April 1988 UILU-ENG-88-2218

COORDINATED SCIENCE LABORATORY
College of Engineering

ACTIVE EXPLANATION REDUCTION: An Approach to the Multiple Explanations Problem

Shankar A. Rajamoney
Gerald F. DeJong

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS None
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) UILU-ENG-88-2218	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Coordinated Science Lab University of Illinois	6b. OFFICE SYMBOL (If applicable) N/A
6c. ADDRESS (City, State, and ZIP Code) 1101 W. Springfield Avenue Urbana, IL 61801	7a. NAME OF MONITORING ORGANIZATION Office of Naval Research
8a. NAME OF FUNDING / SPONSORING ORGANIZATION ONR	8b. OFFICE SYMBOL (If applicable)
8c. ADDRESS (City, State, and ZIP Code) 800 N. Quincy St., Arlington, VA 22217	7b. ADDRESS (City, State, and ZIP Code) 800 N. Quincy St., Arlington, VA 22217
	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N-00014-86-K-0309
	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

ACTIVE EXPLANATION REDUCTION: An Approach to the Multiple Explanations Problem

12. PERSONAL AUTHOR(S)

Rajamoney, Shankar A. and DeJong, Gerald F

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM _____ TO _____	April 1988	14

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Machine Learning, Explanation-Based Learning Imperfect Theory Problems, Learning by Experimentation

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The *multiple explanations problem* is central to explanation-based learning from imperfect theories. In this paper, we present a new approach called *active explanation reduction* to deal with this problem. Active explanation reduction involves the purposeful alteration of the world to generate new information. This new information will cause some of the explanations to become inconsistent with reality, thereby eliminating them from further consideration. Active explanation reduction may also be viewed as *experiment design*. This paper presents a theory of experiment design which is based on the principle of *refutation*. The theory describes three strategies for designing experiments - *elaboration*, *discrimination* and *transformation*. The theory and an experiment engine - an implementation of the theory - are illustrated using a detailed example which involves constructing explanations from intractable theories. The relation of the multiple explanations problem to the imperfect theory problems is also described. Finally, active explanation reduction is evaluated based on four criteria - completeness, efficiency, tolerance of unavailable data and feasibility.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code) / 22c. OFFICE SYMBOL

Technical Report

UILU-ENG-88-2218

ACTIVE EXPLANATION REDUCTION:

An Approach to the Multiple Explanations Problem*

Shankar A. Rajamoney†
Gerald F. DeJong

Artificial Intelligence Research Group
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
1101 West Springfield Avenue
Urbana, IL 61801

April 1988

ABSTRACT

This paper also appears in the *Proceedings of the Fifth International Conference on Machine Learning*, Ann Arbor, Michigan, June 1988.

The *multiple explanations problem* is central to explanation-based learning from imperfect theories. In this paper, we present a new approach called *active explanation reduction* to deal with this problem. Active explanation reduction involves the purposeful alteration of the world to generate new information. This new information will cause some of the explanations to become inconsistent with reality, thereby eliminating them from further consideration. Active explanation reduction may also be viewed as *experiment design*. This paper presents a theory of experiment design which is based on the principle of *refutation*. The theory describes three strategies for designing experiments - *elaboration*, *discrimination* and *transformation*. The theory and an experiment engine - an implementation of the theory - are illustrated using a detailed example which involves constructing explanations from intractable theories. The relation of the multiple explanations problem to the imperfect theory problems is also described. Finally, active explanation reduction is evaluated based on four criteria - completeness, efficiency, tolerance of unavailable data and feasibility.

---

* This research was partially supported by the Office of Naval Research under grant N-00014-86-K-0309.

† University of Illinois Cognitive Science/Artificial Intelligence Fellow.

ACTIVE EXPLANATION REDUCTION:
An Approach to the Multiple Explanations Problem

I INTRODUCTION

Explanation-based learning [DeJong86, Mitchell86] is a powerful learning technique which involves 1) constructing an *explanation* for *why* a given example is an instance of the goal concept and 2) generalizing the explanation to obtain an operational goal concept. The construction of the explanation is a knowledge-intensive process requiring a complete and correct domain theory. However, such perfect domain theories are rarely available in practice. Consequently, the success of explanation-based learning hinges on the construction of explanations from imperfect domain theories [Dietterich86, Mitchell86, Rajamoney87].

