Introspective reasoning using meta-explanations for multistrategy learning*

Ashwin Ram and **Michael T. Cox**

College of Computing
Georgia Institute of Technology
Atlanta, Georgia 30332-0280
ashwin@cc.gatech.edu and cox@cc.gatech.edu

Abstract

In order to learn effectively, a reasoner must not only possess knowledge about the world and be able to improve that knowledge, but it also must introspectively reason about how it performs a given task and what particular pieces of knowledge it needs to improve its performance at the current task. Introspection requires declarative representations of meta-knowledge of the reasoning performed by the system during the performance task, of the system's knowledge, and of the organization of this knowledge. This chapter presents a taxonomy of possible reasoning failures that can occur during a performance task, declarative representations of these failures, and associations between failures and particular learning strategies. The theory is based on Meta-XPs, which are explanation structures that help the system identify failure types, formulate learning goals, and choose appropriate learning strategies in order to avoid similar mistakes in the future. The theory is implemented in a computer model of an introspective reasoner that performs multistrategy learning during a story understanding task.

*Appears in R.S. Michalski and G. Tecuci (eds.), *Machine Learning: A Multistrategy Approach, Vol. IV*, pages 349–377, Morgan Kaufman Publishers, San Mateo, CA, 1994.

1 Introduction

It is generally accepted that learning is central to intelligent reasoning systems that perform realistic reasoning tasks, such as understanding natural language stories or solving non-trivial problems (e.g., Schank, 1983). It is impossible to anticipate all possible situations in advance and to hand-program a machine with exactly the right knowledge to deal with all the situations it might be faced with. Rather, during the performance of any non-trivial reasoning task, whether by humans or by machines, there will always be difficulties and failures. Intelligence lies in the ability to recover from such failures and, more importantly, to learn from them so as not to make the same mistake in future situations. In addition, the reasoner needs to identify what it needs to learn, and to focus its learning in order to avoid the combinatorial explosion of inferences and search necessary in complex, unrestricted situations.

A general theory of multistrategy learning must provide a taxonomy of the types of reasoning situations that provide an opportunity to learn, strategies for learning in different situations, and methods for selecting appropriate learning strategies based on what needs to be learned. The central claim of this chapter is that, in order to learn effectively, a system must not only possess knowledge about the world and be able to improve that knowledge, but it also must introspectively reason about how it performs a given task and identify what particular pieces of knowledge it needs to improve its performance at the current task. To see why this must be so, consider three fundamental problems in learning.

Although several researchers have attempted to develop algorithms for machine learning that can acquire knowledge through the generalization of input examples or experiences, little effort has been made to formulate a general solution to the problem of identifying what needs to be learned in the first place. When a reasoner encounters a problem, a contradiction, or an unexpected event, it is clear there is a need to learn some piece of knowledge that, had it been present, would have prevented such a problem from occurring. But from where does the problem arise? Perhaps the information the reasoner was provided to base its conclusion on was incomplete, or perhaps its own knowledge used to interpret such information was incomplete. Perhaps the reasoner used an inappropriate piece of knowledge for the current situation. Perhaps the reasoner chose the wrong strategy to, say, verify a hypothesis, or looked in the wrong place to gather evidence. Identifying the cause of a reasoning failure (or the cause of an expected success) is known as the *blame (or credit) assignment problem* (e.g., Hammond, 1989; Minsky, 1963; Weintraub, 1991).

Traditional approaches in machine learning have assumed that the knowledge to be learned has already been identified by an external agent. In Winston's (1975) arch learning program, for example, the user decided that the concept of an arch was a useful concept to teach the program. Mitchell, Keller, and Kedar-Cabelli's (1986) explanation-based generalization algorithm relies on a "target concept" to be learned, which is supplied as input to the algorithm. In some approaches, the learning process has no target or goal at all; the program has no sense of what it is trying to learn or why it is trying to learn it. Recently, some researchers have argued that the identification of what might be called the "learning goals" of the reasoner is an important aspect of the learning problem (e.g., Hunter, 1990b; Michalski, 1993; Ram, 1991; Ram & Hunter, 1992; Ram & Leake, 1993). This view is consistent with psychological data on question asking in educational contexts (e.g., Scardamalia and Bereiter, 1991), and on goal orientation in learning (e.g., Barsalou, 1991; Ng & Bereiter, 1991) and in focus of attention and inferencing (cf. Zukier's 1986 review). In the authors' own research, it is argued that active, goal-based learning is important for both computational reasons and for cognitive reasons (Ram, 1991; Ram & Hunter, 1992). Furthermore, it is argued that learning goals should not be thought of as external inputs to a learning program, but rather should be generated by the program itself (see also Mitchell, Utgoff, & Banerji, 1983; Laird, Rosenbloom & Newell, 1986).

Given that the reasoner knows that an error occurred and what type of error it is, the second step following blame assignment is to determine what needs to be learned. This is referred to as the problem of *deciding what to learn* or, equivalently, the problem of formulating learning goals to pursue. For any given failure, there could be a large set of possible lessons to be drawn. Choosing the appropriate one depends not only on the type of the reasoning failure, but also on the prior goals and tasks of the reasoner and the current state of its knowledge. Often, a desired piece of knowledge will not be immediately available in the input. In such cases, the reasoner must be able to suspend its learning goals, and reactivate them later when an appropriate opportunity arises.

Finally, if the reasoner knows what to learn, it still needs to identify what method is best suited for performing the desired learning. In many cases, a combination of learning strategies is necessary. For example, if the reasoner is presented with a novel explanation for a problem, it needs to be able both to acquire such an explanation in a general way (explanation generalization) and to remember it again in future situations in which it is likely to be applicable (index learning). Furthermore, a single learning strategy may be applicable in a number of different reasoning situations. For example, the reasoner may need to learn a new index to an explanation, both when the explanation is newly acquired and when the explanation is already known but incorrectly indexed in memory. Identifying appropriate learning strategies is called the *strategy selection problem*, and is particularly important in multistrategy learning systems (e.g., Cox & Ram, 1992; Hunter, 1990a; Reich, 1993).

Machine learning research, for the most part, has focussed on the development of learning algorithms for different situations. Most systems arising from this research are based on the application of a learning algorithm to a well-defined learning problem in the domain of interest. When one attempts to extend these systems to incorporate multiple learning strategies, one runs into an interesting problem: these systems cannot make decisions concerning which learning strategies to use in different circumstances. These decisions are made ahead of time by the system designer. Furthermore, it is difficult to extend these systems to allow them to make their own decisions about which learning strategies to pursue, because the knowledge used to make these decisions is buried in the procedures that make up the system. This further complicates the strategy selection problem.

These issues are addressed using the following approach. First, classes of learning situations are identified based on an analysis of the types of reasoning failures that occur. Next, each type of reasoning failure is characterized by a description of how the conclusions were drawn (a description of a chain of reasoning led up to those conclusions), why these conclusions were drawn (a description of the bases for the processing decisions underlying that chain of reasoning), why the conclusions were faulty (an explanation of why the drawn conclusions were incorrect), what the correct conclusions ought to have been (a description of the desired conclusions), and how the reasoner should have drawn them (a description of a chain of reasoning that would lead to the desired conclusions).[1] Finally, each type of reasoning failure is associated with the learning that needs to occur to avoid such a failure and the strategies that can perform the desired learning.

This information is represented explicitly in the system using a meta-model describing the reasoning process itself. In addition to the world model that describes its domain, the reasoning system has access to meta-models describing its reasoning processes, the knowledge that this reasoning is based on, and the indices used to organize and retrieve this knowledge. A meta-model is used to represent the system's reasoning during a performance task, the decisions it took while performing the reasoning, and the results of the reasoning. If a difficulty or failure is encountered, the system introspectively examines its own reasoning processes to determine where the problem lies, and uses this introspective understanding to improve itself using the appropriate learning strategies.

Many research projects in AI have demonstrated the advantages of representing knowledge about the world in a declarative manner. Using such knowledge, a reasoner can create an explicit representation of events in the world. Such data structures can be used to reason about the world in a systematic fashion. Similarly, it is claimed here that declarative knowledge about reasoning can be beneficial in reasoning about one's own thoughts. Just as reasoning involves the processing of explicit structures representing physical events in the world, introspection involves the processing of explicit structures representing mental events in the head.

Introspective reasoning of this type is central to the solution of the credit/blame assignment problem, the problem of deciding what to learn, and the strategy selection problem and, hence, to learning. Furthermore, introspective reasoning not only guides learning about the world, but can also help the reasoner improve its own reasoning processes. Because the decisions made when choosing reasoning strategies, the results of these decisions, and the trace and final outcome of the reasoning process are all made explicit, the reasoner can learn to make better decisions by analyzing its own behavior and, hence, improve its own reasoning. This allows the reasoner to improve the decisions that underlie

[1] In general, a complete explanation of a reasoning failure would have all these components. In any given situation, however, the reasoner might only be able to construct a partial explanation, which would then determine what the reasoner can learn from that experience.

reasoning and learning and, thus, provides a method for "learning to learn." As noted by Minsky (1985) and others, such learning is the foundation upon which to develop truly intelligent systems.

