D. Sleeman and P. Edwards (eds.), *Machine Learning: Proceedings of the Ninth Inernational Conference*, Aberdeen, Scotland (July, 1992), pp. 123-128.

Multistrategy Learning with Introspective Meta-Explanations

Michael T. Cox
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332-0280
cox@cc.gatech.edu

Ashwin Ram
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332-0280
ashwin@cc.gatech.edu

Abstract

Given an arbitrary learning situation, it is difficult to determine the most appropriate learning strategy. The goal of this research is to provide a general representation and processing framework for introspective reasoning for strategy selection. The learning framework for an introspective system is to perform some reasoning task. As it does, the system also records a trace of the reasoning itself, along with the results of such reasoning. If a reasoning failure occurs, the system retrieves and applies an introspective explanation of the failure in order to understand the error and repair the knowledge base. A knowledge structure called a Meta-Explanation Pattern is used to both explain how conclusions are derived and why such conclusions fail. If reasoning is represented in an explicit, declarative manner, the system can examine its own reasoning, analyze its reasoning failures, identify what it needs to learn, and select appropriate learning strategies in order to learn the required knowledge without overreliance on the programmer.

1 INTRODUCTION

In recent years several machine learning techniques have been proposed. Yet it is problematic, given a particular learning situation, to determine the most appropriate learning strategy. Many learning theories depend upon particular domains and specific classes of problems. The goal of this research is to provide a general representation and processing framework for introspective reasoning about reasoning failures and the selection of appropriate learning strategies for different failure classes. A taxonomy of reasoning failures is being developed toward this end. It is claimed that explicit, declarative representations of reasoning failures allow a reasoning system to examine its own reasoning processes, analyze its reasoning failures, identify what it needs to learn, and select appropriate learning strategies in order to learn the required knowledge.

The learning framework for an introspective system is as follows: First the system performs some reasoning task. As it does so, the system records a trace of the reasoning along with its conclusions and the goal it is pursuing. Included in the trace are the considerations prompting such a decision and the bases for making the decision. Monitoring its progress, the system reviews each reasoning chain in order to detect failures. If a failure develops, the system must not only correct the mistake, but must attempt to learn from the mistake in order to avoid it in the future. The learning which is performed has three phases: Identify what went wrong (blame assignment), decide what to learn, and select an appropriate learning strategy. Blame assignment requires that the system identify both faulty background knowledge (BK) and faulty processing decisions. An introspective agent can use its knowledge of failures in order to understand how it failed to reason correctly in a given situation and hence to learn. This paper will examine three types in the failure taxonomy:

Mis-indexed Structure - The reasoner may have an applicable knowledge structure to deal with a situation, but it may not be indexed in memory so that it is retrieved using the cues provided by the context. In this case the system must add a new index, or generalize an existing index based on the context. If on the other hand, the reasoner retrieves a structure that later proves inappropriate, it must specialize the indices to this structure so that the retrieval will not recur in similar situations (Cox & Ram, 1991).

Novel Situation - A failure can arise when the reasoner does not have an appropriate knowledge structures to deal with a situation. In such cases, the reasoner could use a variety of learning strategies, including explanation-based generalization (EBG) (DeJong & Mooney, 1986; Mitchell, et al., 1986) or explanation-based refinement (Ram, 1992), coupled with index learning (Hammond, 1989; Ram, 1992) for the new knowledge structures.

Incorrect BK- Even if the reasoner has applicable knowledge structures, they may be incorrect or incomplete. Learning in such cases is usually incremental, involving strategies such as elaborative question asking (Ram, 1991, 1992) applied to the reasoning chain, and abstraction or generalization techniques applied to the BK.

Meta-AQUA is a computer program that performs multistrategy learning through self-analysis of its reasoning processes during a story understanding task. In order to perform this kind of reasoning, a new kind of knowledge structure was proposed, called a *Meta-Explanation Pat-*

tern (Meta-XP) (Cox & Ram, 1991). Meta-XPs are similar to explanation patterns (Schank, 1986), and are causal justifications of the reasoning performed by a system that explain how and why the system reasons. These structures form the bases for blame assignment and learning. There are two broad classes of Meta-XPs: Trace Meta-XPs and Introspective Meta-XPs.