One of the important problems due to imperfect domain theories is the *multiple explanations problem*: the explanation construction process yields a set of incompatible explanations for why the given example is an instance of the goal concept. A standard explanation-based learning system cannot handle the multiple explanations problem. Such systems either rely on a single explanation or on multiple, but equally-valid, explanations from the domain theory. Any selection results in a different, but correct, generalization. For multiple valid explanations, the selection of an arbitrary explanation will not have major implications on the learning process. Multiple incompatible explanations pose a difficult problem. The system cannot arbitrarily select an explanation since selecting the wrong explanation will have profound implications on its subsequent problem solving and learning behavior. Nor can it generalize all the explanations because then the problem solving component will have a difficult time selecting the correct generalized rule.

The multiple explanations problem is central to explanation-based learning from imperfect domain theories. Multiple explanations can be due to:

(1) Incomplete Domain Theories: The explanation constructor is forced to make assumptions due to insufficient knowledge. Each assumption can result in a distinct explanation.

(2) Intractable Domain Theories: The explanation constructor makes approximations to make the explanation construction process tractable. But an explanation based on approximations can correspond to multiple explanations from the exact theory leading to problems during the refinement of the approximations.

(3) Incorrect Domain Theories: The domain theory may have incorrect generalizations which can result in multiple explanations.

Consider, for example, a robot that is operating in the real world. Since it is impossible to completely specify the state of the world, it has to operate under incomplete information. Suppose it observes a string increasing in length attached to a wall. In this case, it may come up with a number of explanations for the increase in length of the string: 1) The string is hot and is expanding. 2) The string is growing with time like children do. 3) The string is elastic and is being

pulled. If it were to arbitrarily select an explanation, then it may conclude incorrectly that all strings age and may never learn about elastic strings. Or if it generalizes all the explanations, then when it has to increase the length of a string it may decide to use the first generalized rule. In which case, it may decide to heat the string.

One human method to combat multiple explanations is to actively interact with the environment. The interactions yield new data from the world which may eliminate some of the explanations. Even when there are relatively few explanations, active pruning may be desirable. Humans often perform cheap tests to head off undesirable possibilities. They heft a snowball several times before throwing it or test the swimming pool temperature with a toe before diving in.

This paper proposes a method called *active explanation reduction* as a partial solution to the multiple explanations problem. Active explanation reduction involves the purposeful interaction with the world in such a way that observable behavior will dictate which of a number of incompatible explanations corresponds to reality. This process can be looked upon as conducting experiments in the world. The outcome of these experiments provide additional data which, provided the experiments are suitably chosen, will be inconsistent with a number of explanations. These explanations can then be eliminated from further consideration.

Central to active explanation reduction is the issue of experimental design. The experimental design process must have several important features. First, it must be complete; if there is a way to tease apart different explanations the design system should find it. Second, it must be tolerant of unavailable data. Third, it should be efficient. Each experiment should evenly divide the explanations so that significant information is acquired regardless of the experiment's outcome. Fourth, it should be practical. Lighting a match is not a reasonable way to tell whether a nearby barrel contains water or gasoline.

This paper presents a theory and an implementation of experiment design. The theory describes three strategies for designing experiments - *elaboration*, *discrimination* and *transformation*. The theory and an experiment engine - an implementation of the theory - are illustrated using a detailed example.

II ACTIVE EXPLANATION REDUCTION

Active explanation reduction is a technique for finding the explanation that is consistent with reality from a given set of candidate explanations. Figure 1 illustrates the major components of an active explanation reduction system. Active explanation reduction involves the following steps:

[a] Hypothesis Identification

This step involves the identification of the hypotheses underlying the explanations. The explanation constructor may have made assumptions about the state of the world and generated explanations based on each hypothesis. For example, in the case of the robot and the string, the robot hypothesized that the string may be elastic, or is being heated, or is growing. If no hypotheses were made or if the hypotheses are difficult to isolate then the entire explanation is treated as a hypothesis.

