] takes an active approach to the information-gathering for explanation by varying the depth of its reading of input stories in order to actively seek information reflecting its knowledge goals, but GDIE goes one step further in taking a broader view of the possible interactions between the information-seeking process and the environment. The active approach to verifying explanations in the environment is related to existing models of explanation that use experimentation to choose between candidate explanations [Rajamoney, 1993], but with a crucial difference: in GDIE the action taken is part of the explanation construction process itself, rather than only part of the selection process choosing between the candidate explanations that have already been built.

GDIE in a Broader Context

GDIE is a goal-driven learning process that applies at least four types of learning strategies, according to reasoning about its needs: Learning by gathering information, learning by transforming existing knowledge (e.g., during the inference process used to elaborate explanations), learning by forming new connections between beliefs, and learning by storing and re-indexing the explanations and information-gathering plans that it builds. In principle, the GDIE process should be integrated with other goal-driven learning processes to provide it with access to the learning strategies of the overarching system. For example, inductive generalization might be used to hypothesize new causal rules to be used by the explanation component. Conversely, because explanations are required as the starting point for other learning methods such as explanation-based generalization [Mitchell *et al.*, 1986; DeJong and Mooney, 1986] and explanation-based indexing [Barletta and Mark, 1988; Leake and Owens, 1986], GDIE should also be viewed as a new tool to be added to goal-driven learners' repertoires of methods for forming the explanations required to support desired learning.

In the GDIE model, the choice between types of learning is dictated by mediating between the factors that shape any goal-driven learning process: task goals, knowledge goals, and the environment in which processing occurs. Consequently, the GDIE framework has ramifications for any goal-driven learning process: the theory of how to perform this mediation, as well as the mechanisms needed to enable this task, apply to any goal-driven learning process.

Towards a Computer Model

An application of GDIE to explaining vehicle malfunctions is being implemented at Indiana University by Raja Sooriamurthi. Because a central tenet of GDIE is that explanation interacts with goals beyond explaining for its own sake, the program models explanation construction from the perspective of a driver who explains and repairs vehicle malfunctions to further the goal of reaching a destination under time and resource constraints. Because GDIE integrates reasoning and action, the explanation effort takes place in a dynamic simulated environment reflecting different possible circumstances under which a stranded motorist can attempt to explain and repair vehicle malfunctions.

Each phase of the explanation effort reflects the interaction between task goals, knowledge goals, and explanation construction methods. When initially faced with a vehicle malfunction, the "driver" must first determine whether to explain at all. This depends on system goals, the state of the changing simulated environment, the state of system knowledge, and applicable methods for explaining. For example, if the driver is under extreme time pressure and a taxi becomes available, explaining the problem may not be worthwhile unless very inexpensive explanation methods are available (e.g., asking the mechanic from a tow truck for an educated guess about the nature of the problem).

In general, many alternative explanation construction methods may apply to generating initial hypotheses, such as asking for or looking up possible explanations, using case-based reasoning, or using backwards chaining from the symptoms to be explained). Applying each of the methods may require action in the world (e.g., to seek help or to form a more precise characterization of the symptoms to explain). After the driver has formed an initial explanation, it will in general be necessary both to verify the explanation and to elaborate it until it is sufficiently detailed to suggest a response. Verification and elaboration in turn may require additional reasoning, such as additional chaining or retrieval of cases, or additional information-gathering (e.g., by memory search, knowledge transformation, or by search in the world).

A central issue in developing the system is how to select and manage goal-driven explanation methods and how to allow them to interact with overarching goals. During explanation construction, decisions about how to pursue explanations must reflect estimates of the feasibility and costs of different strategies and their expected usefulness.[1] For example, if an explanation localizes a problem as being internal to the engine, and no tools are available for disassembling the engine, it is not worthwhile attempting to elaborate that explanation in order to determine the specific part that is defective. This reasoning interacts with overarching goals: if it is decided that all possibly relevant repairs would take too long, no attempt is made to determine which repair should be done—the goal of repairing the problem is simply abandoned.

System learning will include not only gathering needed information in the world (e.g., about whether the spark plugs are clean), but storing explanations, verification and repair plans, and the results of reasoning about whether to pursue particular explanations, as well as transforming and reorganizing existing knowledge to facilitate its application in the future.

