Invited paper for the First World Conference on the
Fundamentals of Artificial Intelligence, Paris, July 1-5, 1991.

TOWARD A UNIFIED THEORY OF LEARNING:
An Outline of Basic Ideas

Ryszard S. Michalski
Center for Artificial Intelligence
George Mason University
Fairfax, VA 22030

Abstract

Initial results toward developing a unifying conceptual framework for characterizing diverse learning strategies and paradigms are presented. We outline the *inferential theory of learning* that aims at understanding *competence* aspects of learning processes, in contrast to computational theory that is concerned with complexity aspects. The theory views learning as a goal-oriented process of creating or modifying knowledge representations. Such a process may involve any type of inference (deduction, analogy or induction) or information transmutation (e.g., reformulation, abstraction or copying). Any type of learning can therefore be characterized in terms of the types of such knowledge transformations that occur in a learning process. Several concepts fundamental to understanding learning are analyzed in a novel way and compared, such as analytic vs. synthetic learning, deduction, induction, abduction, abstraction and generalization. It is shown, for example, that inductive generalization, inductive specialization and abduction can be viewed as various forms of general induction, and that abstraction is a form of constructive deduction. Based on these concepts, a general multicriterion classification of learning processes is proposed. The presented ideas have a special significance for the development of a new generation of learning systems, called *multistrategy systems*, that integrate diverse learning strategies in a goal-oriented fashion.

1. INTRODUCTION

In view of an extraordinary proliferation of different methods and approaches to machine learning, there is a strong need for developing a conceptual framework that would clarify their interrelationships, and determine the areas of their most effective applicability. For sources reporting the progress in various areas of machine learning, the reader is recommended to consult, e.g., Laird, 1988; Haussler and Pitt, 1988; Touretzky, Hinton and Sejnowski, 1988; Goldberg, 1989; Schafer, 1989; Segre, 1989; Fulk and Case, 1990; Porter and Mooney, 1990; Kodratoff and Michalski, 1990; Birnbaum and Collins, 1991). The purpose of this paper is to outline the *inferential theory of learning*, which analyzes and characterizes learning processes from the viewpoint of the types of knowledge transformations occuring in them. Since every learning process can be characterized in such terms, the theory offers a general conceptual framework for analyzing diverse learning systems.

Learning has been traditionally viewed as changing behavior due to experience. While such a view is intellectually appealing, it does not give any clear answer to the question of how to build

learning systems. To build an algorithmic model of learning, one needs to explain, in computational terms, why the changes occur, and how they occur in response to different kinds of experience.

To this end, the inferential learning theory postulates that learning is a process of improving knowledge representations by exploring the learner's experience. Such a process can be characterized by the kinds of knowledge transformations that are needed to accomplish the learning goal. These knowledge transformations are done by performing various kinds of inference (deduction, analogy or induction), and/or information transmutations (e.g., copying, reformulation or abstraction) that involve the learner's prior knowledge and the input information. These operations can be done by a learner explicitly, or implicitly, as results of a specific mechanism engaged in the processing of information. Since these operations can be applied in a great variety of ways, learning processes need to be guided by learner's goals, which also can be expressed explicitly or implicitly. The learner's experience can be in the form of sensory observations, facts or knowledge communicated by a source (e.g., a teacher). In sum, learning processes can be characterized in terms of their goals, the types of inference involved, the role of prior knowledge and the types of the input information.

The aims of the inferential theory are to understand the *competence* aspects of learning processes, in contrast to the computational learning theory (e.g., Fulk and Case, 1990), that concerns *computational complexity* of such processes. These competence aspects are concerned with such problems as what kinds of knowledge the learner would be able to learn from what kinds of inputs, how is this accomplished, and how the results of learning relate to what was received from a source and to what the learner already knew. The presented work draws upon the ideas presented earlier in (Michalski, 1983; Michalski and Ko, 1988; Michalski, 1990a). The next section presents basic tenets of the theory. To clearly explain the underlying ideas and research aims, the presentation relies on conceptual explanations and examples, rather than on precise definitions and formal elaborations.

2. BASIC TENETS OF THE THEORY

Any learning process aims at improving the learner's knowledge or skill by interacting with some information source. A key idea of the inferential theory of learning is that this improvement is done through various *knowledge transformations.* Consequently, the inferential theory analyzes learning processes from the viewpoint of the roles and types of knowledge transformations that occur in them.

The underlying tenet of the theory is that learning can be usefully viewed as a process of modifying knowledge structures to achieve a certain goal. These structures represent the learner's current knowledge and abilities. They are modified as a result of an interaction between the learner's prior knowledge, the inputs from an information source, and the learner's goal. These three components define what we call the *learning task*.

According to the theory, the interactions among the components of a learning task can be characterized, at a conceptual level, in terms of knowledge transformations that are required to accomplish a given learning goal. These knowledge transformations are accomplished by applying various kinds of inference - deductive, analogical or inductive, and/or information transmutations - copying, reformulation and abstraction. (Since such information transmutations are usually done according to well-defined and truth-preserving operations, they can be viewed as forms of deduction).

In symbolic learning systems, knowledge transformations are performed in a more or less explicit way, and in conceptually comprehensible steps. In subsymbolic systems (e.g., neural networks), the inferences are performed implicitly, in steps dictated by the underlying computational mechanism. For example, a computer program for learning concepts from examples may involve explicit rules of inductive generalization (e.g., Michalski, 1983). On the other hand, a neural network may produce a generalization of the same input as a result of a sequence of small modifications of the weights of the internode connections. Although these weight modifications do not directly correspond to any explicit inference steps they, nevertheless, can be analyzed and mapped into certain knowledge transformations. For example, Wnek at al. (1990) described a simple method for visualizing target concepts and concepts learned from examples by a neural network, genetic algorithm, and two symbolic learning systems (see also Figure 2).

A learning process is always guided by some underlying goal, otherwise the proliferation of choices of what to learn would quickly overwhelm any realistic learning system. The learning goal can be explicitly defined, or only implicitly defined, by the way the learner processes the input information, by what it pays attention to, etc. The input information ("input") can be observations, stated facts, concept instances, previously formed generalizations, conceptual hierarchies, or some combinations of different types of knowledge. At the beginning of a learning process, the input activates segments of the learner's prior knowledge that are relevant to the learning goal. Such learner's goal-relevant prior knowledge is called *background knowledge* (BK).

The background knowledge can be in different forms, e.g., in a declarative form, which is most useful for explicit reasoning (conceptual knowledge), or in procedural from, as sequences of instructions for executing specific tasks (control knowledge, skills).

Figure 1 illustrates major components and the information flow in a general learning process according to the theory.

