Decision Models: A Theory of Volitional Explanation*

Ashwin Ram

Georgia Institute of Technology
School of Information and Computer Science
Atlanta, Georgia 30332-0280
(404) 853-9372
E-mail: ashwin@ics.gatech.edu

Abstract

This paper presents a theory of motivational analysis, the construction of volitional explanations to describe the planning behavior of agents. We discuss both the content of such explanations, as well as the process by which an understander builds the explanations. Explanations are constructed from *decision models*, which describe the planning process that an agent goes through when considering whether to perform an action. Decision models are represented as *explanation patterns*, which are standard patterns of causality based on previous experiences of the understander. We discuss the nature of explanation patterns, their use in representing decision models, and the process by which they are retrieved, used and evaluated.

1 Issues in explanation

In order to learn from experience, a reasoner must be able to *explain* what it does not understand. When a novel or poorly understood situation is processed, it is interpreted in terms of knowledge structures already in memory. As long as these structures provide expectations that allow the reasoner to function effectively in the new situation, there is no problem. However, if these expectations fail, the reasoner is faced with an *anomaly*. The world is different from its expectations. In order to learn from this experience, the reasoner needs to know *why* it made those predictions. It also needs to explain *why* the failure occurred, i.e., to identify the knowledge structures that gave rise to the faulty expectations, and to understand why its domain model was violated in this situation. Finally, it must store the new experience in memory for future use. Explanation is a central issue in this process of understanding and learning.

The construction of explanations is also known as *abduction*, or *inference to the best explanation*. This process is usually viewed as the chaining together of causal inference rules in order to create a causal chain, in which a proposed set of premises is shown to be causally responsible for the event or fact being explained. However, there are two problems with this view.

The first problem is the familiar one of combinatorial explosion of inferences. Most explanation programs create explanations by chaining together inference rules that describe the causality of the domain. For example, PAM [Wilensky, 1978] used a set of planning rules connecting together typical goals and plans of people, and chained them together to form motivational explanations for actions observed in a story. However, this process is very inefficient in complex domains, where the causal chains may be several steps long.

The second problem is the evaluation of explanations. Since the chaining process is seeking a connection between two concepts, most theories use an evaluation criterion based on the structural properties of this connection. For example, marker passing and spreading activation techniques (which are often proposed as a solution to the combinatorial explosion problem) usually judge the goodness of an explanation by the length of the causal chain. The shortest correct explanation is assumed to be the "best" one. However, the definition of "best" is dependent on the *goals* of the reasoner in forming the explanation and not just on the length or correctness of the causal chain underlying the explanation. In situations where there is no one "right" explanation, the "best" explanation must be more than a causal chain that describes the events in the domain; it must also address the reason that an explanation was required in the first place. This in turn determines what the reasoner can learn from the explanation.

*The research described was conducted while the author was at Yale University, and supported by the Defense Advanced Research Projects Agency and the Office of Naval Research under contract N00014-85-K-0108, and by the Air Force Office of Scientific Research under contracts F49620-88-C-0058 and AFOSR-85-0343.

In addition to processing issues, a theory of explanation must also address the content issues of the nature and representation of explanations. What is an explanation, and what kinds of knowledge does it provide? What is the nature of the causal knowledge that underlies volitional explanations? The answers to these questions depend both on the explanations that we desire to build, as well as on the process that is used to build them.

This paper presents a theory of explanation based on the claim that new explanations are built, not by chaining inference rules together, but rather by reusing explanations that have been encountered in previous situations and are already known to the system [Schank, 1986]. Our view raises several questions:

- *Content and representation:* What kinds of knowledge must an explanation provide? How do we represent this knowledge? What kinds of structures are used to represent explanations in memory? What is the vocabulary out of which these structures are built?
- *Retrieval:* How do we find pre-stored explanations in memory without having to try each one?
- *Evaluation:* How do we determine what kind of explanation is needed, and which explanation is the "best" one in a particular situation?
- *Learning:* How are explanations learned so that they can be reused in the future? What happens when pre-stored explanations don't apply to the current situation?[1]

The theory presented here has been implemented in the AQUA program, a story understanding program which learns about terrorism by reading newspaper stories about unusual terrorist incidents in the Middle East. We will illustrate our ideas with examples taken from this program. Further details may be found in [Ram, 1987; Schank and Ram, 1988; Ram, 1989].