2 Overview of the approach

This chapter proposes a theory of multistrategy learning in which the reasoner models its own reasoning processes explicitly, and analyzes this model after a reasoning experience in order to identify what it needs to learn and to select the appropriate learning strategy from a set of available strategies. The introspective analysis is done using a meta-model of the reasoner's own knowledge and reasoning processes. A theory of content and representation of these meta-models is presented along with a computer model, Meta-AQUA, that implements this theory for a story understanding task.

The theory focuses on failure-driven learning. The term "failure" includes not simply performance errors, but also expectation failures (Schank, 1986), anomalous situations that the reasoner failed to predict, and other types of reasoning failures. Unlike successful processing where there may or may not be anything to learn, failure situations are guaranteed to provide a potential for learning; otherwise, the failure would not have occurred (Minsky, 1985).[2] When a reasoning failure occurs, the system posts a *knowledge goal* which drives the reasoner to explain or otherwise resolve the gaps in its knowledge. Knowledge goals, often expressed as questions, represent the reasoner's goals to learn (Ram, 1989, 1991; Ram & Hunter, 1992). In order to learn from the failure and to avoid repeating the same mistake in the future, the system needs to identify the cause of the failure and then, depending upon the cause, apply a given learning strategy.

The key representational entity in the theory is a *meta-explanation pattern* (Meta-XP), which is a causal, introspective explanation structure that explains how and why an agent reasons, and that helps the system in the learning task. The theory of reasoning and learning is based on these structures. There are two broad classes of Meta-XPs. *Trace Meta-XPs* record a trace of the reasoning performed by a system, along with causal links that explain the decisions taken. *Introspective Meta-XPs* are structures used to explain and learn from a reasoning failure. They associate a failure type with a particular set of learning strategies and point to likely sources of the failure within the Trace Meta-XP. Note that Trace Meta-XPs explain how a system draws its conclusions, while Introspective Meta-XPs explain why these conclusions fail in particular situations.

Meta-XPs form the basis for the introspective reasoning necessary for experience-based, goal-directed learning in situations when the reasoner encounters a reasoning failure. Consider the following three types of reasoning failures:

- **Novel situation:** An expectation failure can arise when the reasoner does not have the appropriate knowledge structures to deal with a situation. The situation is said to be anomalous with respect to the current knowledge in the system. In such a situation, the reasoner could use a variety of learning strategies, including explanation-based generalization (DeJong & Mooney, 1986; Mitchell, Keller, & Kedar-Cabelli, 1986), inductive generalization from input examples (Michalski, 1983), and explanation-based refinement (Ram, 1993), coupled with index learning to place the new structures appropriately in memory.

- **Incorrect background knowledge:** Even if the reasoner has knowledge structures that are applicable to the situation, these knowledge structures may be incomplete or incorrect. Learning in such situations is usually incremental, and involves strategies such as elaborative question asking (Ram, 1993) applied to the reasoning chain and abstraction, generalization and specialization techniques in conceptual memory (Michalski, 1993).

- **Mis-indexed knowledge structure:** The reasoner may have an applicable knowledge structure, but it may not be indexed in memory such that it can be retrieved using the particular 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,

[2] In Meta-AQUA, an unexpected success also counts as a reasoning "failure" because the reasoner was unable to correctly predict the outcome of the task.

the reasoner retrieves a structure that later proves inappropriate, it must specialize the indices to this structure so the retrieval will not recur in similar situations. Learning the right indices to organize knowledge in memory is known as index learning (e.g., Hammond, 1989; Bhatta & Ram, 1991; Ram, 1993).

To learn from such failures, the reasoning system uses a multistrategy learning approach in which it records and analyzes a declarative trace of its own reasoning process using a Trace Meta-XP. The data structure holds explicit information concerning the manner in which knowledge gaps are identified, the reasons why particular hypotheses are generated, the strategies chosen for verifying candidate hypotheses, and the basis for choosing particular reasoning methods for each of these. If the system encounters a reasoning failure, it then uses Introspective Meta-XPs to examine the declarative reasoning chain. An Introspective Meta-XP performs three functions: (1) it aids in blame assignment (determining which knowledge structures are missing, incorrect, or inappropriately applied); (2) it aids in the formulation of appropriate knowledge goals to pursue; and (3) it aids in the selection of appropriate learning algorithms to recover and learn from the reasoning error. Such meta-explanations augment a system's ability to introspectively reason about its own knowledge, about gaps within this knowledge, and about the reasoning processes that attempt to fill these gaps. The use of explicit Meta-XP structures allow direct inspection of the reasons by which knowledge goals are posted and processed, thus enabling a system to improve its ability to reason and learn.

3 Example: The Drug Bust

To instantiate and test the theory, an introspective version of the AQUA system (Ram, 1989, 1991, 1993) called Meta-AQUA has been implemented. AQUA is a question-driven story understanding system that learns about Middle Eastern terrorist activities. Its performance task is to "understand" the story by building causal explanations that link the individual events in the story into a coherent whole, and by building motivational explanations of the actions observed in the story that causally relate the actions to the goals, plans, and beliefs of the actors and planners of the actions. Although AQUA is a general model of case-based learning with multiple learning methods (see Ram, 1993), it falls short of being a general model of multistrategy learning, as defined in this chapter because the knowledge underlying the selection of appropriate learning methods in different situations is not explicitly represented in the system. The Meta-AQUA system extends AQUA by adding a model of introspective reasoning and multistrategy learning using Meta-XP structures. Unlike AQUA, Meta-AQUA does not actually parse the sentences; because this research does not deal with the natural language understanding problem, it is assumed that input sentences are already represented conceptually. To illustrate the type of introspection Meta-AQUA performs and the type of learning that results, consider the following story:

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: The authorities arrested the passenger, charging him with smuggling drugs.

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

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

The line of reasoning that Meta-AQUA pursues in processing this story is as follows. S1 produces no inferences other than the observation that 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 the bark is animate. In this example, the program (incorrectly) believes that dogs bark only when threatened by animate agents. because luggage is an inanimate object,

there is a contradiction, leading to a reasoning failure. This anomaly causes the understander to ask why the dog barked at an inanimate object. This question may lead the system to learn something useful about dogs at some point in the future. Until this question is answered, however, the system can only assume (again, incorrectly) that the luggage somehow threatened the dog.

S3 asserts an arrest scene that reminds Meta-AQUA of a prior incident of weapons smuggling by terrorists. The system then infers the existence of 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 the dog bark?" to be retrieved, and the understanding task is resumed. Instead of revealing the anticipated threatening situation, S4 produces a competing hypothesis. The program prefers the explanation given by S4 over the earlier one because it links more of the story together (e.g., see Alterman, 1985; Ng & Mooney, 1990; Thagard, 1989). The system uses the trace of its reasoning process, stored in a Trace Meta-XP, to review the understanding process. It characterizes the reasoning error as one in which there is (1) an expectation failure caused by the incorrect retrieval of a known explanation ("dogs bark when threatened by animate objects," erroneously assumed to be applicable), and (2) 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 the Introspective Meta-XP `XP-Novel-Situation-Alternative-Refuted`.

This composite Meta-XP characterizes a common class of reasoning errors and consists of three basic Meta-XPs: `XP-Novel-Situation`, `XP-Mis-Indexed-Structure`, and `XP-Incorrect-Background-Knowledge`. `XP-Novel-Situation` directs an explanation-based generalization algorithm to be applied to the node representing the explanation of the bark. Because the detection scene of the drug-bust case and the node representing the sniffing are unified because of the explanation given in S4, the explanation is generalized to drug busts in general. The general explanation is then indexed in memory using an index learning algorithm. `XP-Mis-Indexed-Structure` directs the indexing algorithm to the defensive barking explanation. 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 is specialized so that retrieval occurs only for animate agents, not physical objects in general. Finally, `XP-Incorrect-Background-Knowledge` directs the system to examine the source of the story's anomaly. The solution is to alter the conceptual memory representation so that the constraint on the object of dog-barking instantiations is abstracted to physical objects, not just animate agents.

Though the program is directly provided with an explanation that links the story together, Meta-AQUA performs more than mere rote learning. It learns to avoid the mistakes made during the processing of the story. The application of Meta-XPs allows the system to use the appropriate learning strategy (or, as in the above example, multiple strategies) to learn exactly that which the system needs to know to process similar situations in the future correctly. This is essentially a case-based or experience-based approach, which relies on the assumption that it is worth learning about one's experiences because one is likely to have similar experiences in the future (see, e.g., Hammond, 1989; Kolodner & Simpson, 1984; Ram, 1993; Schank, 1982).

