Reset reproduction of CMU Computer Science report CMU-CS-88-122. Reprinted July 1994.
 Current address: MIT Laboratory for Computer Science, Cambridge, Massachusetts.

On Rationality and Learning

Jon Doyle

Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

February 26, 1988

Abstract: Much work in machine learning views learning primarily as a process of representation. We propose instead to view learning as the rational interpretation of experience. This provides a unifying, precise framework that illuminates both the strengths and weaknesses of current learning methods and the possibilities for uniformly mechanizing different types of learning.

This research was supported by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 4976, Amendment 19, monitored by the Air Force Avionics Laboratory under Contract F33615-87-C-1499. The views and conclusions contained in this document are those of the author, and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Government of the United States of America.

1 Introduction

Considerable progress has been made recently on the problem of machine learning: as a practical technique for use in artificial intelligence systems, as a predictive psychological theory [Rosenbloom et al. 1987], and as a formal computational theory [Valiant 1984]. Many diverse types of learning have been studied, and for each of these a great variety of techniques and processes have been proposed. To some extent, this great diversity in types and techniques of learning stems from the richness of the subject. But it also may indicate the lack of some unifying notion that makes clear how the aims of these many techniques relate to learning and to each other. Lack of a unifying concept will not prevent progress from being made, but it can divert attention to topics of little relevance or poor prospects for success.

This paper restates the problem of learning in terms of the notion of rationality. This formulation has several advantages over earlier conceptions: it ties together several separate strands of research in a coherent view; it offers an explicit formal conception of and approach to learning rather than informal definitions; and it illuminates the sometimes severe limitations of some oft-studied techniques. We first present the rational conception of learning, and then discuss the strengths and weaknesses of some current approaches to similarity-based and explanation-based generalization from this perspective. Finally, we employ results from the theory of rationality to address the question of whether one can hope to uniformly mechanize the many approaches to learning.

2 What is learning?

At first glance, it seems surprising that there should be confusion about what learning is, as two good definitions are widely known. According to Simon [1983], learning "denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time." According to Minsky [1986], learning "is making useful changes in the workings of our minds." One can quibble with each of these definitions, but the quibbles are hard to pursue without making the definitions more precise.

These definitions notwithstanding, many studies in machine learning ap-

pear to be based on very different conceptions of learning. Some authors make no explicit statement of how the techniques they study constitute learning, while others seek more precise and specific definitions of learning than the above, but wind up with very different notions. Michalski [1986], for example, finds Simon's and Minsky's definitions too informal and promptly moves to redefine learning as "constructing or modifying representations of what is being experienced." This conception of learning is *very* different from the first two. Michalski tries to paper over the gulf between them, but does not really succeed, mainly because the third definition is a profoundly mistaken definition of learning, as we see shortly. Fortunately, this definition does not rule Michalski's technical work on learning: see, for example [Stepp and Michalski 1986]. But the conception of learning as representation has been extremely influential, both before and since Michalski articulated it, and characterizes much more of the literature than its competitors.

We suggest that a more illuminating definition is that *learning is interpreting experience by making rational changes of mental state or operation.* This means rationally deciding how to interpret one's sensory events as facts about what is going on, and then rationally deciding whether to change one's mental state in light of this information, and if so, in what way. This definition is still very general, but in this case precise, well-developed theories are available for each of the elements in the definition. For the notion of rationality we employ (for the moment) the standard notion from decision theory, according to which an action is said to be rational for an agent at some instant if it is of maximal expected utility according to the agent's beliefs and preferences about current and future events, where the agent's preferences may be a function of its goals and plans. We will not repeat the formal theory here as good expositions are readily available (e.g. [Jeffrey 1983]). The other element of the definition is the agent's fixed constitution or architecture, which sets out the possible states and changes through which learning takes place. We do not have space to elaborate any examples here. See, for example, [Minton 1988], which presents a system that knows enough about its own architecture to allow it to make essentially rational changes in its search strategies on the basis of its search experiences. In addition, precise definitions of many other sorts of architectures are extant, ranging from automata theory and programming languages to the more interesting conceptions of Bayesian agents [Jeffrey 1983] and knowledge-level architectures (e.g. [Genesereth and Nilsson

1987] and [Doyle 1988a]).