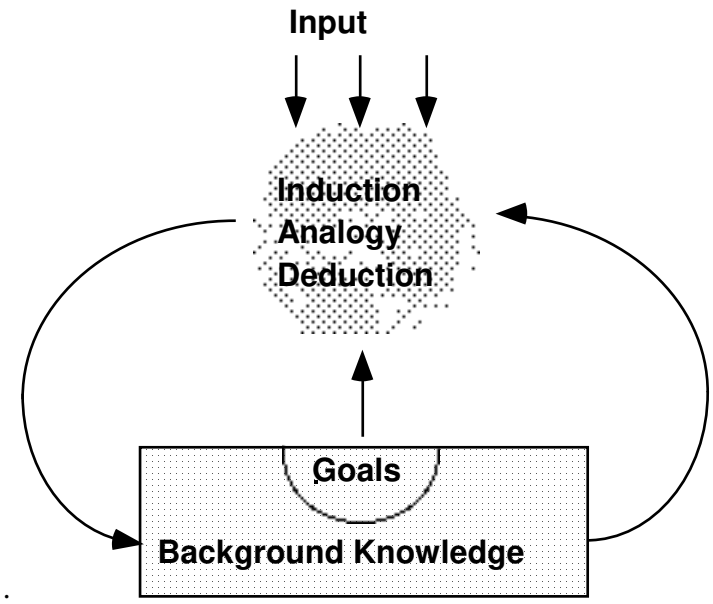

Figure 1. An illustration of a general learning process

In each learning cycle, the learner analyzes the input information in terms of its background knowledge and its goals, and generates new or better knowledge (according to the learning goal) through forms of inference that are required to accomplish the goal. The results are fed back to the learner's "knowledge base," and may be used in subsequent learning processes.

A learning strategy is defined by the type of knowledge transformation that turns the input into the stored knowledge. The lowest strategy is "rote learning" ("direct knowledge implantation") that involves essentially the copying of information from a source into the knowledge base. Such a process requires a proper arrangement of the new knowledge within the learner's current knowledge structure. The next level strategy, “learning from instruction," involves a selection of the relevant parts from the knowledge supplied by the source, its desirable syntactic transformations (to fit the learner's conceptual structure), and a determination of its relationship to the current knowledge. The above strategies are not supposed to change the meaning of the knowledge obtained from the source, and therefore engage only truth-preserving knowledge transformations. In such forms of learning the learner relies primarily on its memory, rather than on its reasoning capabilities.

This does not mean, of course, that rote learning and learning from instruction are not important forms of learning. Their implementation poses a number of interesting problems, such as those concerning knowledge representation, organization, access, and knowledge reformulation. In the case of learning from instruction, there is also a problem of determining what kind of information transmutation is to be made, and what parts of the source knowledge are relevant to the learner's goals. These strategies are also important because they are widely used in human learning, as well in computer systems. For example, building a computer database can be viewed (from the viewpoint of the computer) as a form of direct knowledge implantation. Most of the current methods of knowledge acquisition for knowledge-based systems can be viewed as combinations of the above two learning strategies. Higher strategies of learning require a learner to perform correspondingly more advanced forms of inference, such complex deduction, plausible deduction, analogy and induction.

As mentioned earlier, a learning process depends on the learning task (defined by the available input information, the learner's background knowledge, and the learning goal). The input information comes from an information source, which may be the learner's environment, a teacher, or a learner's own internal process. The prior knowledge relevant to the learning goal directly affects what kind of learning strategy is to be applied. The learning goal can be implicit or explicit, but it is a necessary guide for determining what parts of prior knowledge are relevant, what type of knowledge is desired, and how to evaluate the learned knowledge.

There can be many different types of learning goals, e.g., to solve a problem, to perform an action, to "understand" observed facts, to concisely describe given data, to discover a regularity in a collection of observations in terms of high level concepts, etc. A learner may have more than one goal, and the goals may be conflicting. In such a situation, their relative importance affects the decision about the amount of effort the learner extends in pursuing any of them. A weakness of some machine learning research is that it considers a learning process separately from the learning goal(s), and as a result it is often method-oriented rather than problem-oriented. Studying the role of goals in learning is an important research topic for machine learning.

In summation, the inferential theory states that in order to learn, an agent needs to be able to perform *inference*, and to have *memory* that supplies the background knowledge needed for performing the inference, and to record the results of the inference for future use. Without either

of the two components, the ability to reason and the ability to store and retrieve information from memory, no learning can be accomplished. Thus, one can write an "equation":

Learning = Inference + Memory

It should be noted that the term "inference" is used here in a very general sense, meaning any possible form of knowledge transformation or manipulation, including syntactic and semantic transformations, as well as random searching for a specified entity. The double role of memory, as a supplier of BK, and as a storer of the results, is often reflected in the organization of a learning system. For example, in a neural net, BK resides in both, the structure of the network (in the type of units used, and in the way they are interconnected) and in the initial weights of the connections. The learned knowledge usually resides only in the changed values of the weights. In a decision tree learning system, the BK includes an attribute evaluation procedure. The knowledge created is in the form of a decision tree. In an "ideal" rule learning system, all BK would be in the form of rules, and a learning process would involve both modifying prior rules and/or creating new ones. The limits of what can be learned are determined by what part of BK cannot be changed in a learning process.

Because inferential theory views learning as an inference process, it may appear that it only applies to symbolic methods, and does not apply to subsymbolic or hybrid forms of learning, such as neural net learning, reinforcement learning or genetic algorithm-based learning. It is argued that it does apply to them also, because these methods can also be analyzed from the viewpoint of the types of knowledge transformations performed by them. They can generalize, specialize, similize, reformulate or copy the input information.Figure 2 illustrates this point.

The figure presents "images" of concepts learned by a neural network, a classifier system using a genetic algorithm, a decision tree learning program (C4.5), and a rule learning program (AQ15). Each cell of a diagram represents a single combination of attribute values, i.e., an instance in the description space. The area "target concept" includes all possible instances of the concept to be learned. The area "learned concept" denotes all instances that the would classify as belonging to the concept after the learning process. The set-theoretic difference between the "target concept" and the "learned concept" thus represents "error image." Each instance in this area will be incorrectly classified by the learned concept. By analyzing the images of the concepts learned by different paradimgs, one can determine the degree to which they generalized the original examples, can "see" the differences between these generalizations, etc. (For more details, see Wnek et al., 1990.)

Thus, from the viewpoint of the inferential learning theory, the difference between symbolic and subsymbolic systems is that the latter perform knowledge transformations implicitly, e.g., by modifying weights of connections, rather than explicitly, as in the former. The prior knowledge in these systems is also represented in an implicit way, e.g., by the structure of the network and the initial settings of the weights of the connections, or by initial classifiers in a genetic algorithm-based learning. This knowledge can be re-represented (at least conceptually), in the form of logical expressions or rules, and analysed as any other knowledge. The subsymbolic approaches, obviously, also have the ability to memorize results of their learning. For example, in a neural net, the acquired knowledge is manifested in the new weights of the connections among the net's units.