2 What is an explanation?

The need for an explanation arises when some observed fact doesn't quite fit into the reasoner's world model, i.e., the reasoner detects an *anomaly*. An explanation is a knowledge structure that makes the anomaly go away. To illustrate the nature of such a structure, let us consider some candidate explanations for the following story (New York Times, Nov 27, 1985, page A9) from the domain of the AQUA program:

S-1: Suicide bomber strikes Israeli post in Lebanon.

SIDON, Lebanon, November 26 — A teenage girl exploded a car bomb at a joint post of Israeli troops and pro-Israeli militiamen in southern Lebanon today, killing herself and causing a number of casualties, Lebanese security sources said.

...

A statement by the pro-Syrian Arab Baath Part named the bomber as Hamida Mustafa al-Taher, born in Syria in 1968. The statement said she had detonated a car rigged with 660 points of explosives in a military base for 50 South Lebanon Army men and Israeli intelligence and their vehicles.

Why did Hamida go on the bombing mission?

(1) Because Lebanon is a Middle Eastern country.

(2) To destroy the Israeli military base.

(3) Because she was a religious fanatic.

(4) Because she didn't realize she was going to die during the mission.

Consider (1). This does not seem like an explanation for S-1. The reason isn't that (1) is false, but rather that there seems to be no causal connection between (1) and S-1. Thus it is not sufficient for a proposed explanation to be true; *an explanation must be causally connected to the anomaly.* It must contain a set of premises and a causal chain linking those premises to the anomalous proposition. If the reasoner believes the premises, the proposition ceases to be anomalous since the causal interactions underlying the situation can now be understood.

However, not all causal structures are explanations. For example, (2) is causally relevant to S-1, but it still doesn't feel like an explanation. To understand why, let us make the anomaly in S-1 explicit. The real question isn't "Why did Hamida go on the bombing mission?", but rather one of the following:

S-2: Why was Hamida willing to sacrifice her life in order to destroy the Israeli military base?

S-3: Why did Hamida go on a mission that would result in her own death?

The reason that explanation (2) feels strange is that it misses the point of the question. If the point is made explicit as in S-2, (3) is a possible explanation for the anomaly. Alternatively, if the real question is intended to be S-3, (4) is a possible explanation. The point is that, in order to qualify as an explanation, a causal description must address the underlying anomaly.

To state this another way, *an explanation must address the failure of the reasoner to model the situation correctly.* In addition to resolving the incorrect predictions, it must also point to the erroneous aspect of the chain of reasoning that led to the incorrect predictions. An explanation is *useful* if it allows the reasoner to learn and to improve its performance at its task; the claim here is that *an explanation must be both causal and relevant in order to be useful.* This is important in evaluating explanations to determine the best one for a particular situation.

[1] These issues are beyond the scope of this paper.

3 Explanation patterns

An explanation is a causal chain that demonstrates why the anomalous proposition might have occurred by introducing a set of premises that causally lead up to that proposition. There may be more than one explanation for a situation, depending on the question that the reasoner is interested in. For example, if the system needs to explain the motivations behind the girl's actions in story S-1, it may build what we think of as the *religious fanatic explanation*: The girl was a Moslem fanatic; she was so determined to further the cause of her religion that she was willing to die for it; and she believed that destroying the military base would help her religious cause.