3 Memorization and generalization

To see the advantages besides precision offered by viewing learning in terms of rationality, it helps to examine some of the learning techniques commonly studied. We begin by considering the simple case of rote memorization of experience. This certainly counts as learning according to the representational definition. But the plain fact is that most things are not worth memorizing. For most things one experiences, such as the shape of the 20,000th tree leaf one sees, rote memorization is irrational, simply wasting resources and clogging memory. Similarly, memorization of the logical consequences of one's knowledge as they occur to one is also not genuine learning in most cases, because most consequences are not worth remembering either, whether because they are unlikely to ever be used, or because they are too easy to derive. That is, for most logical consequences of what one knows, memorization offers at best no utility (the memories do not help) and at worst negative utility (time and space are consumed in memorizing them and in discarding them when retrieved). The only cases in which remembering logical consequences is commonly considered learning is when the computational cost of deriving the consequence is too high, so that making the conclusion explicit makes its use, retrieval, or derivation economic. This is the motivation behind explanation-based learning (see below), and underlies some modern number-theoretic cryptographic schemes. In these, knowledge of the coding method, of number theory, of the language of the message, and of the coded message logically entails the identity of the uncoded message. But carrying out this inference is too costly unless the encryption key is known explicitly, so discovering the key is really learning something.

We next consider the problem of learning a concept from examples. This problem has been widely studied by researchers in artificial intelligence and theoretical computer scientists, not to mention mathematicians, psychologists, and many others. We will consider two main approaches: similarity-based generalization, and explanation-based generalization.

3.1 Similarity-based generalization

In Valiant's [1984] formulation of learning a concept from examples, one attempts to find a boolean characteristic function for a concept by examining a sample of the concept's extension (small errors are allowed). Here the concept's extension is a fixed target, and the ideal result of learning is a small boolean predicate that exactly characterizes the concept. This view of generalization is clearly representational learning, because it seeks to logically represent the meaning of a concept. It is called similarity-based generalization because the concept learned attempts to capture the similarities among the observed instances.

Since merely memorizing the items of experience and their logical consequences is not learning, concept learning must go beyond the agent's current information to make assumptions not entailed by current knowledge. The generalization itself may be such an assumption, or it may be derived as a logical consequence from separate assumptions in conjunction with experience. Ordinarily the agent's knowledge is incomplete, in which case there are many possible completions and extensions of the knowledge consistent with the evidence. Naturally, if one does not know whether P or $\neg P$ holds, one might assume either, but cannot consistently assume both. Generalization thus involves a selection of which assumptions to make out of the many possible sets of consistent assumptions. As Mitchell [1982] puts it, generalization is a search process whose product is the right generalization.

In fact, taking a broader view, learning involves making many choices in addition to selection of what conclusions to draw from the selected evidence. It can involve selection of the subject about which things will be learned; of the sources of evidence to be employed; of the criteria for determining relevance of potential evidence; of which bits of evidence are true and which are noise; of which differences among evidence are significant and which are insignificant; of how much evidence to seek; and of when to stop. Because it ignores these many choices, similarity-based generalization strikes many people outside the field as sterile, having little relevance to generalizations as they appear in everyday life. Scientific generalizations provide a good example, especially those appearing in debates about public policy. The most obvious fact about these debates is that the generalizations made depend on the debater's aims. In the first place, the opposing parties may differ on what evidence is relevant to the case. But even when they agree on the evidence, they may differ on how to in-

