Typeset and lightly edited version of a manuscript draft privately circulated following its delivery in a 1981 lecture at Carnegie Mellon University. Manuscript written October 24, 1981; partially typeset November 4, 1981; typesetting completed and postscript added August, 1994.

Some Mathematical Problems in Artificial Intelligence

by Jon Doyle
October 24, 1981 — Pittsburgh, Pennsylvania

The purpose of this essay is to sketch a few mathematical problems arising in artificial intelligence, with the hope that some may appeal to mathematicians, logicians, or theoretical computer scientists. I have thought about these problems off and on for some time, though mostly off, and am now attempting to set some of them down in response to conversations with Merrick Furst and Joseph Bates, and in response to recent interest in non-monotonic logics.

I approach this task with some hesitation. Artificial intelligence is a complex field. As a science, it has major common interests with cognitive psychology, linguistics, philosophy, mathematical and philosophical logic, and the theory of computation. In addition, the engineering side of artificial intelligence has close ties to applied computer science (programming systems, methodology, algorithms, hardware) and to the fields it takes as domains for experimental systems. But more relevant to the purpose of this paper is the infancy of the field itself. Artificial intelligence is at present still largely concerned with problems of formulation rather than problems for solution. Sometimes invented techniques precede formulation of the problems they solve, but much work in AI can be seen as struggling towards a formulation of a problem. This can be seen in almost any conference proceedings—many of the papers have as their major burden explaining their point of view, the way they formulate the problem. Once one has a formulation, the implementations are almost always straightforward, at least in one's imagination. This may be an indicator of the difficulty of the field—one keeps floundering around until a viewpoint is developed that seems to admit a solution, in contrast to the initial viewpoints (or lack of them) implicit in the fundamental questions What is mind? and What is intelligent action? and How might one realize them?

With these disclaimers, I hope the problems I present do not seem too unmotivated. I have tried to pick out problems that seem relatively well established, that have appeared in many formulations over a period of time, or that appear to underlie many approaches held plausible in today's conventional wisdom. I have also attempted to avoid presenting problems that seem to involve much effort along philosophical or psychological lines in their development.

The problems of artificial intelligence discussed here all fall under four categories mapped out by the agent-action and structure-development dimensions. There is more to the field, but I cannot present that here—hopefully another work in progress will reach completion and widen the discussion here. This decomposition of the field is merely an initial approximation intended as an expository aid—for the structure of actions must find a mirror in the structure of agents, the structure of agents must permit the development of agent and action, and the development of agent and its actions is carried out through a series of actions. Expect to see problems arising repeatedly in the list that follows.

As a final warning: this is merely the first draft of this work. It is doubtless incomplete in its coverage, as mentioned above. Also, no strong effort has been applied to ensure the scholarly completeness of its references. In addition, some parts of these problems are currently under active investigation by the author, by others, and by some students. When so mentioned, the author hopes for the courtesy of non-competition on the part of the reader.

1 Conceptual Economy of Representations

We suppose the agent has a set of representations which it uses as its beliefs, desires, and other attitudes. We further suppose these representations can be viewed as a set of axioms in some (first-order) language. In such a view, the *concepts* employed by the agent are subsets of axioms, i.e., subtheories describing some particular domain (e.g., dogs, or arithmetic). The agent's beliefs will include axioms relating the various conceptual subtheories. For example, the subtheories of dogs and mammals might be related by the statement that all dogs are mammals, or that the dog-concept subtheory contains the mammal-concept subtheory.

With this view of the agent's beliefs as decomposed into many possibly overlapping subtheories, a number of questions arise which to my knowledge have received no careful mathematical treatment. The simplest, and most fundamental of these is that of what might be called the conceptual economy of the agent's representations—that is, where the set of concepts employed stands among all alternative conceptualizations and axiomatizations of the same theory. Logicians have of course studied alternative axiomatizations or bases for important theories like predicate calculus, arithmetic, and set theory, with attention to both the simplicity and elegance of the axiomatization, and to the notions captured in the theory. The extension of this problem of importance to artificial intelligence is to find some way of classifying the size and structure of whole theories of the world, not just numbers or sets. The applications and ramifications of this problem will appear in subsequently mentioned problems.