3. TYPES OF INFERENCE

As stated earlier, the inferntial theory postulates that a learner learns by conducting inference to derive the desirable knowledge representation from the input and current BK, and then stores the results for future use. Such a process may involve any type of inference. Therefore, looking from such a viewpoint, a complete learning theory must include a complete theory of inference. Such a theory of inference should be able to account and explain all possible types of knowledge transformations. Figure 3 presents an attempt to schematically illustrate all general types of inference.

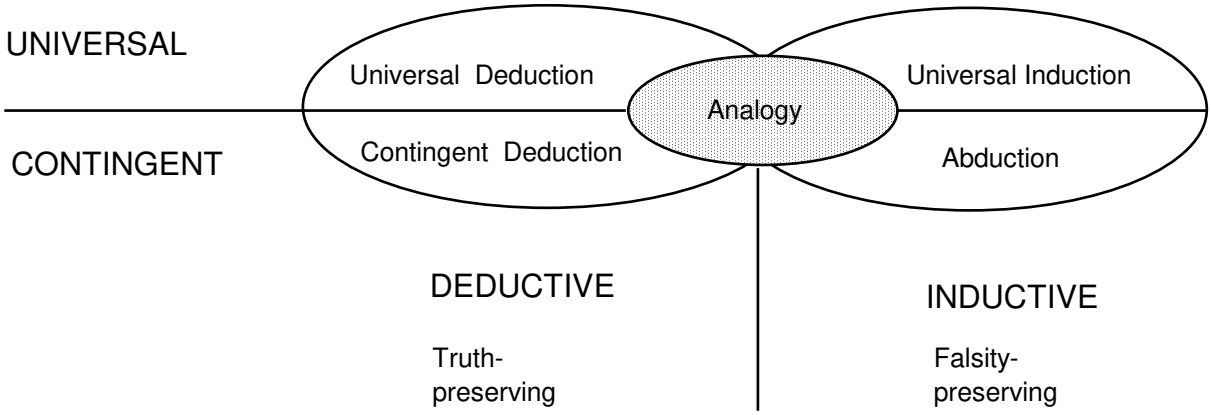

Figure 3. A classification of different kinds of inference

The first major classification is to divide inference types into deductive and inductive. The difference can be explained by considering an entailment:

$$P \cup BK \models C \quad (1)$$

where P denotes a set of statements, called *premise*, KB represents the reasoner's background knowledge (including rules of inference), and C denotes a set of statements, called *consequent*. Deductive inference is deriving consequent C, given premise P and BK. Inductive inference is hypothesizing premise P, given consequent C and BK. If $\models$ is the formal logic entailment (i.e., (1) is a valid formula), then deductive inference can be viewed as "tracing forward" the relationship (1), and induction as tracing "backward" such relationship.

In a general view of deduction and induction, which captures also their approximate or commonsense forms, $\models$ may denote a "weak" entailment, i.e., plausible, probabilistic or partial. The difference between the"strong" (valid) and "weak" entailment leads to another major classification of inference types. Specifically, inferences can be divided into those based on the *universal* (*domain-independent* or*tautological)* dependencies, and those based on *contingent* (*domain-dependent* or *conditional*) dependencies. A universal dependency between knowledge components (e.g., individual statements or sets of statements) represents a necessarily true relationship, i.e., a relationship that must be true in all possible worlds. For example, if the statement "All elements of the set X have the property q" is true, then then the statement "x, an element of X, has the property q" must necessarily be true also. This relationship is true independently of the domain of discourse, i.e., of the nature of elements in the set X.

If a reasoning process involves only statements that are assumed to be true (e.g., axioms, "true" observations, etc.) and universal dependencies, then deriving C, given P, is the universal deduction, and hypothesizing P, given C, is universal induction. For example, suppose that BK is "All elements of the set X have the property q" and the input (premise P) is "x is an element of the set X." Deriving a statement "x has the property q" is universal deduction. On the other hand, suppose that BK is "x is an element of the set X" and the input (observed consequent C) is "x has the property q." Deriving a hypothesis (premise P) that "All elements of the set X have the property q" is universal induction.

Contingent dependencies are relationships that are domain-dependent in the sense that they represent some world knowledge that is not totally certain, or that they may be worlds in which they are not true. They can be in the form of probabilistic dependencies (e.g., Pearl, 1988), plausible mutual dependencies (Collins and Michalski, 1989), partial dependencies, etc. The contingency of these relationships is usually due to the fact that they represent an incomplete information about all the factors in the world that enter a dependency. These relationships may hold with different "degrees of strength." The conclusions from inferences based on contingent dependencies (even using valid rules of inference) are therefore uncertain, and may be characterized by different "degrees of belief" (probabilities, degrees of truth, likelihoods, etc.). They also usually hold in both directions, although not with the same strength in each direction (Colins and Michalski, 1989).

For example, "If there is fire, then there is smoke" is a (bidirectional) contingent dependency, because there could be a situation or a world in which it is false. If one sees fire, then one may derive (deductively) a conclusion that there may be smoke. This conclusion, however, is not certain. In a reverse direction of reasoning ("tracing backward" this dependency), if one observes smoke, then one may hypothesize (abductively) that there is fire. It is also an uncertain inference. Therefore, it may appear that there is no principal difference between contingent deduction and abduction.

These two types of inference are different if one assumes that |= in (1) represents a causal ordering, i.e., P is viewed as a cause, and C as a consequence. Contingent deduction derives a plausible consequent, C, of the causes represented by P. Abduction derives plausible causes, P, of the consequent C. Contingent deduction can thus be viewed as "tracing forward," and abduction as "tracing backward" such contingent, causally ordered, dependencies.

In sum, both contingent deduction and abduction are based on contingent domain-dependent relationships. Contingent deduction produces likely consequences of given causes, and abduction produces likely causes of given consequences. If a dependency is truly symmetrical (e.g., A <=> B), then the difference between contignent deduction and abduction ceases to exist.

Universal deductive inference is strictly truth-preserving, and universal induction is strictly falsity-preserving (if C is not true, then the hypothesis P cannot be true either). A universal deduction thus produces a provably correct (valid) consequent from a given premise. A universal induction produces a hypothesis that provably entails the given consequent (though the hypothesis itself may be false). Contingent deduction is truth-reserving, and abduction is falsity-preserving only to the extent to which the contingent dependencies involved in reasoning are true.

The intersection of the deduction and induction (i.e., an inference that is both truth-preserving and falsity-preserving for universal or true dependencies), represents an equivalence-based inference. Analogy can be viewed as an extension of such equivalence-based inference, namely, as a similarity-based inference. Every analogical inference can be characterized as a combination of deduction and induction. Induction is involved in the derivation of an analogical match (i.e., in determination of the properties and/or relations that are similar between the analogs), and deduction uses the analogical match to derive unknown properties of the target analog. Therefore, in the diagram, analogy occupies the central area.

