To appear in A. Ram & D.B. Leake, editors, *Goal-Driven Learning*, MIT Press/Bradford Books.

Chapter 21

Learning as Goal-Driven Inference

Ryszard S. Michalski and Ashwin Ram

1 Inferential Theory of Learning

A remarkable aspect of human learners is that they are able to apply a great variety of learning strategies in a flexible and goal-oriented manner and to dynamically accommodate the demands of changing learning situations. In contrast, most research in machine learning has been concerned with single learning strategy methods that employ one primary type of inference within a specific representational or computational paradigm.

In addition to the fact that such systems do not model human cognition adequately, single strategy methods also suffer from practical problems such as lack of flexibility and narrow range of applicability. The learning goal in such "monostrategy" learning systems is defined implicitly by what they can actually do. For example, a decision tree learning program can, given appropriate examples, learn a decision tree. It cannot, however, take a set of rules and create more abstract rules, or take a decision tree and learn statistical dependencies among different attributes. It cannot even formulate the need to do so; it does not really know what it is trying to learn or why it is trying to learn it.

Developing an adequate and general computational model of adaptive, multistrategy, and goal-oriented learning is, therefore, a fundamental long-term objective for machine learning research for both theoretical and pragmatic reasons. In this chapter, we outline a proposal for developing such a model based on two key ideas. First, we view learning as an active process involving the formulation of learning goals during the performance of a reasoning task, the prioritization of learning goals, and the pursuit of learning goals using multiple learning strategies (Hunter, 1990; Ram, 1989, 1991; Ram & Cox, 1994; Ram & Hunter, 1992). However, while previous efforts have focused on the process of generating learning goals (Cox & Ram, 1994; Ram, 1991; Ram & Cox, 1994) and on the planful pursuit of learning goals (Cox & Ram, 1994; Hunter, 1990; Ram & Hunter, 1992; Redmond, 1992), little attention has been paid to the fundamental learning actions that actually carry out the inferences necessary to learn. In order to develop a general theory of

learning, it is important to understand the "primitives" of learning, that is, to develop a principled theory of the learning actions involved in acquiring and transforming knowledge.

The second key idea, then, is to model learning as a kind of inference in which the system augments and reformulates its knowledge using various types of primitive inferential actions (Michalski, 1993b). These primitive inferences, known as *knowledge transmutations*, include generalization, abstraction, explanation, and similization, and their counterparts, specialization, concretion, prediction, and dissimilization. For example, the familiar kind of learning known as "concept formation" is based on the inductive generalization transmutation, a kind of inference in which the system extends a description of a given set of instances so as to include other instances in that description. Knowledge transmutations can be combined in a flexible and dynamic manner to yield the desired learning behavior as specified by the system's learning goals. This view of learning is known as the *inferential theory of learning*, since it views learning fundamentally as a process of inference (Michalski, 1993b).

In contrast to broad-grained characterizations of learning in terms of traditional machine learning and information gathering algorithms (such as explanation-based generalization, inductive generalization, or database lookup, typically represented as "modules" in a multistrategy learning "toolbox"), the inferential theory of learning proposes a taxonomy of the types of inferences that underlying various forms of learning. It is interesting to note that many of the single strategy learning algorithms previously proposed in the machine learning literature can be modeled as combinations of these knowledge transmutations. Thus, in addition to serving as the basis of a computational model of multistrategy learning, knowledge transmutations can be used as a theoretical framework to characterize and analyze machine learning algorithms in general.

In this chapter, we will focus on the inferential theory of learning and its role in goal-driven learning. In particular, we will view learning as a guided or planful search through a *knowledge space*—the space of knowledge representations that the learner can represent or potentially generate. This search is actively guided by the *learning goals* of the system. The operators of the search are instantiations of generic types of *knowledge transmutations*, each capable of changing knowledge in some fundamental manner. Learning, then, is the goal-directed transformation of knowledge; this transformation is carried out through the basic inferential processes that are embodied in knowledge transmutation operators. The early ideas underlying the development of the inferential theory of learning go back to Michalski (1983). A classification and illustration of different learning strategies based on the criteria developed in the theory can be found in Michalski (1993a).