- Frankel, Bar-Hillel, and Levy, Foundations of Set Theory
- Fahlman, NETL: A system for Representing and Using Real-World Knowledge

2 Relative Expressibility of Conceptualizations

One application of the agent's understanding of the conceptual economy of its beliefs is for it to modify its conceptualizations to employ more economical representations. (We discuss this question further in a later problem.) But any consideration of alternative conceptualizations must take into account their relative expressibility as well as their sizes. Relative expressibility is simply the question of how the notions of one theory may be defined in another. We may know that two theories have identical sets of theorems, in each theory, the axioms of the other theory may either be simple consequences or simple definitions, or may require long proofs or extremely complex definitions. What is lacking is any classification of such relative expressibility problems. For example, one might ask if there are sets of theorems which admit a bound on the complexity or logical and computational translations over all pairs of alternative axiomatizations? Are there sets of theorems which allow no such bound? Given two theories with the same theorems, is the complexity of translation of one into the other comparable to the translation in the other direction? (I.e., are there one-way trapdoor translations?)

A related question concerns the problem of telling whether one has alternative axiomatizations in the first place. It is of course not decidable whether two sets of axioms have the same consequences, for this would involve deciding that each of the axioms of one set are theorems of the other. Can one classify any decidable or otherwise "easy" cases of the alternative-axiomatization question?

Related to this determination of alternativeness is the question of equivalence of ontologies. Often one comes up with candidate alternative theories by making multiple descriptions of the same things in the world. In fact, this use of such multiple descriptions is popular and powerful in AI because of pessimistic intuitions about the answers to the questions posed above. But there are many reasons for being very cautious about candidate alternative axiomatizations produced by repeated descriptions of the world. The problem here involves all the questions of the reducibility or independence of one domain over another, the possibility that the same object may admit non-mutually-expressible descriptions. Debates have raged about reductions of chemistry to physics, of biology to chemistry, of psychology to biology, and of economics to psychology. Can a mathematical theory of relative expressibility say anything about these problems?

Hands Off: One of Merrick Furst's students is about to complete a thesis on these questions—wait a month and Merrick promises the full story.

- Pratt, Powers of bases for propositional calculus
- Fodor, Special Sciences
- Fodor, Computation and Reduction
- Putnam, Philosophy and our Mental Life
- Armstrong, A Materialist Theory of Mind
- Ryle, The Concept of Mind
- Sussman, SLICES: At the boundary of analysis and synthesis
- Steele and Sussman, Constraints
- Minsky, A Framework for Representing Knowledge
- Quine, Ontological Relativity and other essays

3 Meta-theory and Conceptual Economy

The use of meta-theoretical devices in representations is becoming more widespread in AI for a number of reasons. One example is the use of meta-theoretical statements to relate the dog-concept and mammal-concept theories rather than implications. A simple example of the added power of this approach is the family resemblance problem. Suppose one has a description of the prototypical member of a family, but each member differs from the prototype in one small regard. If one can relate theories only by implication, one cannot relate the theories describing the prototype and the individual members, so description of the prototype allows no conceptual economy. On the other hand, if one has meta-theoretical tools available, one can relate the prototype to the individuals with statements like "Individual I shares all axioms with prototype P except for statement F." This reduces the size of the description of the family members to the size of the prototype *plus* the number of members, rather than the size of the prototype *times* the number of members.

This example may seem contrived, but I believe it accurately captures many of the difficulties known about the so-called *natural kind* terms of ordinary language, nouns like cat, gold, lemon, etc. If this is true, one can expect such family resemblance problems to riddle any conceptualization of the world. The mathematical problem is to characterize the relative economies and expressibility of meta-theoretical and non-meta-theoretical theories.

- Wittgenstein, Tractatus Logico-Philosophicus (I think)
- Putnam, Is Semantics Possible?
- Doyle, A Model for Deliberation, Action, and Introspection
- Doyle, A Theory of Memory (draft)

4 Defaults and Conceptual Economy

A related problem to the previous one, perhaps even a special case, is that of the economic power of default statements. Defaults are general rules for adopting conclusions which allow any particular conclusion to be defeated. (We treat the logical status of defaults later in problem 18.) To build on our previous example, we might phrase the description of the prototypical family member entirely in terms of defaults: by default a big nose, by default red hair, etc. Then the theories of each actual family member might be specified as a copy of the theory of the prototype, plus an axiom defeating some particular default. This formulation appears to have the advantage of allowing the separation of the statements of the individual-prototype sharing and individual peculiarities in a way not convenient in the pure meta-theoretical version given above. AI experience has indicated that any substantial description of the world must involve many rules for drawing such default conclusions.