As mentioned above, universal induction produces a premise that (together with BK) tautologically implies a given consequent. The tautological implication stems from the set-superset relationship. There are two types of universal induction: *inductive generalization* and *inductive specialization*. Inductive generalization is a widely known form of induction. For example, given that "*bean 1*, *bean 2,* and *bean 3* from a bag B are white" one may hypothesize that "All beans in bag B are white." Clearly, if the hypothesized premise "All beans in bag B are white," is true, then the given consequent, i.e., *bean 1*, *bean 2*, and *bean 3* from bag B are white, must necessarily be true.

Inductive specialization is a less known form of induction. To illustrate this form, suppose, for example, that we are told that

"There is a house in Virginia designed by Jefferson." (2)

Suppose that knowing (2), and having background knowledge that Fairfax is a town in Virginia, an agent hypothesizes that

"There is a house in Fairfax designed by Jefferson." (3)

This would be an example of inductive specialization. To see that this is a form of induction, notice that if (3) is true, then (2) must also be true (assuming the background knowledge is true).

In sum, induction is a process opposite of deduction, whose aim is to produce justifiable premises that entail consequents, or, in other words, justifiable explanations for the given facts.

These explanations can be in the form of generalizations, causal explanations, or both. The term "justifiable" is important here because induction is an underconstrained problem, and just "reversing" deduction could lead to an unlimited number of alternatives. For this reason, the "symmetry" between deduction and induction is only partial. To take into consideration the above idea, the previously given description of inductive inference based on (1) can be further elaborated. Namely, given a consequent C (observations, facts, rules, etc.), and background knowledge BK, the reasoner searches for a hypothetical premise P, consistent with BK, such that

$$P \cup BK \models C \tag{4}$$

and which satisfies the *hypothesis selection criteria.*

In different contexts, the selection criteria are called a *bias* (e.g., Utgoff, 1986), a *comparator* (Poole, 1989), or *preference criteria* (Michalski, 1983). These criteria are necessary for any act of induction because for any given consequent and a non-trivial hypothesis description language there could be a very large set of distinct hypotheses that can be expressed in that language, and that satisfy the relation (4). The selection criteria specify how to choose among them. Ideally, these criteria should reflect the properties of a hypothesis that are desirable from the viewpoint of the reasoner's (or learner's) goals. Often, these criteria (or bias) are hidden in the description language used (e.g., the language may be limited to only conjunctive statements involving a given set of attributes), or implied by the mechanism performing induction (e.g., a method that generates decision trees is automatically limited to using only operations of conjunction and disjunction in the hypothesis representation).

Generally, these criteria reflect three basic desirable characteristics of a hypothesis: *accuracy*, *utility*, and *generality*. The accuracy expresses a desire to find a "true" hypothesis. Because the problem is logically underconstrained, the "truth" of a hypothesis cannot be guaranteed in principle. To satisfy the entailment (4), a hypothesis has to be *complete* and *consistent* with regard to the input facts (Michalski, 1983). In some situations, however, an inconsistent and/or incomplete hypothesis may give a better overall predictive performance than a complete and consistent one (e.g., Quinlan, 1990; Bergadano et al., 1990). The utility requires a hypothesis to be simple and easily applicable for performing an expected set of tasks. The generality criterion expresses the desire to have the hypothesis applicable to a wide range of new cases. This facilitates the use of the hypothesis for prediction.

While the above described view of induction is by no means universally accepted, it seems consistent with some long-standing scientific thoughts on this subject going back to Aristotle (e.g., Adler and Gorman, 1987; see also the reference under Aristotle). Aristotle, and many subsequent thinkers, e.g., Bacon (1620), Whewell (1857), Cohen (1970) and others, viewed induction as a fundamental inference type that underlies all processes of creating new knowledge. They did not assume that the knowledge is created only from the low-level observations, without use of prior knowledge, and based only on universal dependencies.

Based on the role and amount of background knowledge, induction, as defined above, can be divided into *empirical induction* and *constructive induction*. In empirical induction there is little background knowledge, and the generated hypothesis is typically expressed using the same terms (attributes, relations, etc.) as the statements in the input (a consequent to be explained).

For example, the hypothesis may use the attributes selected from among those that are used in describing the instances in the input to induction. For this reason, such induction is sometimes called *selective* (Michalski, 1983). In contrast, a constructive induction would use background

knowledge to generate additional, more problem-oriented terms or concepts, and use them in the formulation of the hypothesis.

To illustrate different kinds of induction, below are few examples. To test if the inferences are inductive, one needs to see if, given BK and the hypothesis, the input is a logical consequence.

Empirical inductive generalization (background knowledge limited)

Input: The "A girl's face" is a beautiful painting. The "Lvow cathedral" is a beautiful painting.

BK: "A girl's face" and "Lvow cathedral" are paintings by Dawski.

Hypothesis: All paintings by Dawski are beautiful.

Constructive inductive generalization (backgound knowledge intensive)

Input: The "A girl's face" is a beautiful painting. The "Lvow cathedral" is a beautiful painting.

BK: "A girl's face" and "Lvow cathedral" are paintings by Dawski. Dawski is a known painter. Paintings are pieces of art. Beautiful pieces of art by a known painter are expensive.

Hypothesis: All paintings by Dawski are expensive.

Inductive specialization

Input: John lives is Virginia.

BK: Fairfax is a town in Virginia.

Hypothesis: John lives in Fairfax.

Abduction

Input: There is smoke in the house.

BK: Fire causes smoke.

Hypothesis: There is a fire in the house.

General (constructive) induction: (e.g., generalization plus abduction)

Input: Smoke is coming from John's apartment.

BK: Fire causes smoke. John's apartment is in the Golden Key building.

Hypothesis: The Golden Key building is on fire.

As mentioned earlier, in the most general formulation of induction, the union of BK and a hypothesis may only weakly entail the consequent. In such cases, the hypothesis could be logically inconsistent and/or incomplete with regard to the given input.

4. INFORMATION TRANSMUTATIONS: ABSTRACTION

As mentioned before, to derive desirable knowledge from a given input, a learner may perform various information transmutations. Such operations may, e.g., change the measurement units, reduce the size of the domain of the attributes, or generally reduce the amount of information conveyed by the input. These operations are not supposed to change the inherent meaning of the information, i.e., these operations are supposed to be truth-preserving. Thus, in the formal sense, these operations can be classified as forms of deductive inference. Three types of information transmutation can be distinguished:

1. Copying

The relevant information in the input is identified, and directly copied into the knowledge base of the learner. There is not change in the form of information. The learner or teacher must determine what parts of the input are relevant to the goal of learning.

2. Reformulation

The input information is transformed into another logically equivalent form. For example, mapping examples represented in a regular coordinate system into a radial coordinate system is a form of reformulation. A translation of the input description from one description language to another description language is also a form of reformulation. For example, a set of regular expressions is translated into an equivalent set of rules.

3. Abstraction

Abstraction reduces the amount of detail in a description of an entity (an object, or a class of objecs). It often transfers a description from one language to another that is more suitable for expressing the properties of the entity that are more relevant to the reasoner's goal. The purpose of abstraction is to reduce the amount of information about an entity in such a way that information relevant to the learner's goal is preserved, and other information is discarded.