A *Trace Meta-XP* (TMXP) records a trace of the reasoning performed by a system along with both the causal linkages explaining the decisions taken and the goal the system was pursuing during such reasoning. TMXPs are similar to Carbonell's (1986) derivational analogy traces, except that the underlying reasoning processes may be based on a reasoning model other than search-based problem solving. TMXPs declaratively represent the mental processes employed in making a processing decision, record both the information that initiated the decision and the information that the decision was based on, and explain how given conclusions are drawn.

An *Introspective Meta-XP* (IMXP) is a structure used both to explain why reasoning processes fail and to learn from reasoning failure. It associates a failure type with a particular set of learning strategies by providing a *knowledge goal*, or a goal to learn (Ram, 1991; Ram & Hunter, to appear). IMXPs also point to likely sources of the failure within the TMXP.

This paper concentrates on the representation and use of Introspective Meta-XPs in the learning theory. Section 2 presents the overall representation of IMXPs. Section 2.1 provides a representation for base IMXPs. Section 2.2 discusses the knowledge goals and plans generated by core IMXPs. Section 2.3 illustrates the theory with a processing example from Meta-AQUA using a composite Meta-XP. Section 3 closes with a discussion of some issues.

2 REPRESENTATION OF INTROSPECTIVE META-XPS

Whereas a Trace Meta-XP explains how a failure occurred, providing the sequence of mental events and states along with the causal linkage between them, an Introspective Meta-XP explains why the results of a chain of reasoning are wrong. The IMXP posits a causal reckoning between the events and states of the TMXP. In addition, an IMXP provides a learning goal specifying what needs to be learned. Then, given such an explanation bound to a reasoning chain, the task of the system is to select a learning strategy to reduce the likelihood of repeating the failure.

An IMXP consists of six distinctive parts:

- The IMXP type class
- The failure type accounted for by the IMXP
- A graph representation of the failure
- Temporal ordering on the links of the graph
- An ordered list of likely locations in the graph where the processing error may have occurred.
- A corresponding list of knowledge goals that can be spawned in order to repair the failure.

There are three classes of IMXPs: base, core, and composite. *Base* types constitute the blocks with which *core* IMXPs are built. We have identified six types in the base class: successful prediction, inferential expectation failure, incorporation failure, belated prediction, retrieval failure, and input failure. The core types are representations of the failure types described by the failure taxonomy, such as Mis-indexed Structure, Novel Situation and Incomplete-BK. Core types are combined to form *composite* IMXPs that describe situations encountered by reasoning agents, such as the example of section 2.3.

The internal structure of an IMXP consists of nodes, representing both mental states and mental events (processes), and the causal links between them. Enables links point from precondition states to processes; results links join processes with resultant states; and initiates links connect two states. The graph gives both a structural and a causal accounting of what happened and what should have happened when processing information.

Introspective Meta-XPs generate *knowledge goals*, which represent the system's learning goals. Knowledge goals help guide the learning process by suggesting strategies that would allow the system to learn the required knowledge. There are two classes of knowledge goals (Ram, 1991; Ram and Hunter, to appear). A *knowledge acquisition goal* constitutes a desire for knowledge to be added to the BK. A *knowledge organization goal* indicates a desire to adjust the indices which organize the BK. Using such indices the system can efficiently retrieve appropriate structures with which an input can be understood.

The knowledge goals spawned by an introspective examination of a reasoning failure are achieved by the use of learning plans, similar to those described by Hunter (1990). The plans are implemented as action sequences which call various learning algorithms. Because the knowledge goals have pointers to the trace of the introspective reasoning, they have access to the TMXPs and IMXPs involved in the analysis of the failure.

2.1 BASE CLASS IMXPS

The three types of failures discussed in the introduction (Mis-indexed Structure, Novel Situation and Incorrect BK) can be accounted for by the complementary notions of omission error and commission error. Commission errors stem from reasoning which should not have been performed or knowledge which should not have been used. Omission errors originate from the lack of some reasoning or knowledge.