4 The explanation-based understanding task

The central task in Meta-AQUA is to build causal explanations that provide conceptual coherence to the story by linking the pieces of the story together. The approach to explanation construction is case-based, and is based on Schank's theory of explanation patterns (XPs) in which explanations are built by applying known XPs to the events in the story (Schank, 1986; Ram, 1990a, in press). Expectation failures arise when the world differs from the system's expectations. For example, the system may be faced with an anomalous situation in which the XP that the system believes to be applicable turns out to be contradicted in the story. When the system encounters an anomalous situation, it tries to retrieve and apply a known explanation to the anomalous concept. The process of explanation generates

questions, or knowledge goals, representing what the system needs to know in order to be able to explain similar situations in the future, thus avoiding repeated similar failures (Ram, 1991, 1993).

Explanation patterns are similar to justification trees, linking antecedent conditions to their consequences. An XP is essentially a directed graph of concepts, connected with `results`, `enables` and `initiates` links. A `results` link connects a process with a state, while an `enables` link connects a precondition state to a process. An `initiates` link connects two states. These three links are sufficient to represent goal-based stories of the type discussed here, although other causal links (e.g., `disenables`) may be added without invalidating the approach.

The set of sink nodes in the graph is called the PRE-XP-NODES. These nodes represent what must be present in the current situation for the XP to apply. One distinguished node in this set is called the EXPLAINS node. It is bound to the concept that is being explained. Source nodes are termed XP-ASSERTED-NODES. All other nodes are INTERNAL-XP-NODES. For an XP to apply to a given situation, all PRE-XP-NODES must be in the current set of beliefs. If they are not, then the explanation is not appropriate to the situation. If the structure is not rejected, then all XP-ASSERTED-NODES are checked. For each XP-ASSERTED node verified, all INTERNAL-XP-NODES connected to it are verified. If all XP-ASSERTED-NODES can be verified, then the entire explanation is verified. Gaps in the explanation occur when one or more XP-ASSERTED-NODE remains unverified. Each gap results in a question or knowledge goal, which provides the system with a focus for reasoning and learning.

The background knowledge used in the current implementation consists of a frame-based conceptual hierarchy, a case library of past episodes, and an indexed collection of XPs. For the task of story understanding, Meta-AQUA employs the algorithm outlined in figure 1. First, the outer loop inputs a sentence representation and checks to see if the concept can answer a prior question. If it can, the reasoning associated with the question is resumed. Otherwise, the concept is passed on to the understanding algorithm. The understanding algorithm consists of four phases: (1) question identification, (2) hypothesis generation, (3) verification, and (4) review/learning.

The first phase looks for questions associated with the concept by checking the concept for interesting characteristics. Meta-AQUA considers explanations, violent acts, and anomalies to be interesting.[3] Explanations and violent acts are detected by the concept type of the input. Anomaly detection is performed by comparing the input to the conceptual definitions found in the conceptual hierarchy. If a concept contradicts a constraint, then a constraint anomaly exists, and a question is posed. Such questions represent the knowledge goals of the program. If no anomaly is detected, then the concept is instantiated, and control passes back to the top level.

If a knowledge goal is posted, then the understander attempts to answer the question by generating a hypothesis. The basis of this decision, i.e., what knowledge is relevant in making the determination, is then recorded in the Trace Meta-XP. Strategies for hypothesis generation include application of known explanation patterns (Kass, Leake & Owens, 1986; Ram, 1990a, in press; Schank & Leake, 1990), case-based reasoning (e.g., Hammond, 1989; Kolodner & Simpson, 1984; Ram, 1993, in press), and analogy (e.g., Falkenhainer, 1990; Kedar-Cabelli, 11988). If none of these applies, then the process is suspended until a later opportunity.

When a hypothesis is generated, it is passed to the verification subsystem. Strategies for hypothesis verification include the devising of a test (currently not implemented; see, e.g., Rajamoney, 1989), comparison to known concepts, and suspension of the reasoning task until further information is available (Ram, 1991).

The system reviews the chain of reasoning after the verification phase is complete. The review process examines the Trace Meta-XP to check whether there was a reasoning failure. If a failure occurred, the review process searches for an introspective explanation. If an Introspective Meta-XP is retrieved, it is applied to the failure. Meta-XP application is analogous to XP application. Meta-AQUA first checks the PRE-XP-NODES to determine if the Meta-XP is applicable to the failure situation. If the Meta-XP is applicable, the XP-ASSERTED-NODES of the Meta-XP are checked to see if they are in the set of current beliefs. If so, the learning algorithm(s) associated with the Meta-XP is executed. If there are XP-ASSERTED-NODES not in the set of current beliefs, then a question is posed on the Meta-XP itself.

Because learning is moderated by the XP application algorithm, it is necessary to represent the explanation-based understanding process outlined above in a declarative manner. This allows matching and syntactic functions to be

[3] A better approach to determining interestingness, based on the goals of the reasoner, is discussed by Ram (1990b).

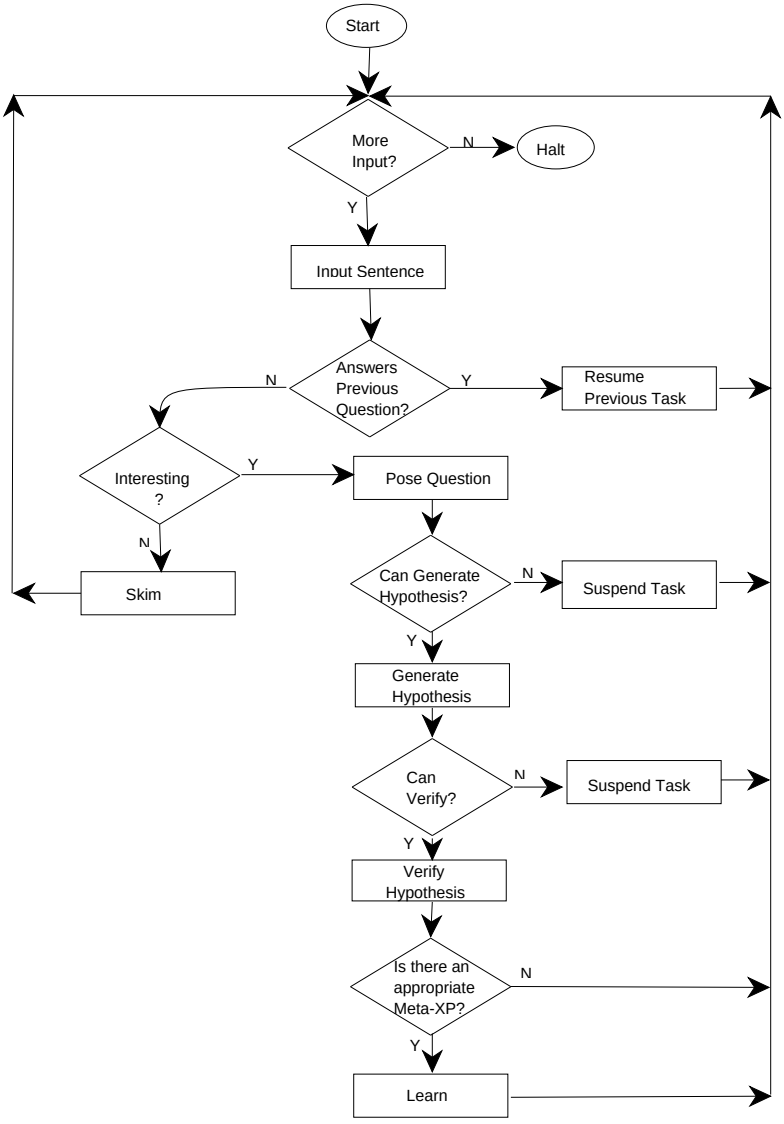

Figure 1: Meta-AQUA control flow

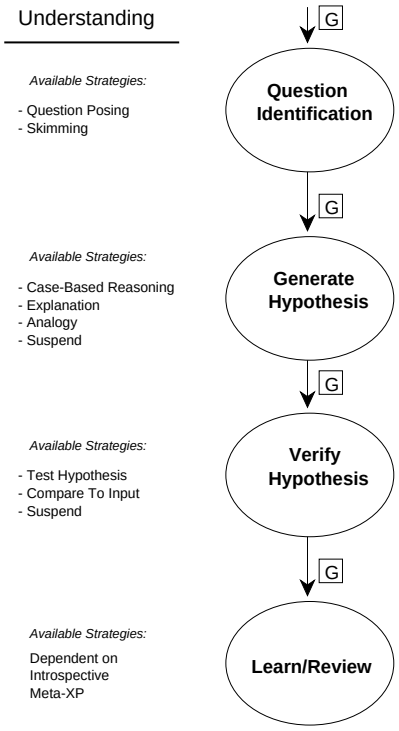

Figure 2: Phases of understanding: Decide-compute chains represented in Trace Meta-XPs

applied to the prior reasoning. Further, it allows the system to pose knowledge goals about aspects of the reasoning process itself.