This view of learning seeks to characterize the capabilities of learning systems in terms of the inferential processes that underlie learning in these systems, and in terms of the influence of learning goals on these inferential processes. As such, it raises issues such as the types of knowledge transmutations that occur in different learning processes, the validity of knowledge obtained through different types of learning, the role of prior knowledge in learning, and the influence of learning goals on the learning process. Unlike traditional models of machine learning, but like the other models discussed in this book, our view of learning emphasizes the use of multiple types of inferences and the importance of learning goals in learning.

2 Learning as Goal-Guided Inference

Learning is triggered by the needs of the performance tasks being pursued by the system. As these tasks are pursued, the system decides what it needs to learn, that is, it identifies learning goals which, if satisfied, would improve its ability to pursue its tasks. This decision is made through an introspective analysis of the system's knowledge and reasoning processes as they are brought to bear on the task (Ram & Cox, 1994). If, as is usually the case, the system has several learning goals active at the same time, the system must also reason about the dependencies and priorities of these goals (Cox & Ram, 1994). To facilitate this, goals are represented in a goal dependency network (Stepp & Michalski, 1986; Michalski, 1993b). It will also often be the case that learning goals will not be immediately satisfiable; thus learning goals must also be indexed in the system's memory in a manner that allows them to be opportunistically retrieved at the appropriate time (Ram, 1989, 1991; Ram & Hunter, 1992).

Having decided what to learn, the system then pursues the desired learning by performing the necessary inferences and then storing the result. Based on the system's learning goals, its background knowledge, and available input, it selects and combines available knowledge transmutations in a dynamic manner to yield learning "plans" appropriate to the given learning situation. The result of this process is some new or newly reformulated knowledge that is then stored in the system's memory. This stage of learning, then, can be characterized by the equation "Learning = Inferencing + Memorizing," where "inferencing" refers to the process of goal-guided search through the knowledge space defined by knowledge transmutations (Michalski, 1993b). More formally, the search can be defined in terms of the following transformation, starting from the system's learning goals and ending with the desired knowledge that satisfies these goals:

Given:

- Learning goals (G)
- Input knowledge (*I*)
- Background knowledge (*BK*)
- Knowledge transmutations (*T*)

Determine:

- Output knowledge *O*, that satisfies goals *G*, by applying transmutations from the set *T* to input *I* and/or the background knowledge *BK*.

Figure 1 illustrates the general process of learning in which external input is transformed, in the context of goals and background knowledge and using a variety of inferential mechanisms, into new knowledge to be learned. In order to develop computational models of machine learning in this framework, one must develop a taxonomy of inferential mechanisms as well as methods for formulating goals, for representing goals as well as the interaction and interdependencies between goals, and for using goals to guide the inferential learning process. Let us discuss these questions in more detail, starting with the general issue of the nature of learning goals.

Insert Figure 1: A framework for a general learning process.

3 Learning Goals

In its basic sense, a learning goal is a specification of the knowledge or skill (performance procedure) that the learner wants to acquire. Learning goals may be externally provided or internally generated, general or specific, domain-independent or domain-dependent, one-time or recurrent. A system's learning goals are used to determine how its background knowledge should be modified in order to perform the desired learning.

Learning goals are a necessary component of any learning process. Given an input, and some nontrivial background knowledge, a learner could potentially generate an unbounded number of inferences (Ram & Hunter, 1992; Rieger, 1975). Many of these inferences, while "correct" in a purely logical sense, may not be useful in performing the overall tasks of the system. In fact, as has been demonstrated by several researchers, learning may sometimes even cause the performance of the system to deteriorate (e.g., Etzioni, 1990; Minton, 1990; Tambe, Newell, & Rosenbloom, 1990). To limit the proliferation of choices, and to ensure that the learning that occurs is actually useful, the learning process must be constrained and/or guided by the goals of the system (Hunter,

1990; Ram, 1989; Ram & Hunter, 1992). While these arguments have been made on computational grounds, the conclusions are supported by psychological evidence for goal-driven learning in humans (e.g., Barsalou, 1991; Ng & Bereiter, 1991; Steinbart, 1992). Similar arguments apply to the use of goals to focus inference generation for understanding, explanation, and diagnosis (e.g., Birnbaum & Collins, 1984; Hunter, 1990; Leake, 1991; Ram, 1990, 1991).