We have identified two types of commission errors: *Inferential expectation failures* typify errors of projection. They occur when the reasoner expects an event to happen in a certain way, but the actual event is different or missing. *Incorporation failures* result from an object or event having some attribute which contradicts some restriction on its values. Additionally, three omission errors have been identified: *Belated prediction* occurs after the fact. Some prediction which should have occurred did not, but only in hindsight is this observation made. *Retrieval failures* occur when a reasoner cannot remember an appropriate piece of knowledge. In essence it represents forgetting. *Input failure* is error due to lack of some input information. To construct the three core types described in this paper, representations for expectation failure, retrieval failure, and incorporation failure are needed.

2.1.1 Inferential Expectation Failure

To illustrate representations of the base types, let node A be an actual occurrence of an event, an explanation, or an arbitrary proposition. The node A (see Fig. 1)[1] results from either a mental calculation or an input concept. Let node E be the expected occurrence. The expected node E `mentally-results` from some reasoning trace enabled by some goal, G. Now if the two propositions are identical, so that $A = E$, or $A \supset E$, then a successful prediction has occurred.[2] Failures occur when $A \neq E$. This state exists when either A and E are disjoint, or conflicting assertions within the two nodes conflict. For example, A and E may represent persons, but E contains a relation specifying gender = male, whereas A contains the relation gender = female. Inferential expectation failures occur when the reasoner predicts one event or feature, but another occurs instead. The awareness of expectation failure is initiated by a `not-equals` relation between A and E.

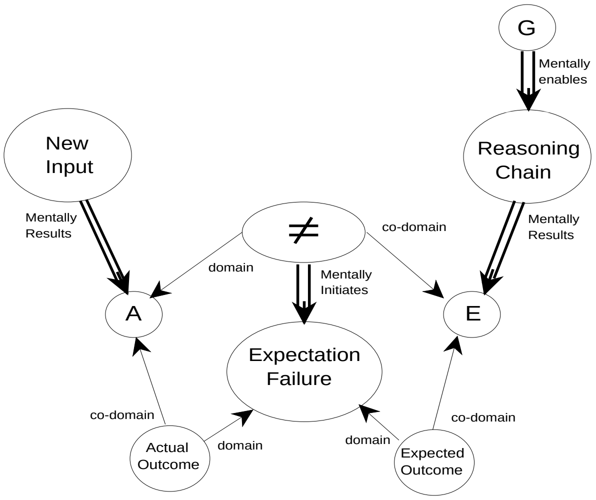

Figure 1: Expectation Failure

2.1.2 Retrieval Failure

Instead of an expectation (E) being present, it is absent with retrieval failure due to the inability of the system to retrieve a knowledge structure that produces E (see Fig. 2). To represent these conditions, Meta-AQUA uses non-monotonic logic values of `in` (in the current set of beliefs) and `out` (out of the current set of beliefs) (Doyle, 1979). Extended values include `hypothesized-in` (weakly assumed in) and `hypothesized` (unknown). Thus absolute retrieval failure is represented by A [truth = `in`] = E [truth = `out`]. The relation that identifies the truth value of E as being out of the current set of beliefs `mentally-initiates` the assertion that a retrieval failure exists. Cuts across links in the figure signify causal relations for which the truth slot of the link is also `out`.

2.1.3 Incorporation Failure

When the incorporation of some input into memory fails due to conflict with the BK, an incorporation failure exists. The conflict produces a `not-equals` relation between the actual occurrence and a conceptual constraint. This relation `mentally-initiates` the anomaly (Fig. 3). Such anomalies are used to identify questions to drive the reasoning and learning processes.