5 Representation of Trace Meta-XPs

The AQUA system embodies a theory of motivational explanation based on decision models (Ram, 1990a, in press), which represent the decision process that an agent goes through in deciding whether to perform an action. The agent considers its goals, goal priorities, and the expected outcome of performing the action, and then decides whether to perform the action. Meta-AQUA extends the model to account for introspective reasoning about knowledge goals. A set of states, priorities, and the expected outcome of a reasoning strategy prompt the reasoner to make a strategy decision. Based on its general knowledge, inferences that can be drawn from this knowledge, and the current representation of the story, the reasoner chooses a particular reasoning strategy. Once executed, a strategy may produce further questions and hypotheses. Each execution node explicitly represents its main result (the structure returned by the function) and its side-effect.

These decide-compute combinations are chained into threads of reasoning such that each one initiates the goal that drives the next. Though the chains can vary widely, the chains for the task of question-driven story understanding take the form shown in figure 2. This reasoning process is recursive in nature. For example, if a hypothesis generates a new question, then the reasoner will spawn a recursive regeneration of the sequence. When insufficient knowledge exists on which to base a decision, a useful strategy is to simply defer making the decision. The reasoning task is suspended and later continued if and when the requisite knowledge appears. This is a form of opportunistic reasoning (Birnbaum & Collins, 1984; Hammond, 1988; Hayes-Roth & Hayes-Roth, 1979; Ram, 1989, 1991).

A Trace Meta-XP, representing the trace of the reasoning process, is a chain of decide-compute nodes (D-C-NODES). These nodes record the processes that formulate the knowledge goals of a system, together with the results and reasons for performing such mental actions. As such, the trace of reasoning is similar to a derivational analogy trace as described by Carbonell (1986). Such a Meta-XP is a specific explanation describing the reasoner's choice of a particular reasoning method or strategy, and the results of executing that strategy. Like an XP, the Meta-XP can be a general structure applied to a wide range of contexts, or a specific instantiation that records a particular thought process. One distinguishing property of Trace Meta-XPs is the fact that a decision at one stage is often based on features in previous stages. For example, the decision of how to verify a hypothesis may be based on knowledge used to initially construct the hypothesis. This property is particularly true of the learning stage, which by definition is based on prior processing.

An understanding system may attempt to retrieve and apply a Meta-XP for reasoning and learning in much the same way that regular XPs are used for explanation. If the antecedent conditions of the Meta-XP exist, then the structure will point to an appropriate learning algorithm without having to analyze all current states in the story representation. This approach provides significant speedup in learning, relying on past successes and failures instead of reasoning from first principles. For example, even though some subquestions on an erroneous hypothesis are verified, Meta-XPs will direct the search for the blame on the basis of the decision to use a given hypothesis generation strategy, not on the basis of the verification strategy.

6 Representation of Introspective Meta-XPs

An Introspective Meta-XP is a data structure used to explain why a particular solution or conclusion fails and to learn from a reasoning failure. Thus an Introspective Meta-XP performs three functions: it aids in (1) assigning blame (determining which knowledge structures are missing, incorrect, or inappropriately applied); (2) determining what knowledge goals the system should formulate; and (3) selecting appropriate learning algorithms to recover and learn from the reasoning error. These functions are accomplished by associating a failure type with likely causes of the reasoning failure and a particular set of learning strategies used to ensure that the failure does not recur. Introspective Meta-XPs have the following components:

- Failure type (one of the types of failures in the taxonomy below)
- Explanation of failure (graph structure of the failure)
- Pointers to locations in the graph likely to be responsible for the failure
- Specification of knowledge goals to be created (to prevent failure from recurring)
- Temporal information for the explanation (temporal orderings on the nodes and the links in the graph)

The graph structure of the Introspective Meta-XP is similar to that of a justification tree or an explanation pattern (Schank, 1986; Ram, 1990a, in press), linking antecedent (mental) conditions to their (mental) consequences. The major difference between a Meta-XP and an XP is that, although both are explanatory causal structures, an XP proposes a causal justification for a physical observation (such as why blowing up a bomb causes objects in its vicinity to be destroyed) or a volitional action (such as why a Lebanese teenager would volunteer for a suicide bombing mission in which he was sure to be killed), whereas a Meta-XP explains how and why an agent reasons in a particular manner. Thus, the representation of a Meta-XP must be able to account for reasoning failures and successes. The three types of reasoning failures discussed in the introduction (novel situations, incorrect or incomplete background knowledge, and mis-indexed knowledge structures) can be represented using the following basic types of successes and failures.

6.1 Successful prediction

The basic types of failures that make up the components of an Introspective Meta-XP graph are expectation failure, retrieval failure, and incorporation failure. A fourth type, successful prediction, does not produce any learning in

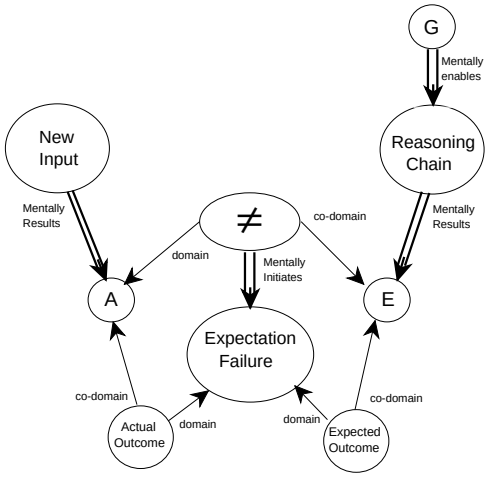

Figure 3: Expectation failure: A $\neq$ E.

Meta-AQUA, but is represented as a mental event in the same manner as failures. To illustrate the representation, let node A represent an actual occurrence of an event in the world, an explanation, or an arbitrary proposition. Let node E represent the expected occurrence. The node A `mentally-results` from either a mental calculation or an input concept. 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 is a superset of E, then a successful prediction has occurred. If on the other hand, A is a subset of E, then there are more questions remaining on the predicted node E. In such cases, Meta-AQUA waits for more information before it introspects.

6.2 Expectation failure

Failures occur when A $\neq$ E. This state exists when either A and E are disjoint, or there are conflicting assertions within the two nodes. For example, A and E may represent persons, but E contains an attribute description specifying gender = male, whereas A contains the attribute description gender = female. Inferential expectation failures occur when the reasoner predicts one event or feature, but another occurs instead. The representation of an expectation failure is shown in figure 3. The EXPLAINS node of the Meta-XP (the distinguished node that the Meta-XP explains) is the one marked "Expectation Failure". The system's awareness of the expectation failure is `mentally-initiated` by the not-equals relation between A and E.

6.3 Retrieval failure

Retrieval failure has a similar structure, although the difference here is that instead of an expectation (E) being present, it is instead absent because of the inability of the system to retrieve the knowledge structures that would predict E (see figure 4). To represent these conditions, Meta-AQUA uses standard non-monotonic logic values of `in` (in the current set of beliefs) and `out` (out of the current set of beliefs) (Doyle, 1979), augmented with `hypothesized-in` (weakly assumed in), `hypothesized-out` (weakly assumed out), and `hypothesized` (unknown) (Ram, 1989). 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 frame is `out`.

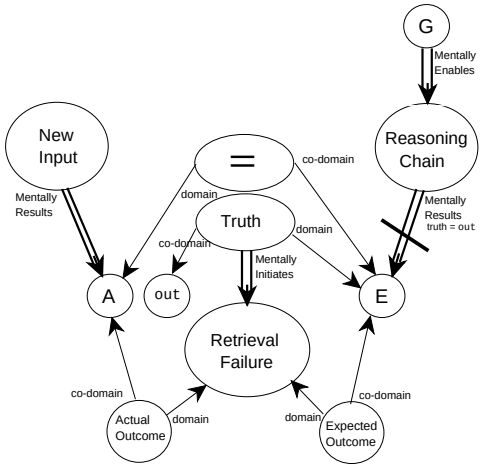

Figure 4: Retrieval failure: A [truth = in] = E [truth = out].

6.4 Incorporation failure

An incorporation failure occurs when the reasoner fails to incorporate a fact or assertion in the conceptual representation of the input into its background knowledge. For example, a fact may contradict a belief that the system holds. The difference between an incorporation failure and an expectation failure is that in the latter, the system has an explicit expectation that is violated, whereas in the former, the new input is presented before the contradictory belief is explicitly identified by the system.