The premise of this explanation is that the girl was a religious fanatic. If the reasoner believes or can verify the premises of an explanation, the conclusion is said to be explained. Explanations are often verbalized using their premises. Thus in normal conversation this explanation would be stated succinctly as "Because she was a religious fanatic." However, the real explanation includes the premises, the causal chain, and any intermediate assertions (such as the girl's belief that the bombing would help her religious cause) that are part of the causal chain.

How might a reasoner construct such an explanation? PAM [Wilensky, 1978] used a set of planning rules connecting typical goals and plans of people, and chained them together to form explanations such as the above. However, this is too inefficient in complicated situations, where the causal chains could be several steps long. To get around this problem, AQUA uses pre-stored explanations for stereotypical situations. These explanations represent standard patterns that are observed in these situations, and hence are called *explanation patterns* [Schank, 1986].

An explanation pattern (XP) is a stock explanation for a stereotypical situation. For example, *religious fanatic does terrorist act* is a standard XP many people have about the Middle East terrorism problem. One might think of them as the "scripts" of the explanation domain.[2] When a reasoner encounters a situation for which it has a canned XP, it tries to apply the XP to avoid detailed analysis of the situation from scratch.

This approach is known as *case-based explanation*, since previous cases or explanations known to the reasoner are used to help in the construction of new explanations. Explanatory cases in AQUA are based on the theory of explanation patterns described by [Schank, 1986], to which we add a theory of the representational structure and content of the XPs used in story understanding.

Explanations can be divided into two broad categories, physical and volitional.

3.1 Physical explanations

Physical explanations link events with the states that result from them, and further events that they enable, using causal chains similar to those of [Rieger, 1975] and [Schank and Abelson, 1977]. Physical explanations answer questions about the physical causality of the domain. For example, if the system had never read a story about a car bombing before, it might encounter an anomaly: "How can a car be used to blow up a building?" The answer to this question is a physical explanation:

(1) A car is a physical object.

(2) A car can contain explosives.

(3) A car can be propelled by driving it.

(4) Explosives can be blown up by the sudden impact of a car colliding with a building.

(5) A building can be blown up by blowing up explosives in its immediate vicinity.

Thus the explanation is that the bomber drove an explosive-laden car into the building, the impact caused the explosives to detonate, which caused the building to blow up.

3.2 Volitional explanations

Volitional explanations link actions that people perform to their goals and beliefs, yielding an understanding of the *motivations* of the characters. For example, the system might detect a different anomaly on reading story S-1, such as "Why would someone commit suicide if they are not depressed?" An explanation for this question, such as the religious fanatic explanation, must provide a motivational analysis of the reasons for committing suicide. For this reason, volitional explanations are also called motivational explanations. Although the basic structure of volitional explanations is the same as that of physical explanations, the vocabulary used to represent the causal chain is very different.

Volitional explanations fall into two broad categories:

1. **Abstract explanation patterns** for why people do things. These are standard high-level explanations for actions, such as "Actor does action because the outcome of action satisfies a goal of the actor."

2. **Stereotypical explanation patterns.** These are specific explanations for particular situation, such as "Shiite Moslem religious fanatic goes on suicide bombing mission."

[2] Unlike scripts, however, XPs are flexible since they contain a description of the *causality* underlying a situation in addition to a description of the situation itself. This allows XPs to be useful in novel situations, while retaining the advantages of pre-stored structures in stereotypical situations. The incremental elaboration of XPs in novel situations is discussed in [Ram, 1989; Ram, 1990b].

For example, an explanation of type 1 for the suicide bombing story could be "Because she wanted to destroy the Israeli base more than she wanted to stay alive." An explanation of type 2 would be simply "Because she was a religious fanatic." The internal causal structure of the latter explanation could then be elaborated to provide a detailed motivational analysis in terms of explanations of the first type if necessary.