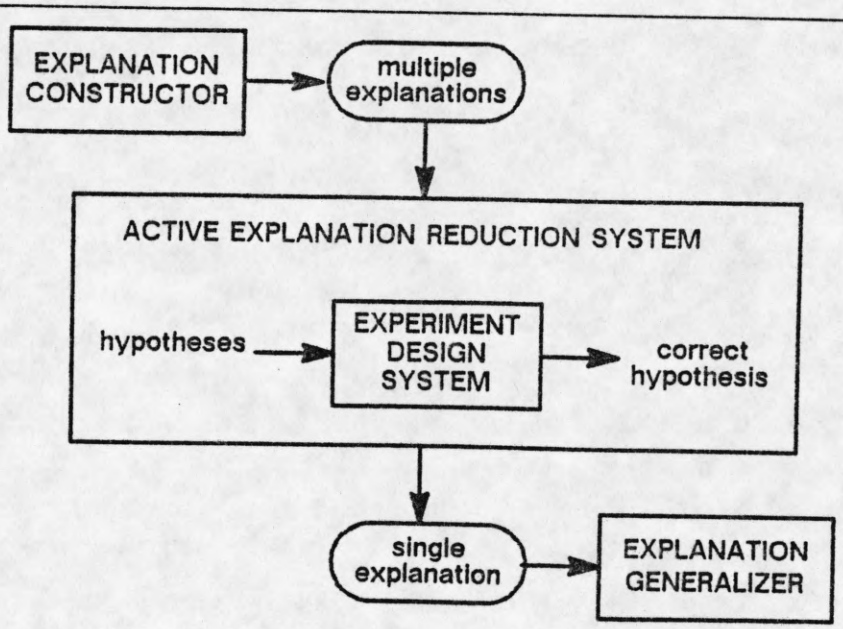

Figure 1: The block diagram for the active explanation reduction system.

[b] Experiment Design

Experiments are designed to test the hypotheses directly or indirectly by testing their ramifications. One of the ramifications of a hypothesis is the original explanation itself. Experiments yield new information from the real world. The next section describes in detail the experiment design process.

[c] Explanation Refutation

The information obtained from the experiments can be used to reject hypotheses and the explanations that were constructed based on those hypotheses. A hypothesis is rejected if it entails ramifications not consistent with the results of the experiments.

A. Strategies for Experiment Design

There are three strategies for designing experiments to refute hypotheses:

[a] Elaboration:

In a given domain, there are usually some quantities that are easily measurable. In *elaboration*, the quantity to be measured is selected according to the ease with which it can be measured. Hypotheses that predict values for the quantity that are not compatible with the measured value can be immediately refuted. This strategy does not guarantee that the designed experiment will refute a hypothesis since all the hypotheses may support the observed value or may not make any predictions regarding the value of the quantity. Elaboration does not involve comparing hypotheses during the design of experiments since it is costly. Consider the robot and the string example. Elaboration would recommend noticing the color of the string since it is easily

measurable. But since none of the hypotheses predict a change in color of the string, the experiment will not refute any of the hypotheses. On the other hand, if a set of hypotheses predicted different changes to the level of water in a container, then elaboration would successfully eliminate some hypotheses since it would recommend noticing the change in the level.

[b] Discrimination:

In *discrimination*, a quantity is selected only if its measurement will help in the refutation of hypotheses. Two values are said to be *discriminable* if a measurement can distinguish between the two values. Discrimination involves the measurement of a quantity which satisfies two criteria: 1) A number of hypotheses should predict different and incompatible values for the quantity. 2) These values should be discriminable. A quantity that satisfies these two criteria is called a *discriminant*. A discrimination experiment is guaranteed to refute hypotheses if the values of the discriminant can be grouped into sets of discriminable and mutually incompatible values and each set has at least one hypothesis that predicts a value from that set. Discrimination is more effective than elaboration, but may be much more expensive.

[c] Transformation:

It may not be possible to identify the correct hypothesis even after the space of measurable quantities for the scenario has been exhausted. *Transformation* of scenarios is a powerful technique for experiment design. It involves modifying the original scenario in a well-specified manner, such as, changing the values of the properties of the components of the scenario (changing the concentration or amount of a solution), replacing components (replacing a container with a heat-insulated container) or re-organizing the manner in which the components are put together (separating two containers originally connected by a pipe). The techniques of elaboration and discrimination can then be used on the new scenario to refute hypotheses. There are three important consequences of creating a new scenario:

(1) **Divergence of values**: Hypotheses that previously predicted identical or indiscriminable values for a quantity may predict different or discriminable values in the new scenario. Consider an example where the level of the water in a container is observed to be constant. This may be explained by two hypotheses: there are two equal flows - a flow of water into the container and a flow of water out of the container or there are no flows. If the scenario is transformed into one in which one of the pipes attached to the container is made bigger (by opening an attached valve wider) then each hypothesis predicts different and discriminable values for the level of the water in the container. If the first hypothesis is true and there is a flow of water through the pipe, then this flow will be increased in the transformed scenario. The two flows will no longer be balanced resulting in a change of level of the water in the container. However, the transformation will have no effect if the second hypothesis is true. The prediction in this case is that the level of water will remain constant.

(2) **Emergent observations**: Quantities that could not be previously observed or measured become observable or measurable in the new scenario. Consider, for example, tracing an unknown path of flowing water. If a dye is added to the original water, then the color changes

in the water will enable the path to be traced.