Incorporation failures represent an important class of anomalies that provide a basis for identifying knowledge goals that drive reasoning and learning. For example, a constraint anomaly occurs when an attribute of the conceptual representation of some proposition conflicts with a constraint defined in the background knowledge. The conflict produces a not-equals relation between the actual occurrence and the conceptual constraint. This relation `mentally-initiates` the anomaly (see figure 5).

7 Associating knowledge goals and learning strategies with Introspective Meta-XPs

Based on the above representations of basic types of failures, the three types of reasoning failures discussed in section 2 can be represented, along with associated knowledge goals and learning strategies for these situations, using Introspective Meta-XP structures.

7.1 Mis-indexed knowledge structure

This type of failure occurs when the reasoner's knowledge structures are not indexed in memory in a manner that allows them to be retrieved in appropriate situations using the particular cues provided by the context. There are two variants of this type of failure, represented by the Introspective Meta-XPs `XP-Erroneous-Association` and `XP-Missing-Association`. `XP-Erroneous-Association` is based on an expectation failure structure, because this situation arises when an index associates a contextual state with a knowledge structure in the background knowledge that produces incorrect inferences in this context. In such cases, a knowledge goal is spawned to modify the index so that the index will still allow Meta-AQUA to retrieve the knowledge structure when appropriate, but not in future instances similar to the current situation in which it was erroneously retrieved. (Because this leads to

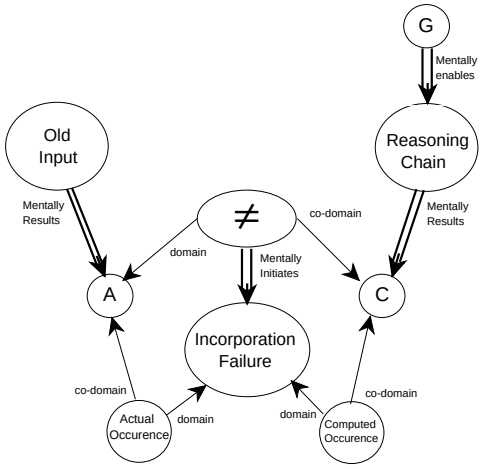

Figure 5: Incorporation failure: Contradiction, constraint conflict, or other anomaly.

existing knowledge being reorganized rather than new knowledge being learned, this type of knowledge goal is called a *knowledge organization goal* and the complementary type a *knowledge acquisition goal*.) The learning strategy used for such a goal is to execute a specialization algorithm that produces a more discriminating index. Because the knowledge organization goal has pointers to a declarative representation of the reasoning which produced the goal, the learning strategy has access to the context of the error.

Indexing failures because of missing associations are represented using `XP-Missing-Association`, which is based on a retrieval failure structure. This situation arises when an appropriate knowledge structure was not retrieved because there was no index to associate the context with that structure. This knowledge structure is represented by a node M in the system's background knowledge that must be `in`. The knowledge goal associated with the Introspective Meta-XP is to find the node M. If the knowledge goal is achieved, the associated learning strategy directs an indexing algorithm to examine the indices of M, looking for an index compatible with the index calculated for A, the node representing the input assertion currently being processed. If found, this index is generalized so that the current cues provided by the context of A will allow Meta-AQUA to retrieve E. If no such index is found, a new index is created. Furthermore, if an appropriate knowledge structure 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 until such time as M can be found.

7.2 Novel situation

Novel situations are structurally similar to the missing association variant of the mis-indexed structure situation, except that the node M (and thus its associated index I) has a truth value of `out`. In other words, a novel situation is one in which there is no knowledge structure in memory that can be retrieved and reasoned with to allow the system to expect a concept that matches A.

Novel situations, represented by `XP-Novel-Situation`, occur when M is missing (`truth` = `out`) and the E node's truth value is either `hypothesized-in` or `out`. When a novel situation is identified, Meta-AQUA creates a knowledge acquisition goal to learn a new explanation from the event, and a knowledge organization goal to learn an appropriate index for the new explanation. The learning strategy used in this case is to perform explanation-based generalization on the node A to create a knowledge structure that can be applied to a wider set of future situations. The strategy also directs an indexing algorithm to the newly created explanation so that it will be retrieved in similar situations in the future.

7.3 Incorrect background knowledge

A third class of reasoning failures arises from incorrect background knowledge, which occurs when the system's world model is not accurate or consistent with the real world. In the current implementation, only one situation of type `XP-Incorrect-Background-Knowledge` is represented. This situation arises when there is an inconsistency with a known fact and a constraint specified in the background knowledge (for example, a constraint in the ISA-hierarchy), leading to an incorporation failure. In this situation, Meta-AQUA spawns a knowledge goal to modify the erroneous constraint in memory. The learning strategy in this case is to check whether the two assertions (the fact and the constraint) are conceptual siblings. If this is true, Meta-AQUA performs abstraction on the constraint, generalizing it to its parent node. The constraint is then marked as being `hypothesized-in` on the basis of induction. The reasoning chain that led to this hypothesis is indexed off the hypothesis so that it can be retrieved when the constraint is invoked in future situations. The hypothesis is verified if the anomalous assertion is re-encountered in later situations.

7.4 Combinations and extensions

An additional class of reasoning failures arises from the inferences used to base a decision on during the hypothesis generation phase. The error is found by searching all hypothesis generation D-C-NODES on the path from the EXPLAINS node of A to the node E, performing elaborative question asking (Ram, 1991, 1993). This case has not yet been represented declaratively. Meta-AQUA reasons about it using a general search heuristic for blame assignment.

Figure 6 shows the composite Meta-XP that is used to direct learning in the drug bust example from section 3. The Meta-XP combines an `XP-Novel-Situation`, an `XP-Mis-Indexed-Structure`, and an `XP-Incorrect-Background-Knowledge`. A2, the actual outcome, is bound to the explanation from S4, whereas E, the expected outcome, is bound to the explanation that dogs bark at objects that threaten them. C is bound to the constraint that dogs bark at animate objects. The index in memory, I, is bound to the index used to retrieve the abstract explanation instantiated as E.

In general, Introspective Meta-XPs are built out of reasoning chains involving successful predictions, expectation failures, retrieval failures, and incorporation failures. Table 1 illustrates some of the combinations that are representable using this set of building blocks, along with the associated learning strategies. Note that the node A is assumed `in` for all entries. In addition, for the two combination Meta-XPs in the table, E' represents the concept that should have been predicted but was not, and M' represents the memory item that should have been retrieved but was not.

8 Related work

Several past studies have called for meta-level reasoning (e.g., Collins & Birnbaum, 1990; Davis, 1980; Stefik, 1981; Wilensky, 1981), and several conferences have been held on the subject (e.g., Brazdil & Konolige, 1990; Maes & Nardi, 1988). However, traditional approaches to meta-reasoning stress knowledge about a system's own knowledge and reasoning in the form of rules dealing with belief, preference heuristics for operator selection, or constraints and defaults on the types of values an attribute may assume, and do not deal with the process of introspective reasoning itself. Some researchers (e.g., Maes, 1987; Pollock, 1989) distinguish between meta-level knowledge and introspection, that is, between knowledge about one's facts and knowledge about one's motivations and processes. It is argued here that introspective access to explicit representations of knowledge and of reasoning processes is necessary in a learning system that can make sophisticated decisions about what and how to learn, and that this ability, in turn, is an essential part of a general theory of multistrategy learning.

Carbonell's (1986) derivational analogy may be viewed as a form of introspective case-based reasoning. However, although his derivational analogy traces are similar to Trace Meta-XPs, there are several differences between the two approaches. The underlying reasoning processes represented in Trace Meta-XPs may be based on a reasoning model other than search-based problem solving. Trace Meta-XPs take into consideration competing reasoning goals and their