Volitional explanations thus correspond to the filling out of the "belief-goal-plan-action" chain [Schank and Abelson, 1977; Wilks, 1977; Wilensky, 1978; Schank, 1986], although we need to expand the vocabulary of this chain in order to model such explanations adequately [Ram, 1989]. A volitional explanation relates the actions in which the characters in the story are involved to the *outcomes* that those actions had for them, the *goals*, *beliefs*, *emotional states* and *social states* of the characters as well as priorities or *orderings* among the goals, and the *decision process* that the characters go through in *considering* their goals, goal-orderings and likely outcomes of the actions before deciding whether to do those actions. A detailed volitional explanation involving the planning decisions of a character is called a *decision model*, and is illustrated in figure 1.

Decision models provide a theory of motivational coherence for stories involving volitional agents. When a decision model is applied to the actions of a given character in a story, it focusses attention on faulty assumptions or inconsistencies identified in the application of the decision model to the story. These inconsistencies signal anomalies, which must be explained by determining whether different parts of the decision model (e.g., the goals of the agent, his beliefs about the outcome, or his volition in deciding to perform the action) are actually present as assumed.

For example, the religious fanatic explanation is based on the following decision model:[3]

1. Explains: Why `volitional-agent` A did a `suicide-bombing` M, with `results` =
   - (1) `death-state` of A
   - (2) `destroyed-state` of `target`, a `physical-object` whose `owner` is an opponent religious group.
2. Premises:
   - (1) A believes in the `religion` R.
   - (2) A is a `religious-fanatic`, i.e., A has `high-religious-zeal`.
3. Internals:
   - (1) A is religious and believes in the religion R (an `emotional-state`, perhaps caused by a `social-state`, such as upbringing).
   - (2) A is strongly zealous about R (an `emotional-state`).
   - (3) A wants to spread his religion R (a `goal`, initiated by (1) and (2)).
   - (4) A places a high priority on his goal in (3), and is willing to sacrifice other goals which we would normally place above the religion goal (a `goal-ordering`, initiated by (1) and (2)).
   - (5) A believes that performing a suicide bombing against opponent religious groups will help him achieve his goal in (3) (a belief or `expected-outcome`).
   - (6) A knows that the performance of a suicide bombing may result in a negative outcome for him (an `expected-outcome`).
   - (7) A weighs his goals (3), goal-orderings (4), and likely outcomes (5) and (6) (a `consideration`).
   - (8) A decides to do the suicide bombing M (a `decision`, based on the considerations in (7)).
   - (9) A does the suicide bombing M (an action or `mop`, whose `actor` is A).
   - (10) The suicide bombing has some outcome for A, which is either positive or negative as viewed from the point of view of A's goals and goal-orderings (a `self-outcome`).

[3]`Typewriter font` represents actual vocabulary items used by the AQUA program. Further details of the representation may be found in [Ram, 1989].

The representation of the religious fanatic explanation is shown in figure 2. The decision model has the following components:

The outcome of an action: Every action results in some set of states that may or may not be beneficial to the people involved in that action, depending on their goals at that time. The outcome of an action, therefore, must be modelled *from the point of view of a particular volitional agent* involved in that action. The most common volitional participants are `actor` and `planner`, but any role involving a volitional agent must potentially be explained.

The decision process: Every agent involved in an action makes a *decision* about whether to participate in that particular volitional role (`actor`, `planner`, `object`, etc.) in the action. Such decisions represent the *planning* process that the agent underwent prior to the action. A complete model of this process requires a sophisticated vocabulary of goals, goal interactions, and plans, such as that of [Wilensky, 1983] or [Hammond, 1986]. There are three basic kinds of decisions:

1. **Choice:** The agent *chooses* to participate or not to participate in a given volitional role in some action. The explanation must describe why he made this choice.
2. **Agency:** The agent is *induced* to participate or not to participate in a given volitional role in an action. This is similar to the previous case in that the agent "enters" the action of his own volition. The difference is that here the agent is acting under the agency of another agent. Thus the

```
           isa                          isa
STATE  <------  MENTAL-STATE  <-------  CONSIDER
                                        . GOALS
                                        . GOAL-ORDERINGS
                                        . EXPECTED-OUTCOME
                                        ----------------------
                                           ||
                                           || mentally-enables
                                           ||
           isa                          isa    vv         isa
PROCESS  <-----  MENTAL-PROCESS <-------  DECISION  <------  DECISION-TO-ENTER
                                          - CHOOSES          DECISION-NOT-TO-ENTER
                                            FORCED
                                            INDUCED
                                          ----------
MOP                                           ||
-----                                         ||
||                          RELATION          || mentally-results
|| results                  --------          ||
||                             -              ||
||                             | isa          ||
vv      isa                    |              vv
STATE  <------  OUTCOME     VOLITIONAL-ROLE-RELATION
COLLECTION      -------     ------------------------
                   -           -          /          \
                   |           |       domain    co-domain
                   |           |          |          |
               co-domain    domain        v          v
        isa        \          /          MOP     VOLITIONAL-AGENT
RELATION <--------  SELF-OUTCOME
                    ------------
                    |   |   |
                    |   |   POS-SELF-OUTCOME
               isa  |   |
                    |   MIXED-SELF-OUTCOME
                    |
                    NEG-SELF-OUTCOME
```

Figure 1: The structure of volitional explanations. A `volitional-agent` participates in some `volitional-role` in a `mop`, which then `results` in an `outcome` (a collection of `states`). Prior to this, the `volitional-agent` undergoes a `decision` process in which he considers his `goals`, `goal-orderings` and `expected-outcome`, which then `mentally-results` in the `volitional-role-relation` being considered being true (in) or false (out) depending on the outcome of the decision.

---

explainer must be able to model inter-agent interactions [Schank and Abelson, 1977; Wilensky, 1983; Ram, 1984].

3. **Coercion:** The agent is *forced* to participate or not to participate in a given volitional role in an action. This case arises when an agent is physically coerced into participation or non-participation.

Considerations in decisions: The system also needs to reason about what an agent was considering as he made a particular decision. Considerations model the goals and beliefs of an agent, along with orderings among these goals and expected outcome of the action being considered. Considerations are composed of three constituents: (1) `goals` considered by the agent while deciding whether or not to participate in an action, (2) `goal-orderings`, the agent's prioritization of these goals, and (3) the `expected-outcome`: the agent's beliefs about what the outcome of the action is likely to be. This is represented by the `consider` node in figure 1.

Each of these constituents may itself need to be explained further. For example, the system might question the social or mental (e.g., emotional) states that `initiated` a particular goal or goal-ordering in an agent, or how a particular belief about the outcome of an action came about. Explanations, therefore, may need to be *elaborated* according to the demands of the story and the goals of the system.

4 Structure of explanation patterns

AQUA has several XPs indexed in memory, representing its causal knowledge of the terrorism domain. These XPs are represented as graph structures (as illustrated above) with four main components:

1. **PRE-XP-NODES:** Nodes that represent what is known before the XP is applied. One of these nodes, the EXPLAINS node, represents the particular action being explained.
2. **XP-ASSERTED-NODES:** Nodes asserted by the XP as the explanation for the EXPLAINS node. These comprise the premises of the explanation.
3. **INTERNAL-XP-NODES:** Internal nodes as-

```
                                   isa
          EMOTIONAL-STATE  <-------  HIGH-RELIGIOUS-ZEAL
                  |                    object A            RELIGION R
                  | isa              -------------------   ----------
                  |                                    ||      ||
                  |                                    initiates
       isa        v              isa                   ||      ||
STATE <------ MENTAL-STATE <------- CONSIDER          vv      vv
                                    . goals :         GOAL(A,GS)
                                    .                     |
                                    .                     | goal-ordering
                                    .                     v
                                    .                 NOT(GOAL(A,BS))
                                    .
                                    . KNOWS-RESULT
                                    ------------------------------
                                      ||
                                      || mentally-enables
                                      ||
MENTAL-   isa             isa         vv
PROCESS <----- DECISION <------- CHOOSES-TO-ENTER
                                 ----------------
                                      ||
                                      || mentally-results
                                      ||
          isa                domain   vv         isa
RELATION <----- SELF-OUTCOME ---------> ACTOR ------> VOLITIONAL-
                ------------            -----         ROLE-RELATION
                     |                  /   \
                 co-domain          domain co-domain
                     |                /       \
                  OUTCOME            M         A
                  -------            |         |
                  /  |  \      SUICIDE-BOMBING  RELIGIOUS-FANATIC
             agent  mop  members       / results
               /     |     \          /
              A      M    (GS and BS)
              |      |     |          |               isa
RELIGIOUS-FANATIC  SUICIDE- |      DEATH-STATE  ------>  STATE
                   BOMBING  |       . object A
                            |                  isa
                       DESTROYED-STATE  ------>  STATE
                       . object TARGET
```

