Machine Learning: Proceedings of the Ninth International Conference, Workshop on Computational Architectures, Aberdeen, Scotland, 1992.

An Architecture for Integrated Introspective Learning

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

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

S. Narayanan
School of Industrial & Systems Engg.
Georgia Institute of Technology
Atlanta, Georgia 30332-0205
sn@chmsr.gatech.edu

Abstract

This paper presents a computational model of integrated introspective learning, which is a deliberative learning process in which a reasoner introspects about its own performance on a reasoning task, identifies what it needs to learn to improve its performance, formulates learning goals to acquire the required knowledge, and pursues its learning goals using multiple learning strategies. We discuss two case studies of integrated introspective learning in two different task domains. The first case study deals with learning diagnostic knowledge during a troubleshooting task, and is based on observations of human operators engaged in a real-world troubleshooting task at an electronics assembly plant. The second case study deals with learning multiple kinds of causal and explanatory knowledge during a story understanding task. The model is computationally justified as a uniform and extensible framework for deliberative learning using multiple learning strategies, and cognitively justified as a plausible model of human deliberative learning.

1 Introduction

This paper presents a computational model of integrated introspective learning, which is a deliberative or strategic learning process in which a reasoner introspects about its own performance on a reasoning task, assigns credit or blame for its performance, identifies what it needs to learn to improve its performance, formulates learning goals to acquire the required knowledge, and pursues its learning goals using multiple learning strategies. Our model makes the following claims about the nature of learning: (i) that *learning is active*, and goal-driven processes underlie much of the learning that occurs during the performance of analytical tasks in complex, real-world domains; (ii) that *learning is experiential* and occurs incrementally through the performance of a reasoning task; (iii) that *learning is opportunistic*, and learning goals that are not immediately satisfiable are remembered so that the reasoner can recognize and use opportunities to pursue them; (iv) that *learning is diverse* and involves multiple different strategies for acquiring new knowledge, modifying existing knowledge, and reorganizing the knowledge base; and finally (v) that *learning is introspective* and involves reflecting on one's own performance, monitoring the state of one's own knowledge, and analyzing one's own reasoning processes.

We also describe two computer systems that implement this theory. The systems are active and goal-driven, starting out with an incomplete understanding of a novel domain and learning through experience using multiple learning strategies. Their learning goals are functional to the purpose of the systems, and are identified during the pursuit of the performance task. The systems reason about the best way to perform a task, introspectively analyzes their own successes and failures in performing their tasks, reasons about what they need to learn, select appropriate learning strategies to acquire that information, and invoke the learning algorithms which then cause them to acquire new knowledge, modify existing knowledge, or reorganize memory by re-indexing knowledge in memory. Our 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.

2 The nature of learning

In our model, learning is viewed as an active, experiential, opportunistic, multistrategy, and introspective process. Let us discuss these properties at greater length.

Active learning: Traditional approaches in machine learning have assumed that the knowledge to be learned has already been identified by an exter-

nal agent (e.g., the "target concept" in explanation-based generalization [Mitchell *et al.*, 1986]). 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 and pursuit of what might be called the *learning goals* or *knowledge goals* of the reasoner is an important aspect of the learning problem (e.g., [Hunter, 1990; Michalski, 1992; Ng & Bereiter, 1991; Ram, 1989; Ram, 1991; Ram & Hunter, 1992]) We argue that active, goal-based learning is important for both computational reasons as well as for cognitive reasons. Thus credit or blame assignment (e.g., [Hammond, 1989; Weintraub, 1991]), formulating knowledge goals, asking questions, focussing attention, and pursuing learning actions are essential components of our learning model.

Experiential learning: Learning is an incremental process of theory formation in which the reasoner accumulates experience in some task domain. Through this experience, it learns to avoid the mistakes made during the performance task. Introspective analysis of a reasoning experience in a given situation allows the reasoner to use an appropriate learning strategy (or, as in our examples, multiple learning strategies) to learn exactly that which it 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 since one is likely to have similar experiences in the future (see, e.g., [Hammond, 1989; Kolodner & Simpson, 1984; Ram, 1992; Schank, 1982]).