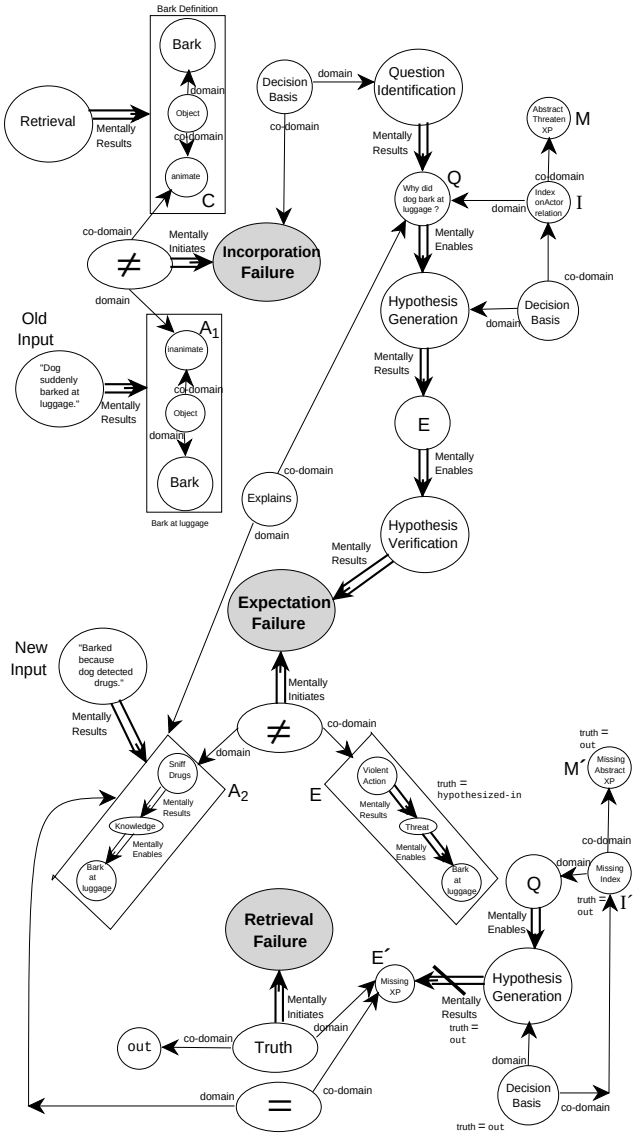

Figure 6: `XP-Novel-Situation-Alternative-Refuted`

	E	E′	M	M′	RC	SP	EF	RF	
Successful Prediction	in	ϕ	in	ϕ	in	in	out	out	No learning
Novel Situation	out	ϕ	out	ϕ	out	out	out	in	EBG on A Index A by context
Retrieval Failure	out	ϕ	in	ϕ	out	out	out	in	Generalize index on M
Novel Situation + Expectation Failure	hypo-in	out	in	out	out	out	in	in	EBG on A Index A by context Specialize index on M
Expectation Failure + Retrieval Failure	hypo-in	out	in	in	in	out	in	in	Specialize index on M Generalize index on M′

Table 1: Truth values of nodes in Introspective Meta-XP representations. ϕ = don't care; `hypo-in` = hypothesized-in; RC = reasoning chain; SP = successful prediction; EF = expectation failure; RF = retrieval failure.

relative priorities, and include a representation of the expected outcome of a processing decision. Finally, Meta-XP theory augments the idea behind derivational traces by adding Introspective Meta-XPs to represent common reasoning situations and reasoning failures, along with associated knowledge goals and learning strategies.

The meta-explanations in this approach are also similar to self-explanations (Chi & VanLehn, 1991; Pirolli & Bielaczyc, 1989; VanLehn, Jones, & Chi, 1992). This research in psychology shows that formulating self-explanations while understanding input examples significantly improves the subjects' ability to learn from the examples. One difference between the two approaches is that self-explanations are explanations about events and objects in the world, whereas meta-explanations are explanations about events and objects in the reasoner's "mind." Experimental results in the metacognition literature suggest that introspective reasoning of the kind proposed here can facilitate reasoning and learning. For example, in a study by Alexander (1992), gifted children were found to have a better understanding of their own mental activities. These children, called "attributors," tended to form mental and perceptual explanations about their own behavior. In another study, Carr (1992) showed that strategic, deliberative control of mental processes facilitated the use of decomposition strategies for a mathematics problem-solving task. The improved performance resulted from meta-knowledge about reasoning strategies and knowledge about when a strategy was appropriate to use.

Outside the introspection paradigm, the work presented here builds on previous research in several areas of machine learning, including case-based reasoning, explanation-based learning, and multistrategy learning. Many multistrategy learning systems are simply integrated systems consisting of a cascade of more than one learning algorithm (e.g., Ahmad, Matwin & Ould-Brahim, 1991; Flann & Dietterich, 1989; Hunter, 1993; Shavlik & Towell, 1989). The control is always the same for every input. Usually an initial learning technique is applied, and its output used as the input to the next algorithm. A new generation of systems use more sophisticated schemes, whereby one or many algorithms may apply to different inputs depending on the situation. In these paradigms, selection of the learning algorithm becomes computationally important. One of the greatest benefits of using Introspective Meta-XPs in this type of framework is their ability to apply learning tasks that are appropriate to a given situation without having to blindly search all possible learning choices. However, many non-cascaded multistrategy learning systems apply learning algorithms in a predefined order (e.g., Genest, Matwin & Plante, 1990; Pazzani, 1991). If the first fails, then the next strategy is tried, and so forth. Much effort may be wasted in worst-case scenarios. In the Meta-XP approach, multistrategy learning is viewed as a deliberative, planful process in which the system makes explicit decisions about what to learn and how to learn it. In this sense, the theory is similar to that of Hunter (1990b), with the difference that the approach presented here is based on introspective reasoning using explicitly represented meta-explanation structures.

One problem with many of the traditional approaches to machine learning is that they do not address the blame assignment problem and, therefore, the problem of what the reasoner needs to learn. The research presented here is based on the belief that these problems are an integral part of the learning process and, furthermore, that they should be addressed in a unified learning framework based on introspection and Meta-XPs. The approach is to use an analysis of reasoning failures encountered to determine what needs to be learned. In this respect, the approach is similar to that of Birnbaum, Collins, Freed & Krulwich (1990), Mooney & Ourston's (1993) EITHER system, and Park & Wilkins's (1990) ODYSSEUS and MINERVA systems, but with some important differences. ODYSSEUS only deals with what are called novel situations in Meta-AQUA; when it fails to explain an action, it always assumes that relevant facts or rules are missing from its knowledge base. MINERVA's failure types are closer in spirit to those of Meta-AQUA, but its method relies on an expert to guide the learning process. Furthermore, the types of failures and corresponding learning actions used in EITHER and ODYSSEUS/MINERVA are specific to a particular reasoning paradigm (logic-based deduction and rule-based expert systems, respectively), whereas Meta-AQUA's taxonomy of failure types is not specific to a particular type of reasoning method. Rather than characterize learning actions at the level of "add a new rule" or "generalize the antecedent of an existing rule," Meta-AQUA is based on characterizations of learning strategies such as "generalize a new explanation," which may in turn be implemented in different ways depending on the reasoning paradigm. In addition, none of these approaches deals with the issue of learning indices for knowledge. Birnbaum *et al.* (1990) focus on the process of blame assignment by backing up through justification structures, and not on the declarative representation of types of failures. Furthermore, they do not discuss the use of failure characterizations to select learning strategies in a multistrategy learning system. Finally, none of the above approaches uses declarative characterizations of reasoning failures to formulate explicit learning goals. Despite these differences, it should be emphasized that the above approaches have much in common between them.

Although active, goal-driven reasoning and learning has been studied in education, psychology and social cognition (e.g., Barsalou, 1991; Ng & Bereiter, 1991; Scardamalia & Bereiter, 1991; Zukier, 1986), very little previous work in machine learning has addressed the issue of learning goals (but see Hunter, 1990a, 1990b; Michalski, 1991; Ram, 1989, 1991; Ram & Hunter, 1992; Ram & Leake, 1993). The theory of knowledge goals provides the basis for an integrated, goal-based approach to reasoning and learning across a variety of task domains. Although some search-based learning methods do incorporate the notion of a "learning goal" as the desired target of a search (e.g., Mitchell, Utgoff, & Banerji, 1983; Laird, Rosenbloom & Newell, 1986), knowledge goals are, in contrast, viewed as the basis for deliberative, planful learning behavior. The notion of a "target concept" used some learning systems (e.g., Mitchell, Keller, & Kedar-Cabelli, 1986), again, is less flexible than the knowledge goals underlying the work presented here. A reasoner can actively reason about its knowledge goals and might, for example, decide not to pursue a knowledge goal, suspend a knowledge goal until a later opportunity to satisfy it occurs (perhaps unexpectedly), or explicitly reason about the relative priorities of pending knowledge goals or about methods of achieving knowledge goals.

From the system design point of view, the use of Meta-XPs in reasoning about knowledge goals during story understanding and problem-solving provides a number of benefits. Because Meta-XPs make the trace of reasoning explicit, an intelligent system can directly inspect the reasons supporting specific conclusions. This avoids hiding knowledge used by the system in procedural code. Instead, there exists an explicit declarative expression of the reasons a given piece of code is executed. With these reasons enumerated, a system can explain how it produced a given failure and can retrieve an introspective explanation of the failure. This approach provides a nice framework for integrated learning approaches in which reasons for processing and learning decisions are made explicitly by the system rather than existing only in the minds of the system's designers.

9 Discussion and future research

The use of introspection by applying Meta-XPs to declarative representations of the reasoning process can aid a reasoner's ability to perform blame assignment, and to direct the learning algorithms that allow the reasoner to recover from failures and to learn not to repeat the failure. The use of Meta-XP structures aids in the blame assignment problem, because all points in the reasoning chain do not have to be inspected. This helps in controlling the search

process. Because answers may not be available at the time questions are posed, an opportunistic approach allows the system to improve its knowledge incrementally and to answer its questions at the time the information it needs becomes available. The representation also allows the system to pose questions about its own reasoning.