(3) **Differential discrimination**: Differential discrimination involves the measurement of a quantity that satisfies the following two criteria: 1) There are a number of hypotheses that predict the same value or indiscriminable values for the quantity in the original and the transformed scenarios. 2) The manner in which the value or indiscriminable values was reached in the transformed scenario is different and discriminable. The observations may be reached much faster or more of the observed behavior may occur in the same time span as compared to the other scenario. Thus differential discrimination involves the discrimination of the second order behavior of a quantity across two scenarios. Consider the example of the robot and the string. Suppose the original scenario is transformed into one in which the robot is pulling the string with a specified force. The hypothesis that claims that strings grow with time would predict the same increase in the length of the string in the two scenarios. However, the hypothesis based on the elasticity of the string would predict a greater increase in size in the second scenario due to the additional force exerted by the robot. Elaboration would recommend observing the change in the length of the string (since it is easy to measure). The experiment would result in the refutation of one of the two hypotheses depending on the observations made when the experiment was performed.

B. A Domain-Independent Experiment Engine

An experiment engine has been developed based on the model of experimentation described above. The inputs to the experiment engine are:

(1) A set of hypotheses.

(2) An inference engine that accepts a hypothesis and a scenario and returns a set of predictions supported by the hypothesis for the given scenario.

(3) Domain-dependent knowledge: There are two major sources of domain knowledge:

[a] A set of predicates that describe the quantities of the domain that can be measured or observed, the values that are discriminable, the parameters of a scenario that can be transformed etc.

[b] A set of scenario transformation operators. A scenario transformation operator constructs a new scenario from a given scenario. It can change the quantities of some of the components of the scenario eg. the concentration of a solution, the components of the scenario eg. replacing a solution by a different solution or the manner in which the components are organized eg. isolating two containers which were previously connected by a pipe. These operators endow the experiment designer with the ability to construct new scenarios.

The experiment engine uses elaboration, discrimination and transformation to design experiments. Transformation of scenarios is viewed as a planning problem. The initial state is the given scenario and the goal state is a transformed scenario in which there is a discriminable new observation, divergent values or discriminable first order or second order behavior. The plan is a sequence of transformations resulting in a scenario satisfying the goal criterion. Both a weak

method planning strategy (Breadth-First Search) and a knowledge-intensive strategy (based on Qualitative Process theory [Forbus84]) have been implemented.

III REPRESENTATION OF THE DOMAIN THEORIES

Domain theories are represented using Forbus' Qualitative Process theory [Forbus84]. According to this theory, changes in the world such as boiling, cooling, heating, etc. are due to *processes*. A process is composed of five pieces of information: *individuals* - a set of objects participating in the process; *preconditions* and *quantity conditions* - a set of conditions that must be true for the process to be active and *relations* and *influences* - a set of statements about the world that are true if the process is active. QP theory provides a language for representing domain theories based on processes. Qualitative Process Engine (QPE) [Forbus86] is an implementation of QP theory. It serves as the explanation constructor for the active explanation reduction system and the inference engine for the experiment engine.[1] QPE provides a description of the behavior of the input scenario based on the specified domain theory.

Information required for elaboration and discrimination are supplied by predicates such as:

(easily-measurable (?change amount ?liquid) Scenario)

where "Scenario" refers to a class of scenarios for which this predicate is applicable. In QP theory, the values for changes to quantities such as temperature, pressure, concentration etc are represented qualitatively as increase, decrease, constant or unknown. The predicates supplied to the experiment engine are expressed in these values. For example, the information required by discrimination is supplied in the form of a predicate:

(discriminable (increase amount ?liquid) (decrease amount ?liquid) Scenario).

For domains represented by QP theory, the experiment engine is supplied with a set of transformation schemata. These schemata are general-purpose transformations on scenarios and are indexed by the type of the hypothesis. This is an example of a knowledge-intensive (but not domain-specific) strategy used by the experiment engine in preference to the default breadth-first search strategy. Some of these transformations are:

(1) If a process is hypothesized to cause an observation then transform the scenario into a new one in which one of the preconditions of the process is negated.

(2) If a process is hypothesized to cause an observation then transform the scenario into a new one in which the rate of the process is increased.

(3) If a hypothesized process entails an unobservable effect then transform the scenario into a new one in which the unobserved effect can be observed.

(4) If two processes are hypothesized to balance a quantity then transform the scenario into a one in which the rate of one of the processes is increased and the other rate is not increased.

---

[1] QPE will be integrated with the active explanations reduction system soon. Currently, the explanations are directly provided to the system.