terpret it and on what extra assumptions to use in making generalizations. For example, the same economic statistics and theories that incriminate President Reagan's economic policies in the arguments of Democrats may exonerate his policies in the arguments of Republicans. (Bukharin, whose economic policies incriminated him in Stalin's court and exonerated him in Gorbachev's, should have been so lucky.) In the scientific debates associated with these policy debates, the parties may have aims other than to find the truth of the matter. They may seek to smear reputations, to win elections, or to muddy public opinion so much that a purely political decision can be made. In real-life generalizations, what conclusions one draws can depend on one's aims. But perhaps work on aimless generalization in artificial intelligence should not be faulted much. As Truesdell [1984] points out, most theories in the philosophy of science are similarly aimless, and the ideas on learning in artificial intelligence have been strongly shaped (sometimes unwittingly) by theories in inductive logic and the philosophy of science. (See also [Grabiner 1986]).

In contrast to everyday generalizations, Valiant's procedure is clearly not rational, as it generalizes in a fixed way independent of the domain or context in which the agent operates. The concept learned is independent of any conclusions the agent might desire to reach, and independent of any preferences about assumptions the agent might entertain. It embodies a fixed criterion of (or bias about) what the "right" conclusions are. In some cases this criterion may match up with the agent's preferences, so that its conclusions are rational, but in other cases this obliviousness to the agent's situation can prevent or delay the agent from learning what it needs to know, since generalizations are not always useful. For instance, most people who purchase a camera do not want to know the principles by which it operates, merely the specifics of how to use it. Determined salespeople or enthusiastic relatives who try to "explain" how to use the camera by explaining how cameras work are just asking for trouble. "Forget the theory, just tell me how to work it" is an eminently rational attitude for most people, for the generalization is irrelevant, merely cluttering memory and slowing down learning now and use later.

3.2 Explanation-based generalization

Explanation-based generalization differs significantly from similarity-based generalization by using knowledge to reduce the number of examples that must be examined. For concreteness we will refer to the EBG procedure of [Mitchell

et al. 1986], which uses domain knowledge to transform a "non-operational" target concept definition into an "operational" subconcept definition. To do this, it uses examples of the concept to find the specific knowledge relevant to the transformation, and uses a formal criterion of "operationality" of concepts to tell when it has completed the transformation. In brief, EBG constructs finds a proof from operational items of knowledge that the example satisfies the target concept, and then combines and generalizes the hypotheses of the proof to yield an operational definition of a subconcept.

Like Valiant's procedure, EBG does not learn rationally. It does not address the question of how the agent selects what to learn, and likewise assumes a fixed target concept which it never abandons. Instead, it limits its explanations to deductive proofs, which make the learned concepts logical subconcepts of the target. This means that EBG cannot handle exceptional instances that lead people to change their definitions (e.g. egg-laying mammals). This is not an essential limitation of EBG, but the fact is that deductive proof is too narrow a conception of explanation for general use, since replacement of concepts can be justified or rationalized as rational calculations, even if not as deductive proofs. In this regard EBG is sometimes even less rational than Valiant's procedure, since approximations to concepts can sometimes be more useful than the exact definition.

Even when one limits attention to learning subconcepts of the target, EBG does not learn rationally. EBG's central concern is using a notion of operationality as a guide to learning. EBG offers no theory of how to choose operationality criteria, only a requirement that such criteria express properties of the linguistic or representational form of the concepts. Unfortunately, there is no operationality criterion which expresses rationality, since rationality is a substantial condition applying to the new definition as a whole, not a formal one applying to the particles of the definition. Consider, for example, the case of visual learning. An example image of a cup might be reduced to millions of operational bit-predicates on a retina, but the resultant definition of "cup," with its millions of conjuncts, can hardly be called operational. Moreover, even if a simpler operational definition was possible, EBG has no way to choose between them.