Since abstraction is an important and often misunderstood operation, we give it special attention here. To elaborate further the descritpion obove, abstraction is defined as a process of creating a less detailed representation of a given entity from a more detailed representation of this entity, using truth-preserving operations. The latter means that the set of inferences that can be drawn from an abstract description of the entity is a subset of the inferences that can be drawn from the original description of that entity (given the same BK). In other words, details that are preserved should not suggest any new meaning that is not implied by the original description. To illustrate an abstraction operation, consider a transformation of the statement "My workstation has a Motorola 25-MHz 68030 processor" to "My workstation is quite fast". To make such an operation, the system needs domain-dependent BK that "a processor with the 25-MHz clock speed can be viewed as quite fast," and a rule "If a processor is fast then the computer with that processor can be viewed as fast." Note that the more abstract description is a logical consequence of the original description, but carries less information.

The abstraction process often involves a change in the representation language, from a one that uses more specific terms to a one that uses more general terms, with a proviso that the statements in the a second language must be logically implied by the statements in the first language. A very simple form of abstraction is to replace in a description of an entity a specific attribute value (say, the length in cm) by a less specific value (e.g., the length stated in linguistic terms, such as short, medium and long). A more complex abstraction would involve a significant change of the description language, e.g., taking a description of a computer in terms of electronic circuits and connections, and changing it into a description in terms of the functions of the individual modules.

The term "abstraction" is often confused with "generalization." To illustrate the difference between the two, consider a statement d(S,p), which says that the attribute d takes value p for the set of entities S. We will write such a statement in the form:

$$d(S) = p \tag{5}$$

Changing (5) to the statement d(S) = p', in which p' represents a more general concept (e.g., a parent node in a generalization hierarchy of values of the attribute d), is an abstraction operation. Changing (5) to a statement d(S') = p, in which S' is a superset of S, is a generalization operation. For example, transferring the statement "color(my-pencil) = light-blue" into "color(my-pencil)=blue" is an abstraction operation. Transforming the original statement into "color(all-my-pencils) = light-blue" is a generalization operation. Finally, transferring the original statement into "color(all-my-pencils)=blue" is both generalization and abstraction. In other words, associating the same information with a larger set is a generalization operation; associating a smaller amount of information with the same set is an abstraction operation.

An abstraction process is usually done to serve a certain goal, namely, to express only the information about some entity that is most relevant to a given task. An abstraction then can be viewed as a process of transforming knowledge from one form to another form, so that information relevant to a given goal is preserved, and irrelevant information is removed. Thus, formally, an abstraction is a transformation:

$$D_1(S) \quad \text{---}> \quad D_2(S) \tag{6}$$

such that $$INF_G(D_1, BK) \supseteq INF_G(D_2, BK) \tag{6'}$$

where $D_1(S)$ and $D_2(S)$ are descriptions of the set S in the same or different languages, and $INF_G(D_1)$ and $INF_G(D_2)$ are sets of deductive inferences, relevant to the goal G, that can be drawn about S from D_1 and D_2, respectively using BK. The goal defines what parts of the description are relevant and cannot be removed. Often, the goal of an abstraction process is only implicit. Since an abstraction is a truth-preserving process (from the viewpoint of the goal), it is a form of deduction.

To illustrate these ideas, let us take a *source statement* "John is 6 feet tall, weighs 190 pounds, has blue eyes, and lives in Fairfax." A transformation of this statement into a *target statement:* "John is a big man who lives in Virginia" is an abstraction of the source statement. To make this abstraction one needs to utilize BK that "Being 6 feet tall and weighing 190 pounds classifies one to be called big," and that "Fairfax is a town in Virginia." The implied goal here is that information about the height, weight and the place where a person lives is relevant to the goal of the reasoner, while the eye color is not. The abstracted statement clearly tells us less about John, but whatever can be inferred from it about John, can also be inferred from the original statement (given the same BK). The target statement does not introduce or hypothesize any more information about John. The goal is an important component in a general formulation of abstraction, because an abstraction process may introduce information that is incidental, and should not be taken into consideration while making inferences about the entity under consideration. For example, from an abstract drawing of a person one should not infer that the person is made out of paper.

In summary, generalization transforms descriptions along the set-superset dimension and is falsity-preserving (Michalski and Zemankowa, 1990). In contrast, abstraction transforms descriptions along the level-of-detail dimension, and is truth-preserving. While generalization

often uses the same description space (or language), abstraction typically involves a change in the representation space (or language). The reason why generalization and abstraction are frequently confused may be attributed to the fact that many reasoning acts involve both processes.

An opposite process to abstraction is *concretion.* Given an abstract description, concretion hypothesizes a description that has more details. Concretion is a form of inductive specialization.

In parallel to constructive induction, one may introduce the concept of constructive deduction. By analogy to constructive induction, constructive deduction is a process of deductively transforming a source description into a target description, which uses new, more relevant terms and concepts than the source description. As in constructive induction, the process uses background knowledge for that purpose. Looking at abstraction from this viewpoint, one may classify it as a form of constructive deduction. The latter is a more general concept than abstraction, however, as it also includes any other possible deductive knowledge transformations resulting in descriptions that contain concepts that were not present in the original description. For example, changing the problem representation space may be a from of constructive deduction, but not an abstraction. Also, constructive deduction may involve various forms of probabilistic (e.g.,Pearl, 1988), or plausible reasoning (e.g., Collins and Michalski, 1989). In the latter case, the distiction between constructive induction becomes unimportant, and is a matter of the degree to which different forms of reasoning are stressed.

5. A CLASSIFICATION OF LEARNING PROCESSES

Learning processes can be classified according to many criteria, such as the type of learning strategy used, the type of knowledge representation employed, the way information is supplied to a learning system, the application area, etc. Classifications based on such criteria have been discussed in, e.g., (Carbonell, Michalski and Mitchell, 1983) and (Michalski, 1986).

The inferential learning theory offers a new way of looking at learning processes, and suggests additional classification criteria. It considers learning as a goal-guided inference process that increases either the amount or the effectiveness of learner's knowledge. Therefore, as a major classification criterion it considers the main goal of learning. Based on this criterion, learning processes can be divided into synthetic and analytic. The main goal of synthetic learning is to acquire new knowledge, which goes beyond the knowledge already possessed, or the deductive closure of that knowledge. The primary inference type involved in synthetic processes is induction or analogy. The word "primary" is important, because every inductive or analogical inference also involves deductive inference. The latter form is used, e.g., to test whether a generated hypothesis entails the observations, to perform an analogical transfer of knowledge using the hypothesized analogical match, or to generate new terms using BK.