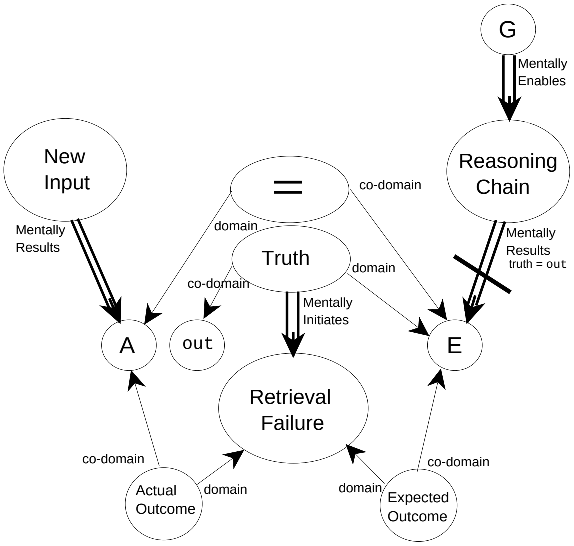

Figure 2: Retrieval Failure

2.2 CORE CLASS IMXPS

2.2.1 Mis-indexed Structure

The core type Mis-indexed Structure has two variants: Erroneous Association and Missing Association. An *Erroneous Association* is represented with inferential expectation failure. An index has associated some context with part of the BK that produced incorrect inferences. A knowledge organization goal is spawned to adjust the index so that it will still retrieve those structures in the BK when appropriate, but not in future instances similar to the current situation. Learning plans are associated with such goals to execute a specialization algorithm producing a more discriminating index. Because the goal has links to a declarative representation of the reasoning which produced it, the algorithm has access to the context of the

[1] Attributes and relations are represented explicitly. The ACTOR attribute of event X with value Y is equivalent to the relation `ACTOR` having `domain X` and `co-domain Y`.

[2] See Cox & Ram (1991) for a summary of interpretation for $A \subset E$.

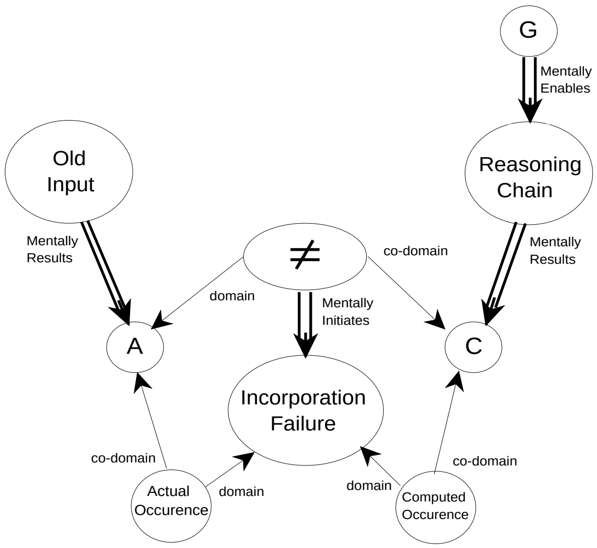

Figure 3: Incorporation Failure

error.

A *Missing Association* is represented by retrieval failure. Here, an appropriate knowledge structure was not retrieved because there was no index to associate the context with the structure. Thus some node M in the BK must be `in`. The goal associated with the IMXP is to find M. If this can be verified, then the plan which found the structure directs an indexing algorithm to examine the indices of M, looking for an index compatible with the index calculated for A. If found, this index is generalized so that the current cues provided by the context of A will retrieve E. If no such index is found, a new index is computed. If M cannot be found, a reasoning question is raised concerning the possibility that M exists. The question is represented as a knowledge goal and indexed by the context of A, and the process is suspended.

2.2.2 Novel Situation

A Novel Situation is structurally like a Mis-indexed Structure (Missing Association variant), except the node M (and thus its associated index) has a truth value of `out`. That is, there is no item in memory that can be retrieved and reasoned with to produce the expectation of a concept like A.

Novel situations occur when A ≠ E and E's truth slot is either `hypothesized-in` or `out`. When Meta-AQUA identifies a novel situation it posts a goal to learn a new explanation of the event. The associated plan is to perform EBG on node A, so that the knowledge can be applied to a wider set of future events. The plan also directs an indexing algorithm to the same node so that the new explanation will be retrieved in similar situations.

2.2.3 Incorrect-BK

Only one instance of the failure type Incorrect-BK is currently represented. This failure is an inconsistency between a known fact and a constraint in the BK. Such failures invoke a knowledge acquisition goal to adjust the constraint in the BK. An associated learning plan then tests whether the two assertions (the fact and the constraint) are conceptual siblings. If this is so, then the program will perform abstraction[3] on the constraint, raising it to its parent on the basis of induction. The constraint is then marked as being `hypothesized-in`. The reasoning chain which led to this hypothesis is indexed off the hypothesis so that the reasoning chain can be retrieved when the constraint is used in future stories. The hypothesis is verified if the anomalous assertion is re-encountered in later situations.