Given a learning goal, the learner determines what parts of prior knowledge are relevant to it, in what form the desired knowledge is to be represented, and how the learned knowledge is to be evaluated. There can be many different types of learning goals, which can be expressed implicitly or explicitly. In humans, for example, many goals are "hardwired" into the system reflecting biological and other needs; these then give rise to specific learning goals in particular situations as the system seeks to fulfill its needs. More generally, learning goals can also arise from intellectual needs such as, for example, the need to find certain information in order to explain an observed anomaly (e.g., Leake, 1992; Ram, 1991). While such goals are often implicitly programmed into machine learning systems by the system's designers, or provided as input by the system's users, it is also possible for a machine learning system to determine its own learning goals (e.g., Birnbaum, Collins, Freed & Krulwich, 1990; desJardins, 1992; Hunter, 1989; Ram, 1989, 1991; Ram & Cox, 1994; Redmond, 1992). For example, if a system engaged in some real-world problem solving task encounters difficulties during the performance of the task, it can reason about the knowledge it brought to bear, and the reasoning processes that it was using, in an attempt to explain why it failed to anticipate or avoid these difficulties. This explanation can be used to identify learning goals which, if satisfied, will result in improved performance in similar situations in the future (Ram & Cox, 1994).

Learning goals can be broadly classified as domain-independent or domain-specific (Michalski, 1993b). *Domain-independent goals* call for a certain type of learning activity, independent of the specific topic of discourse or problem-solving. For example, to acquire a general rule for classifying given facts, to confirm a given piece of knowledge, to derive it from some other knowledge, to concisely describe given observations, to discover a regularity in a collection of data, to find a causal explanation of a found regularity, to acquire control knowledge, to reformulate given knowledge into a more effective or operationalized form, to solve a problem of a given type, to plan what to learn, and so on.

Domain-specific goals call for acquiring a specific piece or type of domain knowledge, and are usually instantiations of domain-independent goals in the context of a specific problem-solving task in a specific domain. Thus, domain-independent learning goals may be viewed as specifications of

abstract types of learning activities. These are instantiated with domain-specific information in the context of the performance task to yield domain-specific learning goals which, in turn, are used to drive learning.

4 Goal Dependency Networks

In general, an intelligent system will have multiple learning goals that are interrelated in a very complex manner. In order to reason about the interactions between learning goals, the system must have some representation of the relationships between these goals, such as their interdependencies and relative priorities. Such a representational structure is called a *goal dependency network* (Stepp & Michalski, 1986; Michalski, 1993b). An example of a goal dependency network is shown in figure 2.

Insert Figure 2: An example goal dependency network for the goals to survive, to be healthy, and to live a vegetarian life-style.

A goal dependency network represents both general and specific goals, and the goal subordination relationships between these goals. Goals are represented as nodes, and the dependencies between nodes as labeled links denoting the types and strengths of the dependencies. Also represented are relevant attributes and predicates, and the attribute relevancy relations between these attributes and the corresponding goals.

In a goal dependency network, the most general and domain-independent goal is to store any given input and any plausible information that can be derived from it. More specific, but still domain-independent goals, specify the need to learn certain types of knowledge. Each of these goals is linked to more specific subgoals, some of which are domain-specific and call for determining some specific piece of knowledge. Since the desired piece of knowledge may not be immediately available or inferable, these goals must be indexed in the system's memory in a manner that allows the system to retrieve them dynamically when the appropriate learning opportunity is encountered (Hunter, 1990; Ram, 1989, 1991; Ram & Hunter, 1992). In this sense, goal-driven learning is analogous to opportunistic planning (Birnbaum & Collins, 1986; Hammond, Converse, Marks, & Seifert, 1993), but is carried out in the domain of knowledge as characterized by the knowledge space.