IV MULTIPLE EXPLANATIONS FROM INTRACTABLE THEORIES

The multiple explanations problem can arise when dealing with the intractable theory problem. Doyle [Doyle86] describes an approach to the intractable theory problem. In his approach, the domain theory is described at different levels of approximation. The most detailed description of the theory is intractable. Each higher approximate theory is more tractable than the lower detailed theories. The problem solver uses the most approximate theory during problem solving. When it fails, it obtains an explanation for the failure from the more detailed theories and uses this explanation to refine the approximate theory. The approximate theories reflect the level of detail required by the problem solver and, at the same time, are tractable. However, his approach makes a critical assumption - there is only one explanation from the detailed theory corresponding to the failure. This is not valid in most cases. In general, there will be a number of explanations for the failure and it will be necessary to identify the correct explanation to correctly refine the approximate theory.

A. A Detailed Example

This example describes the multiple explanations problem due to intractable theories and is from the domain of chemistry. The initial domain theory of the system includes processes for heat flow, electricity flow and chemical reaction. The approximate theory has a naive notion of a chemical decomposition reaction shown in figure 2. The detailed theory breaks this up into three different processes: 1) *Catalytic decomposition* - a catalyst, not changed by the reaction, is required. The rate of the reaction depends on the amount of catalyst in contact with the decomposing substance. 2) *Heat decomposition* - heat is required. The rate of the decomposition depends on the rate at which the heat is supplied. 3) *Electrical decomposition* - an electric current is required. The rate of the reaction depends on the magnitude of the current that is passed through the solution.

Suppose the task is to produce a large quantity of oxygen - a substance that is in great demand in industry. However, initially the problem solving system does not have a method for producing oxygen. It is then shown a scenario where oxygen is produced from water (figure 3). The explanation constructor tries to explain the generation of oxygen using its naive theory of chemical decomposition (figure 2). In this case, it fails since it knows that under normal conditions water does not decompose into hydrogen and oxygen. It then explores the more detailed theory (figure 2) and finds three different explanations (figure 4) corresponding to the known types of decomposition reactions: 1) the decomposition is due to the heating of water, 2) the decomposition is due to the electric current, or 3) the decomposition is due to the presence of platinum which served as a catalyst.

Note that simple-minded approaches to the multiple explanations problem will not work. Suppose the learning system selects an arbitrary explanation and uses that to form a new process. If it has made the wrong choice, for example, using the catalytic decomposition explanation instead of the correct electricity decomposition explanation, then the problem solver will try to generate oxygen by adding platinum to water and will fail. Alternatively, suppose the learning system uses a conjunction of the hypotheses to form the new generalized process. Then the problem solver will