The framework being developed reflects some key differences between situated explanation construction and isolated explanation construction. First, the type of information initially provided to the system—e.g.,

[1] In the system under development, all these estimates will be done by case-based reasoning.

that the car will not start—is not detailed enough to build an explanation to guide a repair. Consequently, an isolated explanation system could not reliably explain the problem: building a useful explanation requires taking action to gather additional information. Second, the cost of explaining is important. Verifying a dead battery is less expensive than verifying a loose wire deep inside the engine—all that testing the battery requires is to turn on the radio—so unless the loose wire is considered a far more probable alternative, the dead battery explanation is reasonable to test first. Third, the explainer's capabilities for using explanations affect how far to pursue explanations and whether to pursue them at all. For example, if the system had no plans for repairing spark plug problems, the spark plug explanation would not be worth pursuing, and taking a taxi might immediately be a better option than attempting to explain. Fourth, the choice of which explanation to pursue changes as information changes during the explanation process. If investigation of one explanation reveals information that changes the promise of another (e.g., if a loose wire is noticed while checking a spark plug), the explanation effort must be re-focused.

Conclusions

AI models of explanation construction generally reflect two assumptions: that at the start of the explanation process, the explainer will have all the information it needs to generate candidate explanations, and that the explanation construction itself can be modeled as pure reasoning using a single fixed method. This paper has argued that explanation should instead be viewed as a goal-driven learning process in which learning goals and task goals interact and in which learning goals are pursued by applying multiple methods to gather, transform, and re-organize information.

In a complex world, explanation generation must be aimed towards satisfying current learning goals. As the explanation process attempts to build explanations satisfying those learning goals, it can itself give rise both to new task goals in the world and to learning goals, requiring satisfying those goals by learning processes involving multiple strategies such as internal inference, memory search, and active information search in the external world. Thus explanation can both serve existing learning goals and generate new learning goals to be pursued through additional goal-driven learning. This paper has described the central issues involved in developing a framework for modeling *goal-driven interactive explanation* (GDIE). The process is shaped by the environment in which explanation is conducted, including both the explainer's internal goals and external environmental factors. In GDIE, explanation and action are combined in a unified goal-driven process of information search, and the reasoning process that traditionally constitutes the entire explanation mechanism is only one of a range of information-seeking actions that can be pursed according to reasoning about the knowledge that is needed and how to acquire that knowledge.

Compared to the traditional *isolated explanation* model, goal-driven interactive explanation significantly facilitates explanation in complex and imperfectly understood domains. First, it increases the range of phenomena that can be explained, by including external information-seeking actions in the explanation process. Second, by making explanation effort sensitive to utility-based considerations, it increases the effectiveness of explanation construction. Third, by reflecting the interplay between new information, environmental factors, and current goals, it increases the usefulness of the resulting explanations by dynamically shifting focus in response to changing needs for information.

In addition, the GDIE framework has wider ramifications for goal-driven learning as a whole. Fundamental GDIE issues include how to reason about three questions: the information a reasoner needs, how to satisfy those needs, and how to recognize and respond appropriately to changes in current circumstances and the learner's knowledge. Each of these is central to managing the learning process in any goal-driven learning system. Consequently, goal-driven interactive explanation is a rich research area not only because of its ramifications for explanation, but for its ramifications for goal-driven learning in other contexts.

References

Barletta, R. and Mark, W. 1988. Explanation-based indexing of cases. In Kolodner, J., editor 1988, *Proceedings of a Workshop on Case-Based Reasoning*, Palo Alto. DARPA, Morgan Kaufmann, Inc. 50–60.

Birnbaum, L.; Collins, G.; Freed, M.; and Krulwich, B. 1990. Model-based diagnosis of planning failures. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, Boston, MA. AAAI. 318–323.

Clancey, W. and Shortliffe, E., editors 1984. *Readings in Medical Artificial Intelligence: The First Decade.* Addison-Wesley, Reading, MA.