2.3 COMPOSITE CLASS IMXPS

Consider an example story processed by Meta-AQUA:

S1: A police dog sniffed at a passenger's luggage in the Atlanta airport terminal.

S2: The dog suddenly began to bark at the luggage.

S3: At this point the authorities arrested the passenger, charging him with smuggling drugs.

S4: The dog barked because it detected two kilograms of marijuana in the luggage.

Numerous inferences can be made from the story, many of which may be incorrect, depending on the knowledge of the reader. Meta-AQUA's knowledge includes general facts about dogs and sniffing, including the fact that dogs bark when threatened, but it has no knowledge of police dogs. It also knows of past weapons smuggling cases, but has never seen drug interdiction. Nonetheless the program is able to recover and learn from the erroneous inferences this story generates.

S1 produces no inferences other than sniffing is a normal event in the life of a dog. However, S2 produces an anomaly because the system's definition of "bark" specifies that the object of a bark is animate. So the program (incorrectly) believes that dogs bark only when threatened by animate objects. Since luggage is inanimate, there is a contradiction, leading to an incorporation failure. This anomaly causes the understander to ask why the dog barked at an inanimate object. It is able to produce but one explanation: the luggage somehow threatened the dog. The BK contains only this reason for why dogs bark.

S3 asserts an arrest scene which reminds Meta-AQUA of an incident of weapons smuggling by terrorists. The system then infers a smuggling bust that includes detection, confiscation, and arrest scenes. Because baggage searches are the only detection method the system knows, the sniffing event remains unconnected to the rest of the story.

Finally, S4 causes the question generated by S2 "Why did

[3] The use of the term abstraction is as defined by Michalski (1991), and can be opposed to that of generalization. The former is an operation on the `co-domain`s of relations, whereas the latter is an operation on relation `domain`s.

the dog bark?" to be retrieved, and the understanding task is resumed. Instead of revealing the anticipated threatening situation, S4 offers another hypothesis. The system prefers the explanation given by S4 over its earlier one. The system characterizes the reasoning error as an expectation failure caused by the incorrect retrieval of a known explanation ("dogs bark when threatened by objects," erroneously assumed to be applicable), and a missing explanation ("the dog barked because it detected marijuana," the correct explanation in this case). Using this characterization as an index, the system retrieves IMXP-Novel-Situation-Alternative-Refuted (see Fig. 4).

This composite Meta-XP consists of three core Meta-XPs: XP-Novel-Situation (centered about "Retrieval Failure"), an Erroneous Association variant of the XP-Mis-indexed-Structure (centered about "Expectation Failure") and XP-Incorrect-BK (centered about "Incorporation Failure"). The plan seeking to achieve the knowledge goal spawned by the XP-Novel-Situation directs an EBG algorithm to be applied to the explanation of the bark (node A2). Since the detection scene of the drug-bust case and the node representing the sniffing are unified due to the explanation given in S4, the explanation is generalized to drug busts in general and installed at the location of node M′. The explanation is then indexed in memory, creating a new index (I′). The plan for the goal of the XP-Mis-indexed-Structure directs an indexing algorithm to the defensive barking explanation (node E). It recommends that the explanation be re-indexed so that it is not retrieved in similar situations in the future. Thus the index for this XP (node I) is specialized so that retrieval occurs only on animate objects, not physical objects in general. The plan achieving the goal of the XP-Incorrect-BK directs the system to examine the source of the story's anomaly. The solution is to alter the conceptual representation of bark so that the constraint (node C) on the object of dog-barking instantiations is abstracted from animate objects to physical objects.

Although the program is directly provided an explanation linking the story together, Meta-AQUA performs more than mere rote learning. It learns to avoid the mistakes made during the story processing. Meta-XPs allow the system to choose appropriate learning strategies in order to learn exactly that which the system needs to know to process similar future situations correctly. A subsequent story, in which a police dog is used to find a marijuana plant in a suspect's home trash bin produces no errors.