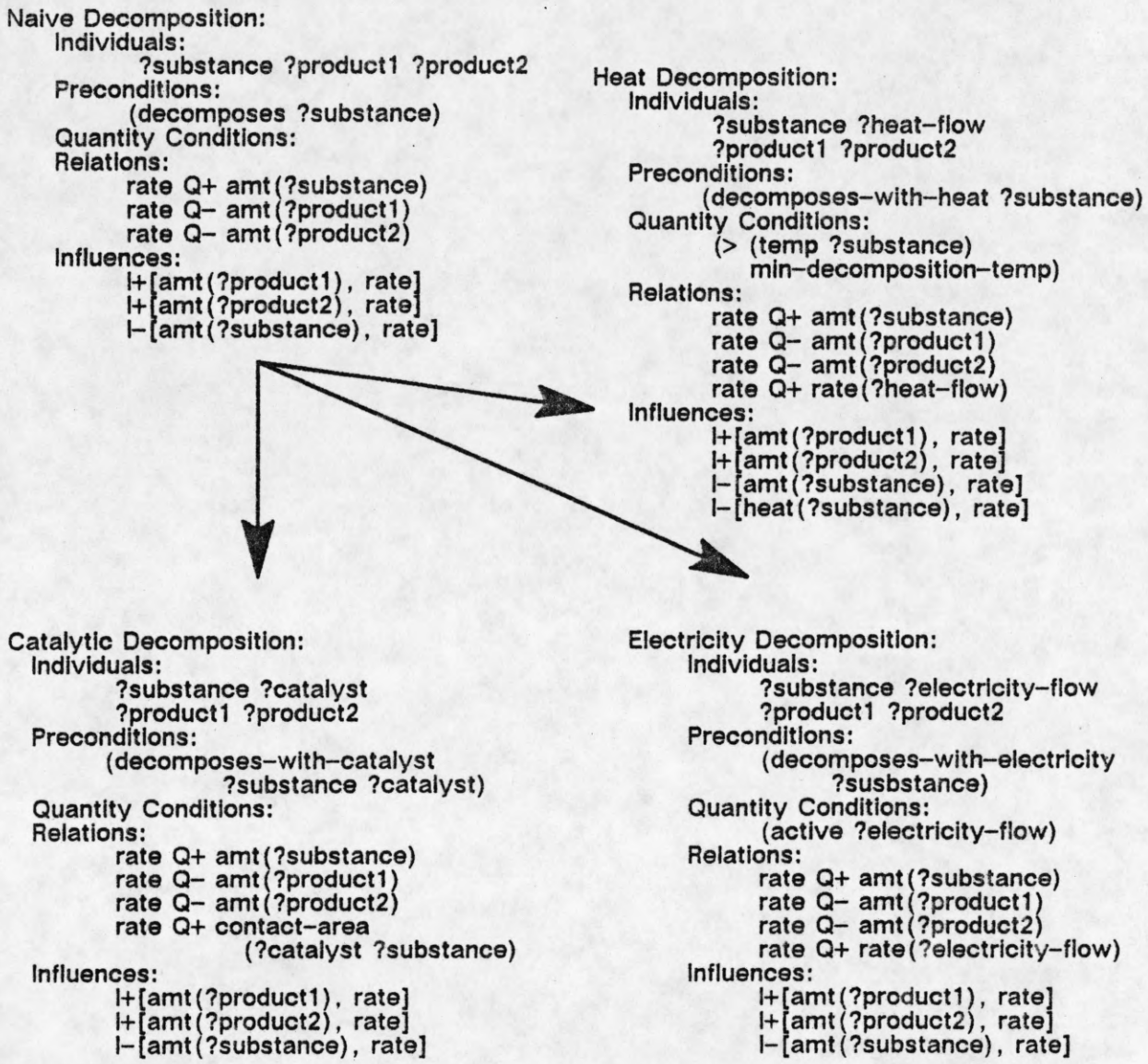

Figure 2: The two different theories used by the explanation constructor. The naive decomposition reaction corresponds to three different reactions in the detailed theory. The heat decomposition reaction requires a heat flow to provide the heat. The catalytic decomposition reaction requires a catalyst. The electricity decomposition reaction requires a direct electric current through the solution. According to QP theory notation, "Qty1 Q+ Qty2" means that Qty1 is directly qualitatively proportional to Qty2, that is, Qty1 is increasing monotonically with Qty2. "I+ Qty1 Qty2" means that Qty1 is directly influenced by Qty2. The direct influences of Qty1 are specified only in the processes that cause Qty1 to change.

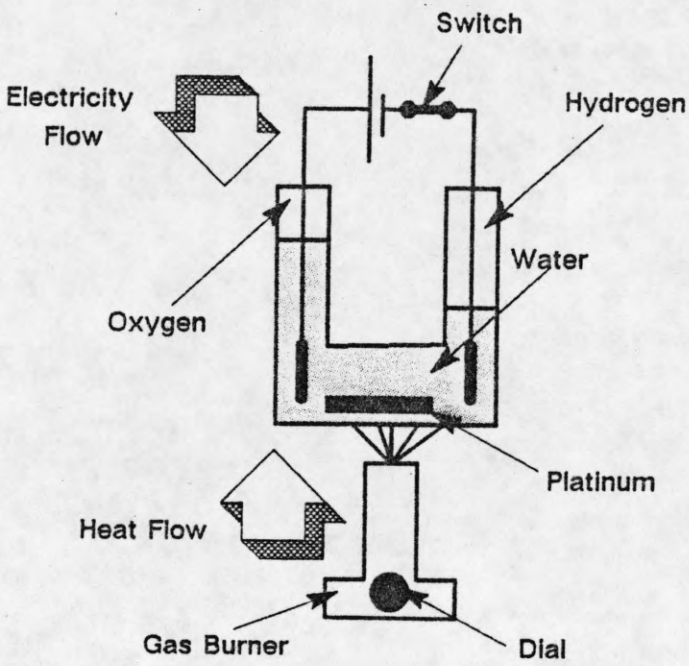

Figure 3: The scenario describing the observed production of oxygen. There are two active processes - a heat flow from a gas burner and an electricity flow through water from an external battery.

be able to generate oxygen but will fail for other products in which the original substance is destroyed by heating or mixing with platinum. Besides, the problem solver will do unnecessary work trying to achieve goals such as generating a heat process and obtaining platinum. Integrating all the hypotheses together using a disjunctive precondition is not a satisfactory solution either. Each process has characteristics that are not shared by others - for example, heat decomposition reaction predicts that the decomposition will proceed faster if the rate of heat supplied is increased and this is not true in the case of the electricity decomposition reaction or the catalytic decomposition reaction. Therefore, it is crucial that the learning system isolate the correct explanation.