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

DeJong, G. 1986. An approach to learning from observation. In Michalski, R.; Carbonell, J.; and Mitchell, T., editors 1986, *Machine Learning: An Artificial Intelligence Approach*, volume 2. Morgan Kaufmann Publishers, Cambridge, MA. 571–590.

desJardins, M. 1992. Goal-directed learning: A decision-theoretic model for deciding what to learn next. In *Proceedings of the Machine Discovery Workshop, Ninth International Machine Learning Conference*, San Mateo. Morgan Kaufmann.

Doyle, A. Conan 1976. The adventure of Silver Blaze. In *Memoirs of Sherlock Holmes.* Schocken Books, New York. chapter 13, 1–16.

Gratch, J. and DeJong, G. 1993. Assessing the value of information to guide learning systems. In *Proceedings of the*

ML-93 Workshop on Knowledge Compilation and Speedup Learning, Amherst, MA. 65–71.

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

Harvey, A. and Bordley, J. 1970. *Differential Diagnosis; The interpretation of clinical evidence.* Saunders, Philadelphis.

Hobbs, J.; Stickel, M.; Appelt, D.; and Martin, P. 1990. Interpretation as abduction. Technical Report 499, SRI International.

Hunter, L. 1990a. Knowledge acquisition planning for inference from large datasets. In Shriver, B.D., editor 1990a, *Proceedings of the Twenty Third Annual Hawaii International Conference on System Sciences*, Kona, HI. 35–45.

Hunter, L. 1990b. Planning to learn. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, Cambridge, MA. Cognitive Science Society. 261–268.

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

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

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

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

Leake, D. and Ram, A. 1993. Goal-driven learning: Fundamental issues (a symposium report). *The AI Magazine* 14(4):67–72.

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

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

Michalski, R. 1993. Inferential theory of learning as a conceptual basis for multistrategy learning. *Machine Learning* 11(2 & 3):111–151.

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

Mitchell, T.; Keller, R.; and 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.

Moore, J. and Swartout, W. 1989. A reactive approach to explanation. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI. IJCAI. 1504–1510.

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

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

Pryor, L. and Collins, G. 1992. Planning to perceive. In *Proceedings of the AAAI Spring Symposium on Selective Perception*, Stanford, CA. AAAI.

Rajamoney, S. 1993. Designing experiments to extend the domain theory. In DeJong, G., editor 1993, *Investigating Explanation-Based Learning.* Kluwer. chapter 5, 166–189.

Ram, A. and Cox, M. 1994. Introspective reasoning using meta-explanations for multistrategy learning. In Michalski, R. and Tecuci, G., editors 1994, *Machine Learning: A Multistrategy Approach.* Morgan Kaufmann. In Press.

Ram, A. and Hunter, L. 1992. The use of explicit goals for knowledge to guide inference and learning. *Applied Intelligence* 2(1):47–73.

Ram, A. and Leake, D. press. Learning, goals, and learning goals. In Ram, A. and Leake, D., editors press, *Goal-Driven Learning.* MIT Press.

Ram, Ashwin 1987. Aqua: Asking questions and understanding answers. In *Proceedings of the Sixth Annual National Conference on Artificial Intelligence*, Seattle, WA. American Association for Artificial Intelligence, Morgan Kaufmann Publishers, Inc. 312–316.

Ram, A. 1990. Knowledge goals: A theory of interestingness. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, Cambridge, MA. Cognitive Science Society. 206–214.

Ram, A. 1991. A theory of questions and question asking. *The Journal of the Learning Sciences* 1(3 & 4):273–318.

Ram, A. 1993. Indexing, elaboration and refinement: Incremental learning of explanatory cases. *Machine Learning* 10(3):201–248.

Redmond, M. 1992. *Learning by Observing and Understanding Expert Problem Solving.* Ph.D. Dissertation, College of Computing, Georgia Institute of Technology. Technical report GIT-CC-92/43.

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

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

Schank, R.C.; Riesbeck, C.; and Kass, A., editors 1994. *Inside Case-Based Explanation.* Lawrence Erlbaum Associates, Hillsdale New Jersey.

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

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

Suchman, L. 1987. *Plans and Situated Actions: the Problem of Human-Machine Communication.* Cambridge University Press, New York.

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

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