The approach relies on a declarative representation of meta-models for reasoning and learning. There are several advantages of maintaining such structures in memory. First, because they represent reasoning processes explicitly, the system can directly inspect the reasons underlying a given processing decision it has taken, evaluate the progress toward a goal, and compare its reasoning to past instances of reasoning in similar contexts. Thus, these traces can also be used in credit/blame assignment, to evaluate how reasoning errors occurred, and to facilitate learning from these errors. Second, because both the reasoning process and the knowledge base are represented using the same type of declarative representations, processes that identify and correct gaps in a knowledge base can also be applied to the reasoning process itself. In other words, a knowledge goal, or a goal to learn, may be directed at the reasoning process as well as at the knowledge base. If causal representations underlying domain theories and introspective representations underlying reasoning processes are both declarative causal structures, the same types of reasoning and learning algorithms can be applied to both.

A third advantage is the potential for speedup learning that is provided by the declarative trace of past reasoning processes. One does not have to replicate the entire sequence of decisions in solving a current problem. Instead, the system may match a current context to past reasoning experiences to retrieve a "macro reasoning operator" to apply to the situation. In addition, related to the previous two points, the ability of a Meta-XP to provide pointers to applicable learning algorithms to be used in given circumstances provides a basis for multistrategy learning. Because Meta-XPs encapsulate reasoning experiences, they can also help the system select appropriate learning strategies based on an analysis of the difficulties encountered during these reasoning experiences. Finally, from a system design perspective, the Meta-AQUA architecture provides a uniform framework for the integration of multiple learning strategies into an intelligent system.

This chapter focussed on the representation and use of Trace and Introspective Meta-XP structures for multistrategy learning in three types of reasoning failure situations. The class of failures represented by `XP-Incorrect-Background-Model` is still under investigation, including the formulation of a representation for deciding when to use the heuristic search briefly mentioned in section 7. Other strategies need to be created as well. The task of knowing when an assertion is incorrect, not just incomplete, is a difficult but interesting research problem.

Additional types of situations that provide the reasoner an opportunity to learn are also being investigated. An important class of failures, for example, is those that occur when the reasoner selects an inappropriate reasoning strategy in a given situation. The Meta-AQUA system is being extended to learn control information by representing Meta-XPs that point to potential problems with the reasoning choices made in each phase. The failure type `Incorrect-Reasoning-Choice` occurs when the reasoner has an appropriate knowledge structure to reason with and index to the structure in memory, but incorrectly chooses the wrong knowledge because the reasoning method it decided to use turned out to be inappropriate or inapplicable. An analysis of the choice of reasoning methods will result in learning control strategies designed to modify the heuristics used in this choice.

Many machine learning systems assume noise-free input, and those that deal with noise seldom analyze the source or causes of the noise. A robust story understander should be able to reason about the validity of input concepts, including the possibility of intentional deception by characters in a story. The class of errors arising from input noise needs to be represented, and corresponding knowledge goals and learning strategies identified. Another interesting extension of this research currently being pursued is combining story understanding with problem solving. Declarative process representations similar to that of story understanding are being developed. Parallel to story understanding sequences of identify question $\Rightarrow$ generate hypothesis $\Rightarrow$ verify $\Rightarrow$ learn/review, problem-solving sequences would be represented as identify problem $\Rightarrow$ generate solution $\Rightarrow$ test $\Rightarrow$ learn/review. Meta-XPs would then be used to reason about and improve the problem-solving process of the reasoner.

10 Conclusions

Meta-AQUA is a computer model of a reasoner that is active and goal-driven, starting out with an incomplete understanding of a novel domain and learning through experience. Its learning goals are functional to the purpose of the system, and are identified during the pursuit of the performance task. The computer program models how an intelligent reasoner reasons about the best way to perform a task; introspectively analyzes its own successes and failures in performing its tasks; reasons about what it needs to learn, selects appropriate learning strategies to acquire that information; and invokes the learning algorithms which then cause the reasoner to acquire new knowledge, modify existing knowledge, or reorganize memory by re-indexing knowledge in memory. The program implements a theory of introspective reasoning using meta-explanations that model typical reasoning situations, types of reasoning failures, and applicable learning strategies. The theory is motivated by cognitive as well as computational considerations, and provides a framework for the development of integrated, multistrategy learning systems for real-world tasks.

Acknowledgments

Research for this article was supported by the National Science Foundation under grant IRI-9009710 and by the Georgia Institute of Technology.

References

Ahmad, A., Matwin, S., & Ould-Brahim, H. "Acquiring the Second Tier: An Experiment in Learning Two-Tiered Concepts," *Proceedings of the First International Workshop on Multistrategy Learning*, R.S. Michalski & G. Tecuci (eds.), pp. 419–426, Harpers Ferry, WV, 1991.

Alexander, J. "Metacognition and Giftedness," Paper presented at the Southeast Cognitive Science Conference, Georgia Institute of Technology, Atlanta, Georgia, January 1992.

Alterman, R. "A Dictionary based on Concept Coherence," *Artificial Intelligence*, Vol. 25, pp. 153–186, 1985.

Barsalou, L. "Deriving Categories to Achieve Goals," in *The Psychology of Learning and Motivation: Advances in Research and Theory*, Volume 27, G.H. Bower (ed.), Academic Press, New York, NY, 1991.

Bhatta, S., & Ram, A. "Learning Indices for Schema Selection," *Proceedings of the Florida Artificial Intelligence Research Symposium*, M.B. Fishman (ed.), pp. 226–231, Cocoa Beach, FL, 1991.

Birnbaum, L., & Collins, G. "Opportunistic Planning and Freudian Slips," *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*, pp. 124–127, Boulder, CO, 1984.

Birnbaum, L., Collins, G., Freed, M., & Krulwich, B. "Model-based Diagnosis of Planning Failures," *Proceedings of the Eighth National Conference on Artificial Intelligence*, pp. 318–323, Boston, MA, 1990.

Brazdil, P.B., & Konolige, K. *Machine Learning, Meta-reasoning, and Logics*, Kluwer Academic Publishers, Boston, MA, 1990.

Carbonell, J.G. "Derivational Analogy: A Theory of Reconstructive Problem Solving and Expertise Acquisition," *Machine Learning II: An Artificial Intelligence Approach*, R.S. Michalski, J.G. Carbonell, & T. Mitchell (eds.), pp. 371–392, Morgan Kaufmann Publishers, San Mateo, CA, 1986.

Carr, M. "Metacognitive Knowledge as a Predictor of Decomposition Strategy Use," Paper presented at the Southeast Cognitive Science Conference, Georgia Institute of Technology, Atlanta, Georgia, January 1992.

Chi, M.T.H., & VanLehn, K.A. "The Content of Physics Self-explanations," *The Journal of the Learning Sciences*, Vol. 1, No. 1, pp. 69–105, 1991.

Collins, G., & Birnbaum, L. "Problem-Solver State Descriptions as Abstract Indices for Case Retrieval," *Working Notes of the AAAI Spring Symposium Series: Case-Based Reasoning*, Stanford, CA, 1990.

Cox, M., & Ram, A. "Multistrategy Learning with Introspective Meta-Explanations," *Machine Learning: Proceedings of the Ninth International Conference*, D. Sleeman & P. Edwards (eds.), ppp. 123–128, Aberdeen, Scotland, 1992.

Davis, R. "Meta-rules: Reasoning about control," *Artificial Intelligence*, Vol. 15, No. 3, pp. 179–222, 1980.

DeJong, G.F., & Mooney, R.J. "Explanation-based learning: An alternative view," *Machine Learning*, Vol. 1, No. 2, pp. 145–176, 1986.

Doyle, J. "A Truth Maintenance System," *Artificial Intelligence*, Vol. 12, pp. 231–272, 1979.

Falkenhainer, B. "A Unified Approach to Explanation and Theory Formation," *Computational Models of Scientific Discovery and Theory Formation*, J. Shrager & P. Langley (eds.), pp. 157–196, Morgan Kaufmann Publishers, San Mateo, CA, 1990.

Flann, N.S., & Dietterich, T.G. "A Study of Explanation-based Methods for Inductive Learning," *Machine Learning*, Vol. 4, pp. 187–226, 1989.

Genest, J., Matwin, S., & Plante, B. "Explanation-based Learning with Incomplete Theories: A Three-step Approach," *Proceedings of the Seventh International Conference on Machine Learning*, B.W. Porter & R.J. Mooney (eds.), pp. 286–294, Austin, TX, 1990.

Hammond, K.J. "Opportunistic Memory: Storing and Recalling Suspended Goals," *Proceedings of a Workshop on Case-based Reasoning*, J.L. Kolodner (ed.), pp. 154–168, Clearwater Beach, FL, 1988.