EBG is an important procedure, but to perform well its inputs must be selected rationally and its outputs must be evaluated rationally. In fact, DeJong's [1983] original criteria for explanation-based learning concerned utility

of the result much more than operationality of the definition's elements. Keller [1987] also recognizes the limitations of formal operational criteria and proposes to redefine operationality to be usability plus utility (by which he seems to mean expected utility), with both usability and utility continuously variable in degree and dynamically changing. This is a step in the right direction. But it seems odd to call this combination "operationality." It would seem more natural to separate the apparently always binary but possibly changing notion of usability from the potentially changing and continuous notion of utility, calling the first a condition of operationality and the second a condition of rationality.

In addition to limitations stemming from using only deductive explanations and formal criteria of operationality, EBG also has no way to handle incomplete or inconsistent domain knowledge. EBG's authors recognize these limitations, and since they intend EBG to be a truly general procedure for learning, they suggest ways in which it might be extended to overcome them. These directions are worth pursuing, but we indicate in section 5 how handling incomplete and inconsistent knowledge poses profound problems that may forever limit the generality of EBG and other learning procedures.

4 Judging rationality

While it is easy to criticize some sorts of memorization and generalization as irrational, it is not always so easy to judge whether some form of mental reorganization is rational or not. The difficulty arises because judgments of rationality are very sensitive to the perspective of the judgment.

In the first place, even a nominally irrational learning method may be appropriate in the context of systems designed for specific purposes. Even if a learning method is irrational in the sense that it ignores the agent's preferences, it can nevertheless be rational for us to employ it as a part of the system's design if we expect that the agent will serve our purposes as well using it as using any other method, even methods more rational according to the agent's perspective. For example, there is a perspective (one which ignores certain computational costs) from which the chunking method of [Rosenbloom et al. 1987] is rational, in that systems employing it move along a demonstrated learning curve. Chunking itself, however, is an entirely mindless operation parasitic on a supposedly rational reasoner.

Secondly, the rationality of an action depends on the time frame over which we evaluate it. It is commonplace that actions rational in the short run may be irrational in the long run, and vice versa. In overall judgments of rationality, the agent must amortize its costs and benefits, taking into account the present value of future consequences of its actions. How this is done depends crucially on the agent's time preferences, on how much the agent prefers satisfying its goals now to satisfying them at different times in the future.

Thirdly, the basic theory of rationality involves only the notions of expectations and utilities, and ignores the familiar notions of goals and plans. As noted above, an agent's preferences may depend on its goals and plans, so that actions rational in the context of one set of goals may be irrational in another. For example, improving one's performance of routine actions is usually rational and so an aim of learning. But if the agent is under threats contingent upon completion of the actions, improving one's performance is no longer rational. Recall how Penelope slowed her weaving when her suitors demanded she admit Ulysses dead.

Finally, the definition of rationality mentions only the results of the action taken, not how it came to be taken. Thus rational learning does not mean that the agent must *calculate* what is rational to do. Of course, mechanization of rational learning can involve calculation of how to change the agent's mental state. This is the natural way to view the long-studied hill-climbing methods. It also serves as a basis for the bucket-brigade algorithm [Holland 1986] and for Minton's [1988] strategy learning system, which collects statistics to estimate expected utilities. Explicit rationality of learning is also reflected in the goal-dependent preference order on generalizations employed by [Stepp and Michalski 1986], in the similarity order on analogies employed by [Carbonell 1983, 1986], and in the preferences guiding shift of bias employed by [Russell and Grosof 1987]. (See [Doyle 1988c] for more on the relation between rationality, similarity, and shift of bias.)

5 Mechanizing rational learning

Since decision theory is a general theory intended to cover all sorts of rational choices, it is natural to ask if it provides a uniform way of mechanizing learning. That is, instead of having different procedures for each type of learning, can we instead use a single procedure which makes rational changes in all aspects

of mental organization? Unfortunately, this appears to be impossible because practical systems must make do with both more and less information than decision theory requires.