Opportunistic learning: An important corollary of the active and experiential nature of learning is that learning is opportunistic. Often, a desired piece of knowledge will not be immediately available in the input, and so the corresponding knowledge goal will not be immediately satisfiable. In such cases, the reasoner must be able to suspend its knowledge goals, and reactivate them later when an appropriate opportunity arises. Because knowledge goals are indexed in memory, it is quite likely that an understander will find information relevant to goals other than the ones that are currently "active." In other words, knowledge goals can be satisfied opportunistically during the course of understanding [Birnbaum, 1986; Dehn, 1989; Hammond, 1988; Ram, 1989; Ram, 1991], leading to opportunistic learning of information previously identified as being useful to obtain (e.g., [Ram, 1992; Ram & Hunter, 1992]). In order for this to happen, the reasoner must be able to remember what it needs to learn, and recognize opportunities to learn the desired knowledge.

Multistrategy learning: There are several things one might learn from any experience, and several different ways of learning these. Once the reasoner has identified 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, 1990; Reich, 1992; Ram & Cox, 1992]).

Introspective learning: There are several fundamental problems to be solved before we can build intelligent systems capable of general multistrategy learning, including: determining the cause of a reasoning failure (*blame assignment*), deciding what to learn (*learning goal formulation*), and selecting the best learning strategies to pursue these learning goals (*strategy selection*). Although previous research has led to algorithms for learning in particular situations, no general theory of learning exists which allows the system to determine its own learning goals and to learn using multiple learning strategies. We claim that a reasoning system that can do this in a general manner must be able to reflect or introspect about its own internals.

Our approach to *integrated introspective learning* is as follows. First, we identify classes of learning situations based on an analysis of the types of reasoning failures that might occur. The taxonomy of reasoning failures follows from the functional architecture of the reasoning system. For each type of reasoning failure, we identify 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, for each type of rea-

[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.

soning failure we identify what needs to be learned to avoid such a failure, and associated learning strategies that can learn the desired knowledge.

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 use this introspective understanding to improve itself using the appropriate learning strategies [Cox & Ram, 1992; Ram & Cox, 1992].

3 An architecture for integrated introspective learning

Figure 1 illustrates a general architecture for an integrated multistrategy reasoning and learning system. The reasoner receives some input from the outside world, and focusses its attention on the part of the input which is relevant or interesting to it. It then uses one or more reasoning strategies, selected using a set of heuristics, to process the input appropriately. The strategies rely on knowledge that is indexed in the system's memory. The reasoner also records a trace of its reasoning process, which is introspectively examined for the purposes of blame assignment, deciding what to learn, and selecting the appropriate learning strategy(s).[2] The key representational entity in our learning 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. There are two broad classes of Meta-XPs. *Trace Meta-XPs* record a declarative trace of the reasoning performed by a system, along with causal links that explain the decisions taken. The trace 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 uses *Introspective Meta-XPs* to examine the declarative reasoning chain. Introspective Meta-XPs are structures used to explain and learn from a reasoning failure. They associate a failure type with learning goals and the appropriate set of learning strategies for pursuing those goals, and point to likely sources of the failure within the Trace Meta-XP. Thus an Introspective Meta-XP performs three functions: it aids in blame assignment (determining which knowledge structures are missing, incorrect or inappropriately applied); it aids in the formulation of appropriate knowledge goals to pursue; and 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 which 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.

4 A taxonomy of reasoning failures

Based on the above architecture, we can characterize the types of reasoning failures that a reasoner might encounter. These are presented in table 1. The term "failure" includes not simply performance errors, but also expectation failures [Schank, 1986], anomalous situations which the reasoner failed to predict, and other types of reasoning failures as well. 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]. Note that an unexpected success also counts as a reasoning "failure" because the reasoner was unable to correctly predict the outcome of the task.

If reasoning is assumed to consist of goal-directed processing of an input using some background knowledge, there are only a limited number of classes of factors that may be responsible for the success or failure of the reasoning process. A failure could stem from the goal, the process, the input, or the background knowledge. Furthermore, if both knowledge and sets of reasoning strategies are organized in order to facilitate access to them, so that appropriate knowledge and strategies can be retrieved and brought to bear on a given situation, the organization of knowledge and reasoning strategies may be responsible for a failure as well. Background knowledge is typically organized by associative links or indices which connect applicability conditions to an appropriate knowledge structure or state, whereas reasoning strategies are typically organized using heuristic rules which link applicability conditions with an operator or process.[3] Finally, if one allows for opportunistic planning and reasoning,

[2] Although learning is, in our view, a type of "reasoning", it is functionally distinct and is thus shown as a separate functional module in figure 1.

[3] If reasoning strategies are represented declaratively, these heuristic rules may be viewed as "indices" to the knowledge structures that encode these strategies.