Figure 2: The religious fanatic explanation pattern. A is the agent, R his religion, M the action he chooses to do, and GS and BS the good and bad outcomes for A as a result of doing that action. A volitionally chooses to perform M knowing both outcomes, the death-state of A and the destroyed-state of the target.

serted by the XP in order to link the XP-ASSERTED-NODES to the EXPLAINS node.

4. **LINKS:** Causal links asserted by the XP. These taken together with the INTERNAL-XP-NODES are also called the internals of the XP.

An explanation pattern states that the XP-ASSERTED-NODES lead to the EXPLAINS node (which is part of a particular configuration of PRE-XP-NODES) via a set of INTERNAL-XP-NODES, the nodes being causally linked together via the LINKS (which in turn could invoke further XPs). In other words, an XP represents a causal chain composed of a set of nodes connected together using a set of LINKS (causal rules or XPs). The "antecedent" (or premise) of this causal chain is the set of XP-ASSERTED-NODES, the "internal nodes" of the causal chain are the INTERNAL-XP-NODES of the XP, and the "consequent" is the EXPLAINS node. The difference between XP-ASSERTED-NODES and INTERNAL-XP-NODES is that the former are merely asserted by the XP without further explanation, whereas the latter have causal antecedents within the XP itself.

5 The explanation cycle

An explanation-based understander must be able to detect anomalies in the input, and resolve them by building motivational and causal explanations for the events in the story in order to understand why the characters acted as they did, or why certain events occurred or did not occur. This process characterizes both "story understanders" that try to achieve a deep understanding of the stories that they read, as well as programs that need to understand their domains in service of other problem-solving tasks. Explanations are constructed by retrieving XPs from memory, applying them to the situation at hand, and verifying or evaluating the resulting hypotheses.

5.1 Anomaly detection

Anomaly detection refers to the process of identifying an unusual fact that needs explanation. The anomalous fact may be unusual in the sense that it violates or contradicts some piece of information in memory. Alternatively, the fact may be unusual because, while there is no explicit contradiction, the reasoner fails to integrate the fact satisfactorily in its memory.

5.2 Explanation pattern retrieval

When faced with an anomalous situation, the reasoner tries to retrieve one or more explanation patterns that would explain the situation. Ideally, an XP should be indexed in memory such that it is retrieved only in those situations in which it is applicable. But this is impossible in practice. For example, consider the applicability conditions for "blackmail." In general, blackmail is a possible explanation whenever "someone does something he doesn't want to do because not doing it results in something worse for him." But trying to show this in general is very hard. Thus, in addition to general applicability conditions, a reasoner must learn specific, sometimes superficial, features that suggest possibly relevant XPs even though they may not completely determine the applicability of the XP to the situation. For example, a classic blackmail situation is one where a rich businessman who is cheating on his wife is blackmailed for money using the threat of exposure. If one read about a rich businessman who suddenly began to withdraw large sums of money from his bank account, one would expect to think of the possibility of blackmail. However, one does not normally think of blackmail when one reads a story about suicide bombing, although theoretically it is a possible explanation.