The explanation constructor gives these three explanations to the active explanation reduction module. The active reduction explanation module retrieves the three hypotheses that the explanation constructor made about the nature of the decomposition of water - it is 1) catalytic decomposition 2) heat decomposition or 3) electrical decomposition. These hypotheses are tested by the experiment engine. All the predictions obtained from the inference engine for the given scenario (the original scenario) are supported by the three hypotheses. Direct elaboration and discrimination fail. The experiment engine now tries to transform the scenario. The experiment engine needs to know if one of the three decomposition reaction processes is active and is causing the generation of the product. Suppose the heat burner cannot be turned off but the heat supplied can be increased or

decreased. The first transformation strategy described in the section 3 cannot be applied to the heat flow process because experimentation does not have the capability to construct a scenario with any of the preconditions of the heat flow process negated. Instead, it uses the second transformation strategy which is based on the rate of the decomposition. It transforms the scenario to a new scenario in which the heat from the burner has been increased. The inference engine now predicts that under the heat decomposition hypothesis oxygen will be generated at a faster rate in the new scenario as compared to the old scenario since the rate of the decomposition depends on the rate at which the heat is supplied. However, under the other two hypotheses both these rates will be the same since the decomposition rate is not affected by heat. Differential discrimination on the new scenario and the original scenario will recommend comparing the rate at which oxygen is generated in each scenario. When they are determined to be the same the heat decomposition hypothesis is refuted.

The experiment engine is then left with two hypotheses: the catalytic decomposition hypothesis and the electrical decomposition hypothesis. The transformation strategy previously used can be reapplied to the catalytic decomposition process. The rate of this reaction depends on the surface area of contact of the catalyst with the reactants. The transformation strategy will suggest a new scenario in which the surface area of contact between water and platinum is

---

```
(increase amt oxygen)
        (I+ (amt oxygen) catalytic-decomposition-rate)
                (active catalytic-decomposition)
                        (decomposes-with-catalyst water platinum)
                                :Hypothesis

(increase amt oxygen)
        (I+ (amt oxygen) heat-decomposition-rate)
                (active heat-decomposition)
                        (decomposes-with-heat water)
                                :Hypothesis
                        (greater-than (temp water) (min-decomposition-temp water))
                                :Premise
(increase amt oxygen)
        (I+ (amt oxygen) electricity-decomposition-rate)
                (active electricity-decomposition)
                        (decomposes-with-electricity water)
                                :Hypothesis
                        (active electricity-flow)
                                <explanation1>
```

Figure 4: The three explanations based on catalytic decomposition, heat decomposition and electrical decomposition of water.

---

increased - for example, by breaking the original platinum pieces into several smaller pieces or by using a larger piece. Under this transformation, the catalytic decomposition hypothesis will predict that oxygen is generated at a faster rate in the second scenario. However, the electrical decomposition hypothesis will predict that the two rates will be the same. An experiment based on differential discrimination will recommend comparing the rates at which oxygen is produced in the two scenarios. When they are determined to be the same the catalytic decomposition hypothesis is refuted.

As it happened, the system chose the first two experiments based on the catalytic and heat decomposition hypotheses (since it had no a priori information about which of the three hypotheses is more likely). The electrical decomposition hypothesis could also have been tested by transformation and discrimination. Based on the first transformation strategy described in the previous section, the original scenario can be transformed into a new scenario in which one of the preconditions of the electrical decomposition process is negated. Suppose, the original scenario is transformed by turning off a switch thereby breaking the path of the electric current. Then the electric decomposition hypothesis will predict that the amount of oxygen will not change - that is, no more oxygen will be generated.

Thus, the experiment engine found that the decomposition reaction is due to the electric current. Doyle's approach and explanation-based learning can now be used to incorporate the generalized version of the explanation into the naive theory. In addition, the system has learned through experimentation that water decomposes when an electric current is passed through it - domain-specific information which it did not have previously.

V EVALUATION OF THE ACTIVE EXPLANATION REDUCTION TECHNIQUE

Experiment design is the central theme behind active explanation reduction. The experiment design system is evaluated based on the four criteria proposed earlier: 1) completeness 2) efficiency 3) tolerance of unavailable data and 4) feasibility.

Completeness

The experiment design system will find an experiment to discriminate between two explanations, if it exists, if 1) The predicates supplied to the system are complete and correct - all quantities that can be observed or measured are known to the system. 2) The transformation operators or the transformation strategies, if supplied, cover the space of scenarios that can be constructed. 3) The inference engine provides all the predictions that are possible for the given scenario. Completeness can be sacrificed for efficiency by 1) using a heuristic search or beam search for the transformation of scenarios and 2) examining only a selected set of predictions from the inference engine.