3 DISCUSSION

Meta-XPs provide a number of computational benefits. Because Trace Meta-XPs make the trace of reasoning explicit, an intelligent system can directly inspect the reasons supporting specific conclusions, evaluate progress towards a goal, and compare its current reasoning to past reasoning in similar contexts. Hiding knowledge used by the system in procedural code is thus avoided. Instead, there exists an explicit declarative expression of the reasons for executing a given piece of code. With these reasons enumerated, a system can explain how it produced a given failure and retrieve an introspective explanation of the failure. Also, because both the reasoning process and the BK are represented using the same type of declarative representations, processes which identify and correct gaps in the BK can also be applied to the reasoning process itself. For example, a knowledge goal may be directed at the reasoning process as well as at the BK. Further, because there is a declarative trace of past reasoning processes, there is the potential for speedup learning as with derivational replay. Finally, the ability of a Meta-XP to provide goals for applicable learning algorithms to be used in given circumstances provides a sound basis for multistrategy learning.

Many multistrategy learners are simply integrated systems consisting of a cascade of more than one learning algorithm (e.g., Flann & Dietterich, 1989; Shavlik & Towell, 1989). For each and every input the control is the same. An initial learning technique is applied such that its output becomes the input to the next technique. Newer systems use more sophisticated schemes whereby various algorithms may apply to different inputs depending on the situation. In these paradigms, selection of the learning algorithm becomes computationally important. One benefit of using IMXPs in this type of framework is their ability to apply learning tasks appropriate to a given situation without having to perform blind search. Many non-cascaded multistrategy learning systems apply learning algorithms in a predefined order (e.g., Genest, et al., 1991; Pazzani, 1991). If the first fails, then the next strategy is tried, and so forth. Much effort may be wasted in worst-case scenarios.

This research has produced a novel, theoretical approach combining multiple learning methods in an integrated manner. This paper focuses on the justifications and technical details. The authors are currently involved in research to evaluate the model's cognitive plausibility as well as the computational benefits of the approach.

Acknowledgements

This research was supported by the National Science Foundation under grant IRI-9009710. The authors thank Sue Farrell for proofing a copy of this paper.

References

Cox, M., & Ram, A. Using Introspective Reasoning to Select Learning Strategies. in Michalski, R. & Tecuci, G. (eds), *Proc. of 1st Intl. Workshop on Multi-Strategy Learning*, 217-230, 1991.

Carbonell, J. G. Derivational Analogy: A theory of reconstructive problem solving and expertise acquisition, in R. Michalski, J. Carbonell, & T. Mitchell (eds.), *Machine Learning: An Artificial Intelligence Approach*, 2, Morgan Kaufmann Publishers, San Mateo, CA., 1986.

DeJong, G., & Mooney, R. Explanation-Based Learning: An Alternative View, *Machine Learning*, 1(2):145-176, 1986.