The main goal of analytic processes is to transform knowledge that the learner already possesses into the form that is most desirable according to the given learning goal. For example, one may know how to type on the typewriter, and through practice learns how to do it more rapidly. Or one may have a complete knowledge of how an automobile works, and therefore can in principle diagnose the problems with it. By analytic learning one can learn how to diagnose various problems more effectively using simple tests. From the viewpoint of the inferential theory, the primary inference type used in analytic learning is deduction. Well-known explanation-based learning methods are forms of analytic learning. Synthesizing a computer program from its specification is another form of this class of learning processes. Other important criteria include the type of input information, the type of primary inference employed, and finally, the role of the learner's background knowledge in the learning process.

Figure 4 presents a classification of learning processes according to all these criteria.

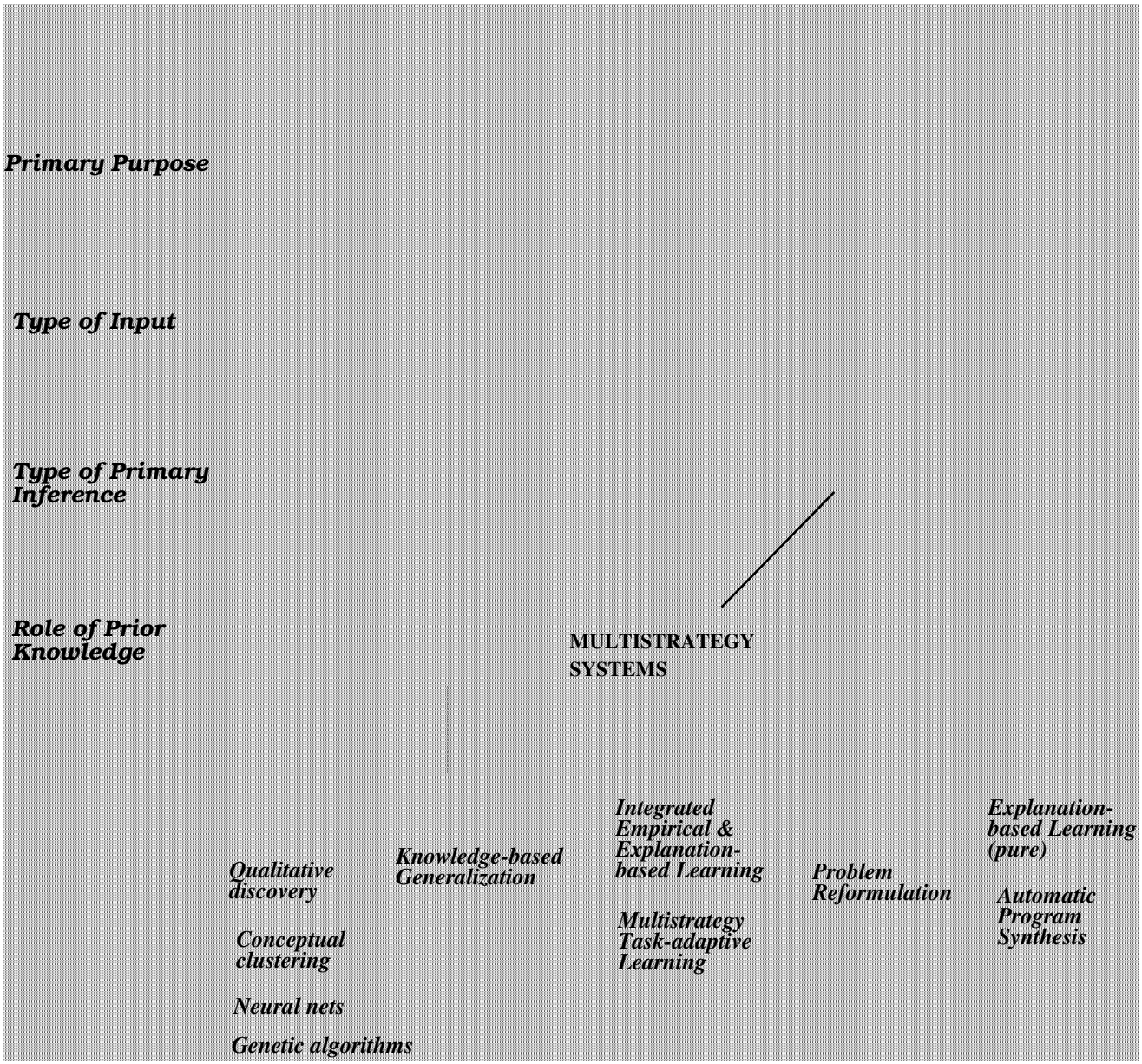

Figure 4. A general classification of learning processes.

This classification shows the basic characteristics of all major machine learning approaches and paradigms. As any classification, this classification is useful only to the degree to which it illustrates important distinctions and relations among various categories. The categories presented are not to be viewed as having precisely delineated borderlines, but rather as labels of central tendencies that can transmute from one to another by differently emphasizing various principal components.

If the input to a synthetic learning method are examples classified by an independent source of knowledge, e.g., a teacher, then we have *learning from examples*. When the input includes facts that need to be described or organized into a knowledge structure by the learner itself, then we

have *learning from observation*. The latter is exemplified by learning by discovery, conceptual clustering and theory formation.

The primary type of inference used in synthetic learning is induction. As described earlier, inductive learning can be empirical (BK-limited) or constructive (BK-intensive). Most work in empirical induction has been concerned with empirical generalization of concept examples using attributes selected from among those present in the descriptions of the examples. Another form of empirical learning includes quantitative discovery, in which learner constructs a set of equations characterizing given data. Learning methods employed in neural nets or genetic algorithms are also viewed as forms of empirical inductive learning. They typically rely on relatively small amounts of BK, and their primary inference type is inductive. This inference, however, is not executed in an explicit way, like in typical symbolic methods, but in an implicit way.

In contrast to empirical induction, constructive induction is knowledge-intensive, as it uses BK to create new and/or high-level characterizations of the input information (e.g., in terms of new attributes not present in the input). As described before, abduction can be viewed as a form of constructive induction, which "traces backward" domain-dependent rules.

To be more complete, let us mention that there are three other classifications of inductive methods, not shown in this classification. One is based on the way facts or examples are presented to the learner. If examples are presented all at once, then we have one-step or batch (non-incremental) inductive learning. If they are presented one by one, or in portions, and the system may have to modify the hypothesis after each input, we have an incremental inductive learning. Another classification is based on whether the input can be assumed to be correct, or that it can have noise, and therefore there is no requirement that a hypothesis must be complete and consistent with regard to the data.

The third classification characterizes methods based on the types of matching instances with concept descriptions. Such a matching can be done in a direct way, or an indirect way. The latter employs a substantial amount of background knowledge. For example, case-based or exemplar-based methods employ matching procedures that allow the system to recognize new examples that do not directly match any past example (e.g., Bareiss, Porter and Wier, 1990). Learning methods based on the *two-tiered concept representation* (Michalski, 1990b; Bergadano et al., 1990) also use sophisticated matching procedures.