Efficiency

The experiment design system will produce the fewest experiments if for each experiment it selects those quantities to measure and those transformations to make that will lead to a maximum number of hypotheses being refuted. However, this will require a priori

information about how the world is going to behave. Instead, during discrimination, if there are many quantities that can act as discriminants, the system selects the discriminant which will divide the hypotheses equally. This will lead to a maximum number of hypotheses being refuted if all of the observations are equally likely. Transformation can be made efficient by using a good transformation strategy that proposes transformations that yield discriminable behavior. For example, the transformation strategy that constructs new scenarios by negating conditions of a process when the process is hypothesized to cause an observation yields discriminable predictions about the observed quantity.

Tolerance of unavailable data

The experiment design system should be capable of constructing other experiments if the present experiment fails to yield any information. If the system is complete and another experiment exists then the system will find it. Redundancy can also be built into the transformation strategy. For example, if negating the precondition is not feasible, then the system can construct a new scenario in which the rate of the process is varied.

Feasibility

The experiment design system should propose only those experiments that are feasible or practical in the real world. The predicates supplied to the experiment design system determine which experiments are feasible. Since, the two basic techniques, elaboration and discrimination, will construct experiments to measure quantities based on these predicates, if they are correct then the experiment is feasible. Transformation also uses the supplied predicates to check if the proposed scenario construction is feasible. For example, while constructing scenarios to vary the rate of the process, the system checks whether the parameter to be varied, such as the dial on the gas burner, is manipulable in the required manner.

VI RELATED WORK

The current work shares a superficial similarity with the BACON system of Langley [Langley81]. While both propose experiments to gain crucial missing information, BACON does not do so in the context of a qualitative model of the world. Similarity with the later systems of NGLAUBER [Jones86] and STAHLP [Rose86] are deeper but less obvious. The purpose of these systems is theory formation and refinement and, therefore, their processing is initiated by rather different world situations. There are also important differences in internal processing. NGLAUBER uses a data-driven clustering algorithm to generate rules for observed data, and STAHLP does not perform experiments but rather relies on minimizing a cost function to decide among competing conjectures. A more recent discovery system, IDS [Nordhausen87], also proposes experiments to gather data. However, the experiments are motivated towards the discovery of new phenomena and, unlike our system, are not directed towards the refutation of well-formed hypotheses. Also, our experimentation design system is based on a theory of experiment design that includes elaboration, discrimination and transformation. Carbonell and Gil [Carbonell87] have recently proposed a system that learns new preconditions and postconditions for STRIPS-like operators by

experimentation. Their experimentation involves comparing states of the world to identify differences. Again, unlike our system, their experiment design is not based on the refutation of hypotheses and a theory of experiment design. Our approach is also related to the *determinations* of [Russell87]. Whereas, Russell assumes the presence of examples to find which disjunction leads to the generalized concept, our experiment design system automatically generates such examples.

VII DISCUSSION

We have outlined a method to cope with multiple explanations which arise in the normal process of explanation construction from imperfect theories. The approach, called *Active Explanation Reduction*, would be invoked by an explanation-based learning system confronted with multiple incompatible explanations. When invoked, the system identifies measurements which would serve to disambiguate among the postulated alternatives. The system then proposes experiments by which these measurements can be obtained. A major portion of the research contribution addresses the problem of *experimental design*.

Active explanation reduction is a general technique that has been used to deal with problems in machine learning and qualitative reasoning. Apart from the intractable theory problems, it has also been used to deal with multiple explanations due to incomplete and incorrect theory problems. The explanations in these cases are based on hypothesized revisions to the imperfect theory and the active explanation reduction technique identifies the correct revised theory. Active explanation reduction has also been applied to problems within qualitative reasoning tasks such as *envisionment* and *measurement interpretation*. Such tasks are swamped by a large number of possibilities due to the ambiguous nature of qualitative reasoning. Experimentation is used to obtain the information required to eliminate those possibilities that are inconsistent with reality, thereby making the reasoning tasks more tractable.

Continuing research includes the construction of a model of theory refinement to deal with the incomplete and incorrect theory problems based on Forbus' QP theory [Forbus84]. The resulting system has a similar motivation to STAHLP [Rose86] but is experimentally oriented. Qualitative reasoning, and QP theory in particular, rely on an accurate description of all the processes of a domain. If the theory is flawed - for example, a process is missing, a precondition is incorrect, or an influence is missing then discrepancies may arise between predictions of the model and observations of reality. In general there will be many different changes to the theory that remove the anomaly. Each change may result in a distinct qualitative theory. The resulting theories form a set of ambiguous hypotheses. We are extending the ADEPT system [Rajamoney88] to integrate our notions of active explanation reduction to experimentally determine which of the various hypotheses correspond to reality.