5 Knowledge Transmutations: Inferential Primitives for Learning

Having formulated learning goals and represented them in its memory structures, a system must perform the learning actions necessary to satisfy its learning goals. Machine learning systems that can combine and use multiple learning methods in learning are known as *multistrategy learning systems* (Michalski & Tecuci, 1993). A central issue in the design of such systems is the repertoire of learning strategies available, and the control methods used to select and combine the appropriate strategies at the appropriate time.

Typical models of multistrategy learning assume that the system's repertoire of learning strategies includes several of the single strategy learning algorithms developed in the machine learning literature, for example, empirical induction or explanation-based generalization, and/or information gathering methods, such as on-line database lookup. These are represented as modules in the system's "toolbox". In contrast, we propose to characterize learning actions at a finer level of detail. In particular, we propose a set of *knowledge transmutations* that can be thought of as the basic inferential primitives for learning (Michalski, 1993b). Knowledge transmutations are operators that make knowledge changes in the knowledge space. The knowledge space is a space of knowledge representations that can represent all possible inputs, all of the learner's background knowledge, and all knowledge that the learner can potentially generate. Learning is modeled as a process of inferential search through this space. This search is guided by the goals, the input, and the background knowledge of the system.

The central property of any knowledge transmutation is the type of underlying inference. Any type of inference can produce some useful knowledge worth remembering for future use; consequently, a complete theory of learning must include a complete theory of inference. To characterize these inferences in a general, language-independent manner, consider the following entailment:

$$P \cup BK \supset C$$

where P stands for a set of statements called the premise, BK stands for a set of statements representing the reasoner's background knowledge, and C stands for a set of statements called the consequent. P is assumed to be consistent with BK.

The *inference type* characterizes the transmutation along the truth-falsity dimension, and thus determines the validity of the knowledge derived by it. *Deductive inference*, or deduction, is deriving the consequent C, given P and BK. *Inductive inference*, or induction, is hypothesizing the premise P, given C and BK. Thus deduction can be viewed as tracing the above entailment "forward", and induction as tracing it "backward". Because this entailment succinctly explains the

relationship between two fundamental forms of inference, it is called the *fundamental equation* for inference. Deduction is *truth-preserving* in that C must be true if P and BK are true, and induction is *falsity-preserving* in that if C is false, then P must be false as well if BK is true. This property applies to every type of induction, including inductive generalization, abduction, inductive specialization, concretion, and so on. For example, if inductive inference produces a statement that characterizes a larger set of entities than the input statement C, it is called an *inductive generalization*; if it reduces the amount of detail in the description of a given set of entities, it is called an *inductive abstraction*; and if it hypothesizes a premise that explains the input, it is called an *inductive abduction*.

In a general view of deduction and induction that also captures their approximate or commonsense forms, the "strong" entailment "$\supset$" may be replaced by a "weak" entailment that includes cases in which C is only a plausible, probabilistic, or partial consequence of P and BK. The difference between strong (valid) and weak (plausible) entailment leads to another major classification of types of inference. Specifically, inferences can be *conclusive* (true in every possible situation) or *contingent* (true in some situations and not in others). These distinctions apply to all types of inference; for example, the strong forms *conclusive induction* and *conclusive abduction* and their weak counterparts *contingent induction* and *contingent abduction*. Figure 3 illustrates all major types of inference in a schematic manner.

Insert Figure 3: A classification of major types of inference.

Finally, each type of inference has a converse. For example, *abduction* hypothesizes explanations of a set of entities, and its converse, *prediction*, derives consequences of the properties of the set of entities. These derivations, as before, might be deductive or inductive, and conclusive or contingent.