Analytic methods can be divided into those that are guided by an example in the process of knowledge reformulation (example-guided), and those that start with a specification (specification-guided). The former category includes *explanation-based learning* (e.g., DeJong and Mooney, 1986), *explanation-based-generalization* (Mitchell, Keller and Kedar-Cabelli, 1986), and *explanation-based specialization* (Minton, Carbonell and Eztioni, 1987; Minton, 1988). If deduction employed in the method is based on axioms, then it is called *axiomatic*. The "pure" explanation-based generalization can be viewed as an example of an axiomatic method, because it is based on a deductive process that utilizes complete and consistent domain knowledge. This domain knowledge plays the role analogous to the axioms in formal theories.

Analytic methods that involve deductive transformations of description spaces are classified as methods of "constructive deduction." A major component of this class is abstraction, as it utilizes background knowledge to create descriptions at a lower level of detail, while preserving the truth

of the description. Other components of this class include transformation of problem representation spaces, determination of a representative set of attributes, etc.

Multistrategy learning systems integrate two or more learning strategies. Among the most widely known of such systems are Unimem (Lebowitz, 1986), Odysseus (Wilkins, Clancey, and Buchanan, 1986), Prodigy (Minton, Carbonell and Etzioni, 1987), DISCIPLE-1 (Kodratoff and Tecuci, 1987), GEMINI (Danyluk, 1987 and 1989), OCCAM (Pazzani, 1988), IOE (Dietterich and Flann, 1988) and POSEIDON (Bergadano et al., 1990). Other examples are in (Segre, 1989, Porter and Mooney, 1990). With few exceptions, most current multistrategy systems are concerned with integrating an empirical method with an explanation-based method. Some, like DISCIPLE, also include an analogical learning. The integration of these methods is typically done in a predefined, problem-independent way. An approach to building a *multistrategy task-adaptive learning* (MTL) is described in (Michalski, 1990a). An MTL system is supposed determine by itself which strategy or a combination thereof is most suitable for a given learning task.

The type of knowledge representation used in a learning system can be used as a separate dimension for classifying learning systems (not shown in Figure 4). That is, one could classify learning systems on the basis of the representation employed, e.g., a logic-style representation, production rules, frames, semantic network, grammar, decision tree, neural network, classifier system, etc., or a combination of different representations. The knowledge representation used in a learning system depends to a large extend on the application domain. It also depends on the type of learning strategy employed, as not every knowledge representation is suitable for every learning strategy. Thus, in parallel to multistrategy systems that combine several strategies, one can also distinguish multirepresentation learning systems that apply different knowledge representations in the process of learning. Such systems might employ various forms of constructive deduction or constructive induction to create and use representations at different levels of abstraction. The latter systems would thus be capable of changing the representation of the original problem statements. The importance of this area has been well acknowledged by pattern recognition researchers (e.g., Bongard, 1970), as well as by AI researchers (e.g., Amarel, 1986; Mozetic, 1989).

Summarizing, reasoning/learning processes can be described from three viewpoints characterizing the relationship between the input to the output:

- the direction of logical relationship: induction vs. deduction.
- the change in the reference set: generalization vs. specialization.
- the change in the level-of-detail: abstraction vs. concretion.

Each viewpoint corresponds to a different mechanism of knowledge transformation that may occur in a learning process, and involves two opposite operations. The operations involved in the first two mechanisms, induction vs. deduction, and generalization vs. specialization, have been relatively well-explored in machine learning. The operations involved in the third mechanism, abstraction vs. concretion (Webster's dictionary defines the latter as being a process of concretizing something) have been relatively less studied. Because these three mechanisms are interdependent, not all combinations of operations can occur in a single learning process (Michalski and Zemankova, 1991).

The above "grand" classification appears to be the first attempt to characterize and relate to each other all major methods and subareas of machine learning. As such it can be criticized on various

grounds. The ultimate goal of this classification effort is to show that diverse learning mechanisms and paradigms can be viewed as parts of one general structure, rather than as a collection of unclearly related components and research efforts.

6. CONCLUSION

The aims of this research are to develop a theoretical framework for characterizing and unifying basic learning strategies and approaches. The proposed inferential theory looks at learning as a process of knowledge transformations. Consequently, it attempts to analyze various methods of knowledge transformations, primarily, various types of inference. A classification of types of inference was proposed that relates to each other such basic types of inference as deduction, induction, abduction and analogy. It has been shown that in addition to widely known inductive generalization, one can also distinguish inductive specialization. Is has been also shown that abduction is a form of general induction, and abstraction is a form of deduction. Based on these concepts, a general classification of learning processes has been proposed.

Many discussed ideas are at an early state of development, and many issues have not been resolved. For example, future research should develop more precise characterization of various concepts discussed, and formal measures for characterizing various knowledge transformations. Another research topic could be to determine how basic operations of various learning algorithms map into the described knowledge transformations.

Concluding, the inferential theory of learning provides a new viewpoint for analyzing and characterizing learning systems. By addressing the logical capabilities and limitations of learning processes, it may ultimately produce a methodology for determining the competence aspects of diverse learning systems, and the areas of their most effective applicability.

Acknowledgments

The author expresses his gratitude to Hugo De Garis, Ken DeJong, Bob Giansiracusa, Heedong Ko, Yves Kodratoff, Elizabeth Marchut, Gheorge Tecuci, Brad Utz, Janusz Wnek and Jianping Zhang for insightful comments on various topics reported in this paper. Thanks also go to Janet Holmes and Susan Lyons for stylistic suggestions and proofreading.

This research was supported in part by the Defense Advanced Research Projects Agency under the grant administered by the Office of Naval Research No. N00014-K-85-0878, and in part by the Office of Naval Research under grants No. N00014-88-K-0397, No. N00014-88-K-0226 and No. N00014-91-J-1351.

References

Adler, M. J., Gorman (Eds..) The Great Ideas: A Synoptic of Great Books of the Western World, Vol. 1, Ch. 39, *Encyclopedia Britannica*, 1987.

Amarel, S., "Program Synthesis as a Theory Formation Task: Problem Representations and Solution Methods," in *Machine Learning: An Artificial Intelligence Approach Vol.II,* Morgan Kaufmann, Los Altos, CA, R. S. Michalski, J. G. Carbonell and T. M. Mitchell (Eds.), 1986.

Aristotle, Posterior Analytics, in *The Works of Aristotle*, Volume 1, R. M. Hutchins (Ed.), Encyclopedia Britannica, Inc., 1987.

Bacon, F., *Novum Organum*, 1620.

Bareiss, E. R., Porter, B. and Wier, C.C., PROTOS, An Exemplar-based Learning Apprentice, in *Machine Learning: AN Artificial Intelligence Approach vol. III*, Morgan Kaufmann, 1990.

Bergadano, F., Matwin, S., Michalski, R.S. and Zhang, J., Learning Two-tiered Descriptions of Flexible Concepts: The POSEIDON System, *Machine Learning and Inference Reports, No. MLI-3,* Center for Artificial Intelligence, George Mason University, 1990.