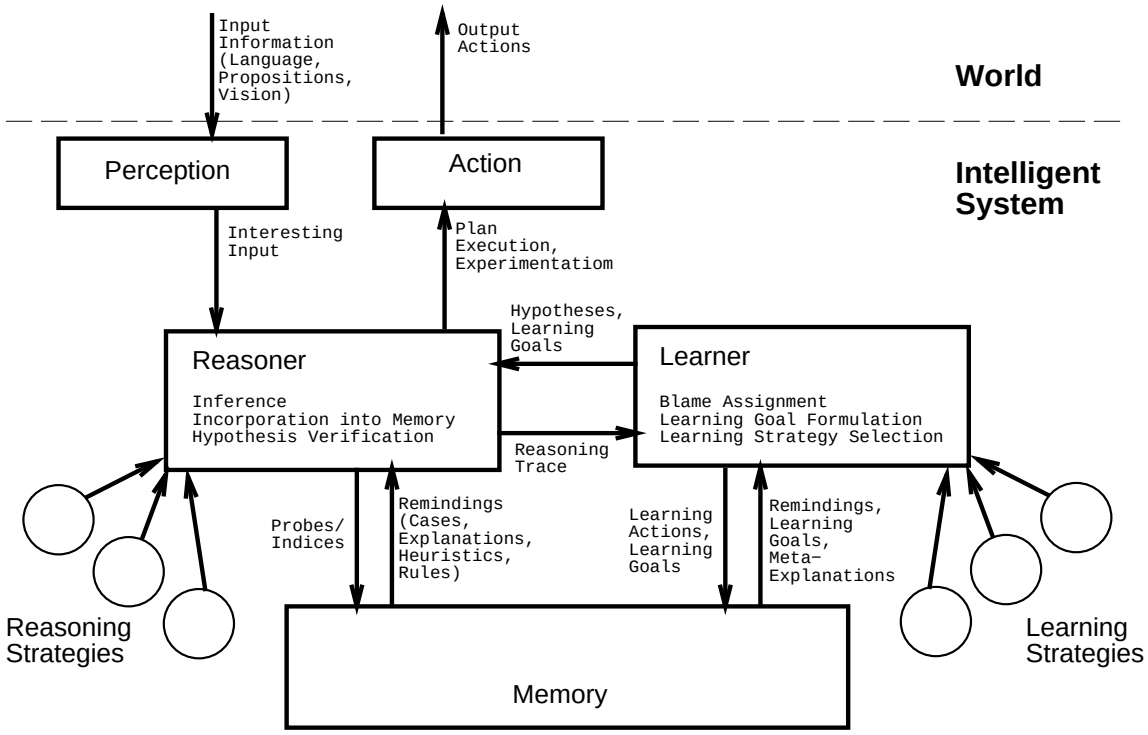

Figure 1: The Integrated Introspective Learning architecture.

goals may also be indexed, or associated with some context, and therefore prone to organizational and retrieval failures.

The dimensions of reasoning failures shown in table 1 identify classes of factors that bear on the blame assignment problem and, hence, on learning. Reasoning failures may be attributed to the reasoner's goals, to the opportunistic selection of suspended goals (or tasks) to pursue, to reasoning strategies or methods used to pursue these goals, to the heuristics used to choose such strategies, to the input, to some piece of background knowledge about the domain, or to the indices used to organize and access background knowledge. Each column in the table represents one of these dimensions, each of which could be missing or incorrect (the failure cases) or correct (the successful case). The last row lists a general characterization of the "source" of the dimension in the integrated introspective learning architecture. Finally, as suggested by the subtable, if a reasoner interacts with other agents in the world, blame for a failure may be attributed to the goals, strategies, input and knowledge of other agents. Thus, for example, noise in the input may actually be due to intentional deception motivated by conflicting goals of an external agent.

Learning is triggered when a reasoning failure is detected. For example, a reasoner may fail to explain a novel input due to incomplete background knowledge, or it might detect an anomaly by noticing a contradiction between the input and its background knowledge. It may fail to use an appropriate reasoning strategy in a given situation, which may cause it to come to an incorrect conclusion or to come to the correct conclusion less efficiently than it could. It may fail to formulate the appropriate reasoning tasks to pursue, or it may select the wrong task to pursue next. These and other types of reasoning failures are detected through an analysis of the Trace Meta-XP that represents the reasoning trace. Reasoning failures are associated with learning goals which specify what the reasoner needs to learn, which, in turn, are associated with learning strategies. An Introspective Meta-XP, then, could be viewed as a "reasoning pattern," representing a trace of a typical reasoning process, the failures encountered during the reasoning process, and the learning necessary in that situation. Examples and further details may be found in [Cox & Ram, 1992; Ram & Cox, 1992].