This framework allows us to describe the complete set of knowledge transmutations that underlie all types of learning. Formally, a knowledge transmutation can be modeled as a transformation that takes as arguments a set of sentences (S), a set of entities (E), and background knowledge (BK), and generates a new set of sentences (S'), and/or a new set of entities (E') and/or new background knowledge (BK') (Michalski, 1993b):

$$T : S; E; BK \rightarrow S'; E'; BK'$$

Figure 4 provides a summary of the different types of transmutations together with the underlying types of inference. Transmutations can be classified into two categories, knowledge-generation transmutations, and knowledge manipulation transmutations. *Knowledge-generation transmutations* operate on the informational content of the input knowledge. For example, they may derive consequences from given knowledge, suggest new hypothetical knowledge, determine relationships between knowledge components, confirm or disconfirm given knowledge, perform mathematical operations on quantitative knowledge, organize knowledge into certain structures, and so on. Knowledge-generation transmutations are performed on statements that have a truth status. *Knowledge-manipulation transmutations*, in contrast, view input knowledge as data or objects to be manipulated, and can be performed on statements or on sets. They include inserting (deleting) knowledge components into (from) knowledge structures, physically transmitting or copying knowledge to/from other knowledge bases, or ordering knowledge components according to some organizational criteria.

Insert Figure 4: Knowledge transmutations and the underlying types of inference.

6 Towards a Computational Model of Goal-Driven Machine Learning

The ideas presented in the previous sections provide a conceptual framework for adaptive, goal-driven, multistrategy learning, which aims at integrating a diverse range of inferential learning strategies into an active, goal-driven learning system. We propose a general learning framework in which the system performs a given task or tasks, monitors its own performance on the task, introspectively analyzes this performance to determine what it needs to learn, formulates explicit learning goals to perform this learning, organizes these goals into a goal dependency network, detects appropriate opportunities for learning, and then performs the desired learning using multiple types of knowledge transmutations in an active, goal-guided search through knowledge space.

This framework combines several perspectives on various aspects of goal-driven learning, technical details of which can be found elsewhere (introspective analysis of reasoning traces: Ram & Cox, 1994; goal dependency networks: Stepp & Michalski, 1986; Michalski, 1993b; opportunistic pursuit of learning goals: Cox & Ram, 1994; Ram, 1989, 1991; Ram & Hunter, 1992; knowledge transmutations: Michalski, 1993b). In this chapter, we have attempted to step back from the technical details of this process to present a broader framework for the design of flexible machine learning systems, focusing in particular on the role of knowledge transmutations as a basis for goal-guided inference.

The proposed framework views learning as an active process of deciding what to learn and how to learn it. This view raises several issues for further research: the origins of learning goals, methods for deciding what to learn, methods for representing goal structures, methods for using learning goals to guide the selection and use of knowledge transmutations, methods for representing knowledge transmutations, and identification and analysis of other types of knowledge transmutations. While we and others have begun to discuss these issues, considerable research is needed to develop computational models of learning and to implement integrated machine learning systems that truly capture the complexities outlined in this chapter. We hope that this chapter will both stimulate researchers to tackle the research issues discussed here as well as provide a framework in which to perform the research.

Acknowledgments

This research was supported in part by the National Science Foundation under grants IRI-9020226 and IRI-9009710, in part by the Office of Naval Research under grant N000014-91-J-1351, and in part by the Defense Advanced Research Projects Agency under the grant N00014-91-J-1854, administered by the Office of Naval Research, and the grant F49620-92-J-0549, administered by the Air Force Office of Scientific Research. Authors are listed alphabetically. We thank David Leake and anonymous reviewers for their comments on an earlier draft of this chapter.

References

Barsalou, L.W. Deriving Categories to Achieve Goals. In G.H. Bower, editor, *The Psychology of Learning and Motivation: Advances in Research and Theory*, *Volume 27*, Academic Press, New York, NY.

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

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

Cox, M.T. & Ram, A. (1994). Choosing Learning Strategies to Achieve Learning Goals. In *Proceedings of the AAAI Spring Symposium on Goal-Driven Learning*, Stanford, CA.

desJardins, M. (1992). *PAGODA: A Model for Autonomous Learning in Probabilistic Domains*. Ph.D. Dissertation, Technical Report 92/678, University of California, Computer Science Department, Berkeley, CA.