The lesser of these problems is that the agents we construct do not have all the information the theory of rationality requires. This incompleteness appears in both expectations and utilities. As informants, we often do not know what probabilities to assign to weird events, nor do we know the true costs of many actions. In particular, we often do not know in all cases how to assign value to time, and so cannot amortize costs and benefits. In practice, however, a number of techniques are available for ameliorating the consequences of incomplete information, including search, defaults, and adaptive estimation techniques.

The more serious difficulty faced in mechanizing rational learning is inconsistency, the occurrence of conflicts among the preferences that the agent must use to select its assumptions and changes. The most obvious examples of such conflicts are manifest in the general maxims which scientists claim to use as guides to formulating theories, maxims like "seek as simple a theory as possible," "seek as general a theory as possible," and "seek as powerful a theory as possible." These maxims are mutually incompatible. The most general theories may not be very powerful, and powerful theories often are not very simple. In each circumstance the theorist must choose which criteria to favor and which to downplay. These choices in turn can depend on the theorist's aims. For example, scientists studying some subject are apt to aim for the most general theory, while engineers studying the same subject are apt to aim for the most powerful theory.

Conflicts among a few general maxims may not be too hard to deal with, but multitudes of conflicting preferences arise naturally in coping with incomplete information. To make a long story short (see [Doyle 1988b] for a fuller treatment), it is quite rational to make routine assumptions by means of stereotypes or default rules. These rules save time and effort in common situations, so most artificial intelligence systems employ hundreds of such rules in representing their knowledge. It is natural to interpret these default rules as preferences of the agent about what assumptions to make, preferences of exactly the same kind appearing in learning about what generalizations to make. However, the cost of making reasoning more efficient in common situations is that stereotypes and default rules can conflict in uncommon situations. Thus

conflicting defaults represent conflicting preferences about assumptions.

The prevalence and individual rationality of these conflicting rules has serious implications for mechanizing rational learning. Each particular method for making assumptions (or for learning) based on default rules thus represents a way of resolving conflicts among preferences. But it is a standard result of decision theory and economics (called Arrow's theorem) that no single method of resolving conflicts can be completely rational without being unreasonable in a certain sense. The upshot is that there is a multitude of ways of proceeding in such cases, and each way has advantages and disadvantages in different circumstances. The literature on political economy has analyzed a number of these, but only a fraction of the possible methods (see [Mueller 1979]). Applied to the case of rational learning, these considerations indicate that there is no universal or general learning procedure. Every specific method of learning will either be irrational in some way, or will have built-in biases which it will exhibit in the choices it makes. Conflicts among different possible biases are real, reflecting a "clash of intuitions" (in the phrase of [Touretzky et al. 1987]) about what should be learned from experience.

6 Conclusion

Viewing learning as rational interpretation of experience captures many of our central intuitions about what learning is. Indeed, something like this view of learning may be close to the psychological truth about human learning, for according to Gazzaniga [1985], the central function of our conscious mind is to compulsively interpret and explain experience. In addition, this view provides a precise unifying framework that explains, justifies, or criticizes many important structures and techniques employed in extant learning systems, and illuminates some of the inescapable limitations all learning systems must face. Though these limitations may leave little hope for finding a good general or universal method of learning, there is nonetheless much room for improving the rationality of methods for learning.

Acknowledgments

I am grateful to Tom Mitchell for valuable comments on a draft of this paper, and thank Jaime Carbonell, Matthew Mason, B. K. Natarajan, Joseph Schatz, and Richmond Thomason for helpful discussions of this topic.

References

Carbonell, J. G., 1983. Learning by analogy: formulating and generalizing plans from past experience, *Machine Learning: An Artificial Intelligence Approach* (R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, eds.), Palo Alto: Tioga, 137-161.

Carbonell, J. G., 1986. Derivational analogy: a theory of reconstructive problem solving and expertise acquisition, *Machine Learning: An Artificial Intelligence Approach, Volume 2* (R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, eds.), Los Altos: Morgan-Kaufmann, 371-392.