5 Two case studies

To ensure generality of the theory, we have performed two case studies in developing reasoning systems in two very different task domains. One system uses "shallow" knowledge to troubleshoot in an electronics assembly plant, and the other uses "deep" causal knowledge to understand natural language stories. Although space limitations preclude detailed discussion

given learning strategy.

Unlike Meta-TS, the knowledge used by Meta-AQUA's reasoning module is deep, explanatory knowledge about physical causality and human motivations. Trace Meta-XPs are instantiated to represent the system's explanation process, and the system's use of this knowledge for the explanation process. Learning strategies are selected using Introspective Meta-XPs. In general, Introspective Meta-XPs are built out of reasoning chains involving successful predictions, expectation failures, retrieval failures and incorporation failures. For example, a common type of failure arises when the reasoner finds an explanation that it thinks is appropriate, but the correct explanation turns out to be a different, novel explanation that the reasoner did not know about. This situation, or "reasoning pattern," is represented by a composite Meta-XP that consists of two basic Meta-XPs: `XP-Novel-Situation` and `XP-Mis-Indexed-Structure`. `XP-Novel-Situation` directs an explanation-based generalization algorithm to be applied to the node representing the novel explanation. The new explanation is then indexed in memory using an index learning algorithm. `XP-Mis-Indexed-Structure` directs the indexing algorithm to the old, incorrectly applied explanation. It recommends that the explanation be re-indexed so that it is not retrieved in similar situations in the future. Further details of the Meta-AQUA system can be found in [Cox & Ram, 1992; Ram & Cox, 1992].

6 Conclusions

The focus of our research is on the integration of different kinds of knowledge and reasoning processes into goal-driven, real-world systems that can learn through experience. In particular, we are interested in modelling the kind of active, goal-driven learning processes that underlie deliberative learning during the performance of complex reasoning tasks. Our model of integrated introspective learning makes several claims about the nature of learning, reasoning and introspection that are supported by research in psychology and metacognition. The meta-explanations in our approach are similar to self-explanations [Chi & VanLehn, 1991; Pirolli & Bielaczyc, 1989]. This research shows that formulation of 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 our meta-explanations are explanations about events and objects in the reasoner's "mind". Experimental results in the metacognition literature suggests that introspective reasoning of the kind we propose here can facilitate reasoning and learning (see, e.g., [Schneider, 1985; Weinert, 1987]).

Our approach can be justified on computational grounds as well. The approach relies on a declarative representation of meta-models for reasoning and learning. There are several advantages of maintaining such structures in memory. Because these structures represent reasoning processes explicitly, the system can directly inspect the reasons underlying a given processing decision it has taken, evaluate the progress towards 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 analyze why reasoning errors occurred, and to facilitate learning from these errors. Furthermore, because both the reasoning process and the knowledge base are represented using the same type of declarative representations, processes which identify and correct gaps in a knowledge base can also be applied to the reasoning process itself. Finally, these knowledge structures provide a principled basis for integrating multiple reasoning and learning strategies, and a unified framework which makes it relatively straightforward to incorporate additional types of failures and additional learning strategies.

Knowledge about reasoning and learning processes is usually encoded as procedures in current AI systems, which are treated as unanalyzable chunks of code as far as the system is concerned. This is problematic both for theoretical reasons as well as for practical ones. To avoid this problem, such knowledge is encoded in meta-explanation structures. Our systems can reason introspectively about their own reasoning process and hence determine both what they need to learn and what learning strategies they should use. The approach is novel because it allows systems to reason about themselves and make decisions that would normally be hard-coded into their programs by the designer, adding considerably to the power of such systems. This ability is central to a general theory of multistrategy learning. To realize this ability, we have developed algorithms for learning and introspection, as well as representational methods using which a system can represent and reason about meta-models describing itself.

Acknowledgements

This research was supported by the National Science Foundation under grant IRI-9009710 and by the Georgia Institute of Technology.

References

[Birnbaum, 1986] L. Birnbaum. *Integrated Processing in Planning and Understanding.* Ph.D. thesis, Yale University, Department of Computer Science, New Haven, CT, 1986. Research Report #489.

[Chi & VanLehn, 1991] M.T.H. Chi & K. VanLehn. The Content of Physics Self-Explanations. *The Journal of the Learning Sciences*, **1**(1):69–105, 1991.

[Cox & Ram, 1992] M.T. Cox & A. Ram. Multistrategy Learning with Introspective Meta-Explanations. In *Machine Learning: Proceedings of the Ninth International Conference*, Aberdeen, Scotland, 1992.