AQUA indexes motivational XPs in memory using typical contexts in which the XPs might be encountered (*situation indices*), as well as character stereotypes representing typical categories of people to whom the XPs might be applicable (*stereotype indices*) [Ram, 1989]. The third type of index is known as the *anomaly index* or *category index*. Recall that in addition to explaining the occurrence of the event, it is important for the XP to address the anomaly which arose from the failure of the reasoner to model the situation correctly. Thus the type of the anomaly provides an index to the type of XP required to build an explanation. For example, if the anomaly was one where an actor performed an action that violated one of the actor's own goals, the reasoner might look for a "goal sacrifice" XP (such as a religious fanatic sacrificing her life for the cause of her religion), or an "actor didn't know outcome" XP (such as a gullible teenager not realizing what the outcome of her action was going to be). However, the category of goal sacrifice XPs would be inappropriate for an anomaly in which the actor failed to perform an action which only had a good outcome for the actor; in this case, a "missed opportunity" XP might be chosen.

5.3 Explanation pattern application

Once a set of potentially applicable XPs is retrieved, the reasoner tries to use them to resolve the anomaly. This involves instantiating the XPs, filling in the details through elaboration and specification, and checking the validity of the final explanations. An XP is instantiated by unifying the EXPLAINS node of the XP with the description of the situation being explained, and instantiating the INTERNAL-XP-NODES and LINKS. If all the PRE-XP-NODES and INTERNAL-XP-NODES of the XP fit the situation, the hypothesis is applicable. If the unification fails, the hypothesis is rejected.[4]

5.4 Hypothesis verification and evaluation

The final step in the explanation process is the confirmation or refutation of possible explanations, or, if there is more than one hypothesis, discrimination between the alternatives. A hypothesis is a causal graph that connects the premises of the explanation to the conclusions via a set of intermediate assertions. At the end of this step, the reasoner is left with one or more alternative hypotheses. Partially confirmed hypotheses are maintained in a data dependency network called a *hypothesis tree*, along with questions (unconfirmed XP-ASSERTED-NODES) representing what is required to verify these hypotheses.

There are five criteria for evaluating the goodness of an explanation:

1. **Believability:** Does the system believe the XP from which the hypothesis was derived? This is not an issue when all XPs in memory are believed, but for a program that learns new XPs, some of which may be incomplete, the believability of the XP is an important criterion in deciding whether to believe the resulting hypothesis.
2. **Applicability:** How well does the XP apply to this situation? Did it fit the situation without any modifications?
3. **Relevance:** Does the XP address the underlying anomaly? Does it address the knowledge goals of the reasoner (i.e., does it allow the reasoner to learn)?
4. **Verification:** How definitely was the explanation confirmed or refuted?
5. **Specificity:** How specific is the XP? Is it abstract and very general (e.g., a proverb), or is it detailed and specific?

Intuitively, a "good" explanation is not necessarily one that can be proven to be "true" (criterion

[4]There is also the possibility of modifying the hypothesis to fit the situation [Schank, 1986; Kass *et al.*, 1986].

4), but also one that seems plausible (**1** and **2**), fits the situation well (**2** and **5**), and is relevant to the goals of the reasoner (criterion **3**).

The relevance criterion is important if the explanation is created for some purpose (and not as an end in itself). The fact that the reasoner encountered an anomaly indicates a need to learn, which could arise in several ways. The reasoner may not have the knowledge structures to deal with a novel situation, or the knowledge structures that the reasoner applies to the situation may be incomplete or incorrect. The domain knowledge may be mis-indexed in memory, i.e., the reasoner may have the knowledge structures to deal with the situation, but it may be unable to retrieve them since they are not indexed under the cues that the situation provides.