Doyle, J. A. A Truth Maintenance System, *Artificial Intelli-*

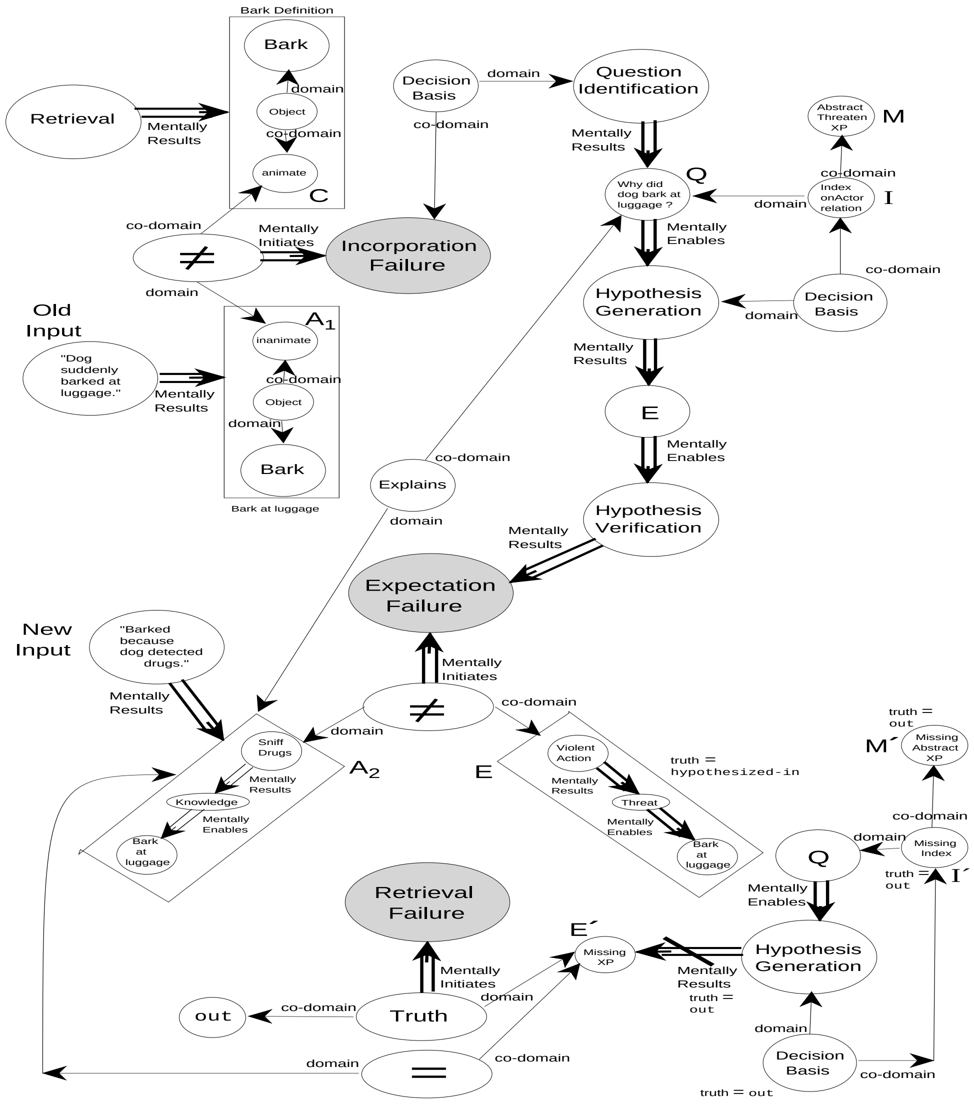

Figure 4: Instantiated IMXP-Novel-Situation-Alternative-Refuted

gence, (12):231-272, 1979.

Flann, N., & Dietterich, T. A Study of Explanation-Based Methods for Inductive Learning. *Machine Learning*. 4:187-266, 1989.

Genest, J., Matwin, S., & Plante, B. Explanation-Based Learning with Incomplete Theories: A three-step approach, in *Proc. of 7th Intl. Conf. on Machine Learning*, Austin, TX, (June), 286-294, 1990.

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

Hunter, L. E. Planning to Learn. in *Proc. of 12th Annual Conf. of the Cognitive Science Society*, Cambridge, MA, (July), 261-276, 1990.

Michalski, R. S. Inferential Learning Theory as a Basis for Multistrategy Task-Adaptive Learning. In Michalski, R. S. & Tecuci, G. (eds.), *Proc. of the 1st Intl. Workshop on Multi-Strategy Learning*, 3-18, 1991.

Mitchell, T., Keller, R., & Kedar-Cabelli, S. Explanation-Based Generalization: A unifying view, *Machine Learning*, 1(1), 1986.

Pazzani, M. Learning to Predict and Explain: An Integration of Similarity-Based, Theory-Driven, and Explanation-Based learning. *The Journal of the Learning Sciences*, 1(2):153-199, 1991.

Ram, A. A Theory of Questions and Question Asking, *The Journal of the Learning Sciences*, 1(3,4), 1991.

Ram, A. Indexing, Elaboration and Refinement: Incremental Learning of Explanatory Cases. To appear in *Machine Learning*. Also available as Tech. Report git-cc-92/04, College of Computing, Georgia Institute of Technology, Atlanta, GA, 1992.

Ram, A. and Hunter, L. The Use of Explicit Goals for Knowledge to Guide Inference and Learning. To appear in *Applied Intelligence*. 2(1).

Schank, R. C. *Explanation Patterns*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

Shavlik, J. W., & Towell, G. G. An Approach to Combining Explanation-Based and Neural Learning Algorithms. *Connection Science*. 1(3), 1989.