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

Hayes-Roth, B., and Hayes-Roth, F. "A Cognitive Model of Planning," *Cognitive Science*, Vol. 2, pp. 275–310, 1979.

Hunter, L. "Classifying for Prediction: A Multistrategy Approach to Predicting Protein Structure," *Machine Learning: A Multistrategy Approach, Vol. IV*, R.S. Michalski & G. Tecuci (eds.), Morgan Kaufman Publishers, San Mateo, CA, 1993.

Hunter, L. "Knowledge Acquisition Planning for Inference from Large Datasets," *Proceedings of the Twenty Third Annual Hawaii International Conference on System Sciences*, B.D. Shriver (ed.), pp. 35–45, Kona, HI, 1990a.

Hunter, L. "Planning to Learn," *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 26–34, Boston, MA, 1990b.

Kass, A., Leake, D., and Owens, C. "SWALE: A Program That Explains," *Explanation Patterns: Understanding Mechanically and Creatively*, R.C. Schank, pp. 232–254, Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

Kedar-Cabelli, S. "Toward a Computational Model of Purpose-directed Analogy," *Analogica*, A. Prieditis (ed.), Kluwer Academic Publishers, Boston, MA, 1988.

Kolodner, J.L., and Simpson, R.L. "A Case for Case-based Reasoning," *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*, Boulder, CO, 1984.

Laird, J.E., Rosenbloom, P.S., and Newell, A. "Chunking in Soar: The Anatomy of a General Learning Mechanism," *Machine Learning*, Vol. 1, pp. 11–46, 1986.

Maes, P. "Introspection in knowledge representation," *Advances in Artificial Intelligence II*, B. Du Boulay, D. Hogg, and L. Steels (eds.), Elsevier Science Publishers, New York, NY, 1987.

Maes, P., and Nardi, D. *Meta-Level Architectures and Reflection*. Elsevier Science Publishers, New York, NY, 1988.

Michalski, R.S. "Inferential Theory of Learning: Developing Foundations for Multistrategy Learning," *Machine Learning: A Multistrategy Approach, Vol. IV*, R.S. Michalski & G. Tecuci (eds.), Morgan Kaufman Publishers, San Mateo, CA, 1993.

Michalski, R.S. "A Theory and Methodology of Inductive Learning," *Artificial Intelligence*, Vol. 20, pp. 111–161, 1983.

Minsky, M. "Steps towards Artificial Intelligence," *Computers and Thought*, E.A. Feigenbaum and J. Feldman (eds.), pp. 406–450, McGraw-Hill, New York, NY, 1963.

Minsky, M. *The Society of Mind*, Simon and Schuster, New York, NY, 1985.

Mitchell, T.M., Keller, R., and Kedar-Cabelli, S. "Explanation-based Generalization: A Unifying View," *Machine Learning*, Vol. 1, No. 1, pp. 47–80, 1986.

Mitchell, T.M., Utgoff, P.E., and Banerji, R. "Learning by Experimentation: Acquiring and Refining Problem-Solving Heuristics," *Machine Learning: An Artificial Intelligence Approach*, R.S. Michalski, J.G. Carbonell, and T.M. Mitchell (eds.), pp. 163–189, Morgan Kaufman Publishers, San Mateo, CA, 1983.

Mooney, R.J., and Ourston, D. "A Multistrategy Approach to Theory Refinement," *Machine Learning: A Multistrategy Approach, Vol. IV*, R.S. Michalski & G. Tecuci (eds.), Morgan Kaufmann Publishers, San Mateo, CA, 1993.

Ng, E., and Bereiter, C. "Three Levels of Goal Orientation in Learning," *The Journal of the Learning Sciences*, Vol. 1, Nos. 3 & 4, pp. 243–271, 1991.

Ng, H., and Mooney, R.J. "On the Role of Coherence in Abductive Explanation," *Proceedings of the Eighth National Conference on Artificial Intelligence*, pp. 337–342, Boston, MA, 1990.

Park, Y., and Wilkins, D.C. "Establishing the Coherence of an Explanation to Improve Refinement of an Incomplete Knowledge Base," *Proceedings of the Eighth National Conference on Artificial Intelligence*, pp. 511–516, Boston, MA, 1990.

Pazzani, M.J. "Learning to Predict and Explain: An Integration of Similarity-based, Theory-driven and Explanation-based Learning," *The Journal of the Learning Sciences*, Vol. 1, No. 2, pp. 153–199, 1991.

Pirolli, P., and Bielaczyc, K. "Empirical Analyses of Self-Explanation and Transfer in Learning to Program," *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, 1989.

Pollock, J.L. "A General Theory of Rationality," *Journal of Theoretical and Experimental Artificial Intelligence*, Vol. 1, pp. 209–226, 1989.

Rajamoney, S. "Explanation-based theory revision: An approach to the problems of incomplete and incorrect theories," Ph.D. thesis, University of Illinois, Department of Computer Science, Urbana, IL, 1989.

Ram, A. "Question-driven Understanding: An Integrated Theory of Story Understanding, Memory and Learning," Ph.D. thesis, Research Report #710, Yale University, Department of Computer Science, New Haven, CT, 1989.

Ram, A. "Decision Models: A Theory of Volitional Explanation," *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 198–205, Cambridge, MA, 1990a.

Ram, A. "Knowledge Goals: A Theory of Interestingness," *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 206–214, Cambridge, MA, 1990b.

Ram, A. "A Theory of Questions and Question Asking," *The Journal of the Learning Sciences*, Vol. 1, Nos. 3 & 4, pp. 273–318, 1991.

Ram, A. "Indexing, Elaboration and Refinement: Incremental Learning of Explanatory Cases," *Machine Learning*, Vol. 10, No. 3, pp. 201–248, 1993.

Ram, A. "AQUA: Questions that Drive the Explanation Process," to appear in *Inside Computer Explanation*, R.C. Schank, A. Kass, & C.K. Riesbeck, Lawrence Erlbaum Associates, Hillsdale, NJ, in press.

Ram, A., and Hunter, L. "The Use of Explicit Goals for Knowledge to Guide Inference and Learning," *Applied Intelligence*, Vol. 2, No. 1, pp. 47–73, 1992.

Ram, A., and Leake, D. "Goal-driven Learning: Fundamental Issues and Symposium Report," to appear in *AI Magazine*, Vol. 14, No. 4, 1993.

Reich, Y. "Macro and Micro Perspectives of Multistrategy Learning," *Machine Learning: A Multistrategy Approach, Vol. IV*, R.S. Michalski & G. Tecuci (eds.), Morgan Kaufman Publishers, San Mateo, CA, 1993.

Scardamalia, M., and Bereiter, C. "Higher Levels of Agency for Children in Knowledge Building: A Challenge for the Design of New Knowledge Media," *The Journal of the Learning Sciences*, Vol. 1, No. 1, pp. 37–68, 1991.

Schank, R.C., and Leake, D.B. "Creativity and Learning in a Case-based Explainer," *Machine Learning: Paradigms and Methods*, J.G. Carbonell (ed.), MIT Press, 1990.

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

Schank, R.C. "The Current State of AI: One Man's Opinion," *AI Magazine*, Vol. 4, No. 1, pp. 3–8, 1983.

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

Shavlik, J.W., and Towell, G.G. "An Approach to Combining Explanation-based and Neural Learning Algorithms," *Connection Science*, Vol. 1, No. 3, 1989.

Stefik, M.J. “Planning and Meta-planning (MOLGEN: Part 2),” *Artificial Intelligence*, Vol. 16, pp. 141–169, 1981.

Thagard, P. “Explanatory Coherence,” *Behavioral and Brain Sciences*, Vol. 12, No. 3, pp. 435–502, 1989.

VanLehn, K.A, Jones, R.M., and Chi, M.T.H. “A Model of the Self-explanation Effect,” *The Journal of the Learning Sciences*, Vol. 2, No. 1, pp. 1–60, 1992.

Weintraub, M.A. “An Explanation-based Approach to Assigning Credit,” Ph.D. thesis, Department of Information and Computer Science, The Ohio State University, Columbus, OH, 1991.

Wilensky, R. “Meta-planning: Representing and using Knowledge about Planning in Problem Solving and Natural Language Understanding,” *Cognitive Science*, Vol. 5, pp. 197–233, 1981.

Winston, P.H. “Learning Structural Descriptions from Examples,” *The Psychology of Computer Vision*, P.H. Winston (ed.), pp. 157–209, McGraw-Hill, New York, 1975.

Zukier, H. “The Paradigmatic and Narrative Modes in Goal-guided Inference,” *Handbook of Motivation and Cognition: Foundations of Social Behavior*, R. Sorrentino and E. Higgins (eds.), pp. 465–502, Guilford Press, Guilford, CT, 1986.