Birnbaum, L. and Collins, G., *Proceedings of the 8th International Conference on Machine Learning*, Northwestern University, Chicago, June 1991.

Bongard, N., *Pattern Recognition,* Spartan Books, New York, 1970 (translation from Russian).

Carbonell, J. G., Michalski R.S. and Mitchell, T.M., An Overview of Machine Learning, in *Machine Learning: AN Artificial Intelligence Approach,* Michalski, R.S., Carbonell, J.G., and Mitchell , T. M. (Eds.), Morgan Kaufmann Publishers, 1983.

Cohen, L.J., *The Implications of Induction*, London, 1970.

Collins, A. and Michalski, R.S., Logic of Plausible Reasoning: A Core Theory, Cognitive Science 13, 1-49, 1989.

Danyluk, A.P., "The Use of Explanations for Similarity-Based Learning," *Proceedings of IJCAI-87*, pp. 274-276, Milan, Italy, 1987.

Danyluk. A. P., "Recent Results in the Use of Context for Learning New Rules," *Technical Report* TR-98-066, Philips Laboratories, 1989.

DeJong, G. and Mooney, R., "Explanation-Based Learning: An Alternative View," *Machine Learning Journal*, Vol 1, No. 2, 1986.

Dietterich, T.G., and Flann, N.S., "An Inductive Approach to Solving the Imperfect Theory Problem," *Proceedings of 1988 Symposium on Explanation-Based Learning,* pp. 42-46, Stanford University, 1988.

Fulk, M. and Case, J. *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, University of Rochester, N.Y., August 6-8, 1990.

Goldberg, D.E., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

Haussler D. and Pitt, L. (Eds.), *Proceedings of the 1988 Workshop on the Computational Learning Theory* (COLT 88), Morgan Kaufmann Publishers, San Mateo, CA, 1988.

Kodratoff, Y. and R. S. Michalski (eds.), *Machine Learning: An Artificial Intelligence Approach Vol. III*, Y. , Morgan Kaufmann, Los Altos, CA, 1990.

Kodratoff, Y., and Tecuci, G., "DISCIPLE-1: Interactive Apprentice System in Weak Theory Fields," *Proceedings of IJCAI-87*, pp. 271-273, Milan, Italy, 1987.

Laird, J.E., (Ed.), *Proceedings of the Fifth International Conference on Machine Learning*, University of Michigan, Ann Arbor, June 12-14, 1988.

Lebowitz, M., "Integrated Learning: Controlling Explanation," *Cognitive Science*, Vol. 10, No. 2, pp. 219-240, 1986.

Michalski, R. S., "Theory and Methodology of Inductive Learning," *Machine Learning: An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell, T. M. Mitchell (Eds.), Tioga Publishing Co., 1983.

Michalski, R.S., Understanding the Nature of Learning: Issues and Research Directions, in *Machine Learning: An Artificial Intelligence Approach Vol. II,* Michalski, R.S., Carbonell, J.G., and Mitchell , T. M. (Eds.), Morgan Kaufmann Publishers, 1986.

Michalski, R.S., Toward a Unified Theory of Learning: Multistrategy Task-adaptive Learning, Reports of Machine Learning and Inference Laboratory MLI-90-1, January 1990a.

Michalski, R.S., LEARNING FLEXIBLE CONCEPTS: Fundamental Ideas and a Method Based on Two-tiered Representation, in *Machine Learning: AN Artificial Intelligence Approach vol. III*, Morgan Kaufmann, 1990b.

Michalski, R.S.and Ko, H., "On the Nature of Explanation or Why Did the Wine Bottle Shatter," *Proceedings of the AAAI Workshop on Explanation-based Learning*, Stanford University, March 1988.

Michalski, R. S. and Zemankova, M., "What is Generalization: An Inquiry into the Concept of Generalization and its Types," to appear in *Reports of Machine Learning and Inference Laboratory*, Center for Artificial Intelligence, George Mason University, 1991.

Minton, S., "Quantitative Results Concerning the Utility of Explanation-Based Learning," *Proceedings of AAAI-88*, pp. 564-569, Saint Paul, MN, 1988.

Minton, S., Carbonell, J.G., Etzioni, O., et al., "Acquiring Effective Search Control Rules: Explanation-Based Learning in the PRODIGY System," *Proceedings of the 4th International Machine Learning Workshop,* pp. 122-133, University of California, Irvine, 1987.

Mitchell, T. M., Keller, T., and Kedar-Cabelli, S., "Explanation-Based Generalization: A Unifying View," *Machine Learning Journal*, Vol. 1, January 1986.

Mozetic, I., Hierarchical Model-based Diagnosis, *Reports of Machine Learning and Inference Laboratory*, No. MLI89-1, 1989.

Pazzani, M.J., "Integrating Explanation-Based and Empirical Learning Methods in OCCAM," *Proceedings of EWSL-88,* pp. 147-166, Glasgow, Scotland, 1988.

Pearl J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.

Poole, D., Explanation and Prediction: An Architecture for Default and Abductive Reasoning, *Computational Intelligence*, No. 5, pp. 97-110, 1989.

Porter, B. W. and Mooney, R. J. (eds.), *Proceedings of the 7th International Machine Learning Conference*, Austin, TX, 1990.

Quinlan, J. R., "Probabilistic Decision Trees," chapter in *Machine Learning: An Artificial Intelligence Approach, Vol. III*, Y. Kodratoff and R. S. Michalski (eds.), Morgan Kaufmann, Los Altos, CA, 1989.

Schafer, D. (Ed.), *Proceedings of the 3rd International Conference on Genetic Algorithms*, George Mason University, June 4-7, 1989.

Segre, A. M. (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning*, Cornell University, Ithaca, New York, June 26-27, 1989.

Touretzky, D., Hinton, G., and Sejnowski, T. (Eds.), *Proceedings of the 1988 Connectionist Models*, Summer School, Carnegie Mellon University, June 17-26, 1988.

Utgoff, P. Shift of Bias for Inductive Concept Learning, in *Machine Learning: An Artificial Intelligence Approach Vol. II,* Michalski, R.S., Carbonell, J.G., and Mitchell , T. M. (Eds.), Morgan Kaufmann Publishers, 1986.

Whewell, W., *History of the Inductive Sciences*, 3 vols., Third edition, London, 1857.

Wilkins, D.C., Clancey, W.J., and Buchanan, B.G., *An Overview of the Odysseus Learning Apprentice*, Kluwer Academic Press, New York, NY, 1986.

Wnek, J., Sarma, J., Wahab, A.A. and Michalski, R.S., COMPARING LEARNING PARADIGMS VIA DIAGRAMMATIC VISUALIZATION: A Case Study in Concept Learning Using Symbolic, Neural Net and Genetic Algorithm Methods, *Proceedings of the 5th International Symposium on Methodologies for Intelligent Systems*, University of Tennessee, Knoxville, TN, North-Holland, October 24-27, 1990.