[Dehn, 1989] N. Dehn. *Computer Story Writing: The Role of Reconstructive and Dynamic Memory.* Ph.D. thesis, Yale University, Department of Computer Science, New Haven, CT, 1989. Research Report #792.

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

[Hammond, 1989] K.J. Hammond. *Case-Based Planning: Viewing Planning as a Memory Task.* Perspectives in Artificial Intelligence. Academic Press, Boston, MA, 1989.

[Hunter, 1990] L.E. Hunter. Planning to Learn. In *Proceedings of the Twelvth Annual Conference of the Cognitive Science Society*, pp. 26–34, Boston, MA, 1990.

[Kolodner & Simpson, 1984] J. Kolodner & R. Simpson. A Case for Case-Based Reasoning. In *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*, Boulder, CO, 1984.

[Michalski, 1992] R.S. Michalski. Inferential Learning Theory as a Basis for Multistrategy Task-Adaptive Learning. In R.S. Michalski & G. Tecuci (eds.), *Machine Learning IV: A Multistrategy Approach*, Morgan Kaufman Publishers, San Mateo, CA, 1992, to appear.

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

[Mitchell *et al.*, 1986] T.M. Mitchell, R. Keller, & S. Kedar-Cabelli. Explanation-Based Generalization: A Unifying View. *Machine Learning*, **1**(1):47–80, 1986.

[Narayanan & Ram, 1992] S. Narayanan & A. Ram. Learning to Troubleshoot in Electronics Assembly Manufacturing. In *Proceedings of the Ninth International Machine Learning Conference, Workshop on Integrated Learning in Real-World Domains*, Aberdeen, Scotland, 1992.

[Narayanan *et al.*, 1992] S. Narayanan, A. Ram, S.M. Cohen, C.M. Mitchell, & T. Govindaraj. Knowledge-Based Diagnostic Problem Solving and Learning in the Test Area of Electronics Assembly Manufacturing. In *Proceedings of the SPIE Conference on Applications of AI X: Knowledge-Based Systems*, Orlando, FL, 1992.

[Ng & Bereiter, 1991] E. Ng & C. Bereiter. Three Levels of Goal Orientation in Learning. *The Journal of the Learning Sciences*, **1**(3&4):243–271, 1991.

[Pirolli & Bielaczyc, 1989] P. Pirolli & K. Bielaczyc. Empirical Analyses of Self-Explanation and Transfer in Learning to Program. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, pp. 450–457, Ann Arbor, MI, 1989.

[Ram & Cox, 1992] A. Ram & M.T. Cox. Introspective Reasoning using Meta-Explanations for Multistrategy Learning. In R.S. Michalski & G. Tecuci (eds.), *Machine Learning IV: A Multistrategy Approach*, Morgan Kaufman Publishers, San Mateo, CA, 1992, to appear.

[Ram & Hunter, 1992] A. Ram & L.E. Hunter. The Use of Explicit Goals for Knowledge to Guide Inference and Learning. *Applied Intelligence*, 1992, to appear.

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

[Ram, 1991] A. Ram. A Theory of Questions and Question Asking. *The Journal of the Learning Sciences*, **1**(3&4):273–318, 1991.

[Ram, 1992] A. Ram. Indexing, Elaboration and Refinement: Incremental Learning of Explanatory Cases. *Machine Learning*, 1992, to appear.

[Reich, 1992] Y. Reich. Macro and Micro Perspectives of Multistrategy Learning. In R.S. Michalski & G. Tecuci (eds.), *Machine Learning IV: A Multistrategy Approach*, Morgan Kaufman Publishers, San Mateo, CA, 1992, to appear.

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

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

[Schneider, 1985] W. Schneider. Developmental Trends in the Metamemory-Memory Behavior Relationship: An Integrative Review. In D.L. Forrest-Pressley, G.E. MacKinnon, & T.G. Waller (eds.), *Metacognition, Cognition, and Human Performance, Volume 1*, Academic Press, New York, NY, 1985.

[Weinert, 1987] F.E. Weinert. Introduction and Overview: Metacognition and Motivation as Determinants of Effective Learning and Understanding. In F.E. Weinert & R.H. Kluwe, (eds.), *Metacognition, Motivation, and Understanding.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

[Weintraub, 1991] M.A. Weintraub. *An Explanation-Based Approach to Assigning Credit.* Ph.D. thesis, The Ohio State University, Columbus, OH, 1991.