Etzioni, O. (1990). *A Structural Theory of Explanation-Based Learning*. Ph.D. Dissertation, Technical Report CMU-CS-90-185, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

Hammond, K., Converse, T., Marks, M., & Seifert, C.M. (1993). Opportunism and Learning. *Machine Learning*, 10(3):279–309.

Hunter, L.E. (1989). *Knowledge Acquisition Planning: Gaining Expertise Through Experience*. Ph.D. Dissertation, Research Report #678, Yale University, Department of Computer Science, New Haven, CT.

Hunter, L.E. (1990). Planning to Learn. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 261–276, Boston, MA.

Leake, D. (1991). Goal-Based Explanation Evaluation. *Cognitive Science*, 15(4):509–545.

Leake, D.B. (1992). *Evaluating Explanations: A Content Theory*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Michalski, R.S. (1983). A Theory and Methodology of Inductive Learning. In R.S. Michalski, J. Carbonell, & T. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach*, pages 3–23, Morgan Kaufman Publishers, San Mateo, CA.

Michalski, R.S. (1993a). Toward a Unified Theory of Learning: Multistrategy Task-Adaptive Learning. In B.G. Buchanan & D.C. Wilkins, editors, *Readings in Knowledge Acquisition and Learning: Automating the Construction and Improvement of Expert Systems*, Morgan Kaufman Publishers, San Mateo, CA.

Michalski, R.S. (1993b). Inferential Theory of Learning as a Conceptual Basis for Multistrategy Learning. *Machine Learning*, 11(2/3):111–151, 1993.

Michalski, R.S. & Tecuci, G. (1993), editors. *Machine Learning: A Multistrategy Approach, Volume IV*, Morgan Kaufman Publishers, San Mateo, CA.

Minton, S. (1990). Quantitative Results Concerning the Utility of Explanation-Based Learning. *Artificial Intelligence*, 42:363–391.

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

Ram, A. (1989). *Question-Driven Understanding: An Integrated Theory of Story Understanding, Memory and Learning*. Ph.D. Dissertation, Research Report #710, Yale University, Department of Computer Science, New Haven, CT.

Ram, A. (1990). Knowledge Goals: A Theory of Interestingness. In *Proceedings of the Twelvth Annual Conference of the Cognitive Science Society*, pages 206–214, Cambridge, MA.

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

Ram, A. (1993). Indexing, Elaboration and Refinement: Incremental Learning of Explanatory Cases. *Machine Learning*, 10:201–248.

Ram, A. & Cox, M.T. (1994). Introspective Reasoning using Meta-Explanations for Multistrategy Learning. In R.S. Michalski & G. Tecuci, editors, *Machine Learning: A Multistrategy Approach, Volume IV*, Morgan Kaufman Publishers, San Mateo, CA.

Ram, A. & Hunter, L. (1992). The Use of Explicit Goals for Knowledge to Guide Inference and Learning. *Applied Intelligence*, 2(1):47–73.

Redmond, M. (1992). *Learning by Observing and Understanding Expert Problem Solving*. Ph.D. Dissertation, Georgia Institute of Technology, College of Computing, Atlanta, GA.

Rieger, C. (1975). Conceptual memory and Inference. In R.C. Schank, editor, *Conceptual Information Processing*. North-Holland, Amsterdam.

Steinbart, P.J. (1992). The Role of Questioning in Learning from Computer-Based Decision Aids. In T.W. Lauer, E. Peacock, and A.C. Graesser, editors, *Questions and Information Systems*, pages 273–285, Lawrence Erlbaum Associates, Hillsdale, NJ.

Stepp, R.S. & Michalski, R.S. (1986). Conceptual Clustering: Inventing Goal-Oriented Classifications of Structured Objects. In R.S. Michalski, J.G. Carbonell, & T.M. Mitchell, editors, *Machine Learning: An Atificial Intelligence Approach, Volume II*, pages 471–498, Morgan Kaufman Publishers, San Mateo, CA.

Tambe, M., Newell, A., & Rosenbloom, P.S. (1990). The Problem of Expensive Chunks and its Solution by Restricting Expressiveness. *Machine Learning*, 5:299–348.