DeJong, G., 1983. Acquiring schemata through understanding and generalizing plans, *Eighth Int. Joint Conf. on Artificial Intelligence*, 462-464.

Doyle, J., 1988a. Artificial intelligence and rational self-government, Pittsburgh: Carnegie Mellon University, Computer Science Department.

Doyle, J., 1988b. On universal theories of defaults, Pittsburgh: Carnegie Mellon University, Computer Science Department.

Doyle, J., 1988c. Similarity, conservatism, and rationality, submitted to *Seventh Natl. Conf. on Artificial Intelligence*.

Gazzaniga, M. S., 1985. *The Social Brain: Discovering the Networks of the Mind*, New York: Basic Books.

Genesereth, M. R., and Nilsson, N. J., 1987. *Logical Foundations of Artificial Intelligence*, Los Altos: Morgan Kaufmann.

Grabiner, J. V., 1986. Computers and the nature of man: a historian's perspective on controversies about artificial intelligence, *Bulletin of the American Mathematical Society* **15**, 113-126.

Holland, J. H., 1986. Escaping brittleness: the possibilities of general-purpose learning algorithms applied to parallel rule-based systems, *Machine Learning: An Artificial Intelligence Approach, Volume 2* (R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, eds.), Los Altos: Morgan-Kaufmann, 593-623.

Jeffrey, R. C., 1983. *The Logic of Decision*, second edition, Chicago: University of Chicago Press.

Keller, R. M., 1987. Defining operationality for explanation-based learning, *Proc. Sixth Natl. Conf. on Artificial Intelligence*, 482-487.

Michalski, R. S., 1986. Understanding the nature of learning: issues and research directions, *Machine Learning: An Artificial Intelligence Approach, Volume 2* (R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, eds.), Los Altos: Morgan-Kaufmann, 3-25.

Minsky, M., 1986. *The Society of Mind*, New York: Simon and Schuster.

Minton, S., 1988. Learning effective search control knowledge: an explanation-based approach, Ph.D. thesis, Computer Science Department, Carnegie Mellon University.

Mitchell, T. M., 1982. Generalization as search, *Artificial Intelligence*, Vol. 18, No. 2, 203-226.

Mitchell, T. M., Keller, R. M., and Kedar-Cabelli, S. T., 1986. Explanation-based generalization: a unifying view, *Machine Learning* Vol. 1, No. 1, 47-80.

Mueller, D. C., 1979. *Public Choice*, Cambridge: Cambridge University Press.

Rosenbloom, P. S., Laird, J. E., and Newell, A., 1987. Knowledge level learning in SOAR, *Proc. Sixth Nat. Conf. on Artificial Intelligence*, 499-504.

Russell, S. J., and Grosof, B. N., 1987. A declarative approach to bias in concept learning, *Proc. Sixth Nat. Conf. on Artificial Intelligence*, 505-510.

Simon, H. A., 1983. Why should machines learn? *Machine Learning: An Artificial Intelligence Approach* (R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, eds.), Palo Alto: Tioga, 25-37.

Stepp, R. E., and Michalski, R. S., 1986. Conceptual clustering: inventing goal-oriented classifications of structured objects, *Machine Learning: An Artificial Intelligence Approach, Volume 2* (R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, eds.), Los Altos: Morgan-Kaufmann, 471-498.

Touretzky, D., Horty, J., and Thomason, R., 1987. A clash of intuitions: the current state of nonmonotonic multiple inheritance systems, *Ninth International Joint Conference on Artificial Intelligence*, 476-482.

Truesdell, C., 1984. Is there a philosophy of science? *An Idiot's Fugitive Essays on Science: Methods, Criticism, Training, Circumstances*, New York: Springer-Verlag, 471-502.

Valiant, L. G., 1984. A theory of the learnable, *Comm. A.C.M.*, Vol. 18, No. 11, 1134-1142.