When an explanation is built, the reasoner needs to be able to identify the kind of processing error that occurred and invoke the appropriate learning strategy. For example, if an incomplete knowledge structure is applied to a situation, the resulting processing error represents both the knowledge that is missing, as well as the fact that this piece of knowledge, when it comes in, should be used to fill in the gap in the original knowledge structure. Similarly, if an error arose due to a mis-indexed knowledge structure, the explanation, when available, should be used to re-index the knowledge structure appropriately. The explanation is therefore constrained by the needs of the learning process [Ram, 1990a].

6 Conclusion

Abduction, or inference to the best explanation, is a central component of the reasoning process. Abduction is viewed, not as a process of chaining together inference rules to produce causal chains, but rather one of case-based reasoning from pre-stored causal chains, known as explanation patterns, associated with prior experiences in memory. This provides a way to control the combinatorial explosion of inferences, but introduces a new set of issues: the content and representation of explanation patterns, the types of indices used to retrieve XPs from memory, the evaluation of candidate hypotheses, and the learning of new XPs.

Evaluation is facilitated by using anomaly characterizations as retrieval indices for XPs. The "best" explanation is not one that is the most "correct," if correctness is even measurable in the domain of interest, but one that is most useful to the process that is seeking the explanation. The anomaly detection process provides retrieval cues that are used to find explanation patterns that are likely to be relevant to the anomaly.

These ideas have been explored in the AQUA program, a computer model of the theory of question-driven understanding. AQUA learns about terrorism by reading newspaper stories about terrorist incidents in the Middle East. The requirements of this task provided constraints on the theory of explanation presented here.

References

[Hammond, 1986] K. J. Hammond. *Case-Based Planning: An Integrated Theory of Planning, Learning and Memory.* Ph.D. thesis, Yale University, Department of Computer Science, New Haven, CT, October 1986. Research Report #488.

[Kass *et al.*, 1986] A. Kass, D. Leake, and C. Owens. *SWALE: A Program That Explains*, pages 232–254. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[Ram, 1984] A. Ram. Modelling Characters and their Decisions: A Theory of Compliance Decisions. Master's thesis, University of Illinois at Urbana-Champaign, Urbana, IL, August 1984. Technical Report T-145.

[Ram, 1987] A. Ram. AQUA: Asking Questions and Understanding Answers. In *Proceedings of the Sixth Annual National Conference on Artificial Intelligence*, pages 312–316, Seattle, WA, July 1987. American Association for Artificial Intelligence, Morgan Kaufman Publishers, Inc.

[Ram, 1989] A. Ram. *Question-driven understanding: An integrated theory of story understanding, memory and learning.* Ph.D. thesis, Yale University, New Haven, CT, May 1989. Research Report #710.

[Ram, 1990a] A. Ram. Goal-Based Explanation. In *Proceedings of the AAAI Spring Symposium on Automated Abduction*, Palo Alto, CA, March 1990.

[Ram, 1990b] A. Ram. Incremental Learning of Explanation Patterns and their Indices. In *Proceedings of the Seventh International Conference on Machine Learning*, Austin, TX, June 1990.

[Rieger, 1975] C. Rieger. Conceptual Memory and Inference. In R. C. Schank, editor, *Conceptual Information Processing.* North-Holland, Amsterdam, 1975.

[Schank and Abelson, 1977] R. C. Schank and R. Abelson. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.

[Schank and Ram, 1988] R. C. Schank and A. Ram. Question-driven Parsing: A New Approach to Natural Language Understanding. *Journal of Japanese Society for Artificial Intelligence*, 3(3):260–270, May 1988.

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

[Wilensky, 1978] R. Wilensky. *Understanding Goal-Based Stories.* Ph.D. thesis, Yale University, Department of Computer Science, New Haven, CT, 1978.

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

[Wilks, 1977] Y. Wilks. What Sort of Taxonomy of Causation Do We Need for Language Understanding. *Cognitive Science*, 1:235, 1977.