The mathematical problem here is to characterize the power of expression of default theories: given a conceptualization in which concepts are expressed using defaults, is there an equivalent theory not involving defaults, and if so, what are their relative complexities? The intuition is that defaults allow substantial economies—one need not say for everything whether some predicate holds or not, but can just say the positive instances and subsume all the negative cases in a default.

- References for Problem 3 plus
- Reiter, Closed-World data-bases
- Reiter, On reasoning by default

- Reiter, A logic for default reasoning
- McCarthy, Circumscription: a form of non-monotonic reasoning
- Reiter, Logical Theory of Databases (in preparation - on this topic)

5 Inferential Economy of Representations

In addition to our assumptions about the attitudes of the agent and their representation, we also suppose the agent has a set of procedures which it can employ in making deductive inferences, or more generally, in taking mental acts that change its sets of beliefs, desires, etc., whether or not these inferences are deductively sound or not. Widening our scope to include the set of procedures of the agent (and assuming this set fixed for the time being—its evolution will be discussed later), we can ask for a classification of the inferential economies of conceptualizations and axiomatizations. The most important problem here is developing some structure in the tradeoffs between complexities of axiomatizations and complexities of proofs. As is well known, compact axiomatizations may require much longer proofs than more redundant and richer axiomatizations. The previous questions discussed ways one might measure the complexity of conceptualizations. Following questions will discuss different possible measures of the complexity of proofs. But given these notions, can one relate then in general, or in particular theories? One very important problem is given specified resource limits (say, on the number of axioms allowed and the length of proofs) can one tailor one's set of concepts to meet these limitations? What are the tradeoffs between effort and foresight? Can precise tradeoff characterizations be used to formulate precise notions of relevance or perceived relevance of one statement to another?

- Hayes, The Naive Physics Manifesto
- Kowalski, Logic for Problem Solving
- Kowalski, Algorithms = Logic + Control
- Rogers, Theory of Recursive Functions and Effective Computability
- Garey and Johnson, Computers and Intractability: A guide to the theory of NP-complete problems
- Rabin, Theoretical impediments to artificial intelligence
- Rescher, Restricted inference and inferential myopia in epistemic logic
- Minsky and Papert, Perceptrons: An introduction to computational geometry
- Davis, Obvious inferences

6 Complexity of Inferences and Arguments

One important problem is characterizing the complexity of inferences and arguments (proofs) in the context of our ideas of the structure of agents. Where ordinary complexity of theory looks at time and space for Turing machines, and space and retrieval time for random access machines, we here want complexity measures relevant to conceptualizations as the memory store. Of course the traditional measures of size apply, but other measures seem important in the setting of an agent trying to decide what to do on the basis of its beliefs. Where ordinary complexity theory assumes the set of possible actions as a given, for AI the decision of what to do is one of the problems. This means other measures of complexity may be of great importance in estimating the time required for the agent to retrieve information, make decisions, and act.

For example, one might look at proofs and arguments as graphs, and study their graph theoretic properties in addition to their size. One might expect the difficulty of constructing an argument to depend both on the variety of information used in it as well as on the intricacy of the ways in which the information is combined. In the former case, one seeks measures of the size of basis of a proof, of the variety of concepts and axioms entering into it, the "working set" or "set of support" of the proof. If one has a measure of relevance on concepts, this could be combined with the idea of working sets to say something about the difficulty of construction of arguments whose working set members may be of distant relevance. For the latter case, one can seek to measure the "locality" of the proof with topological measures of the graph's genus, etc.

- Minsky and Papert, Perceptrons: An introduction to computational geometry
- Abelson, Towards a theory of Local and Global in Computation
- Statman, Structural complexity of proofs
- Denning, The Working Set model for program behavior

7 Meta-theory and Inferential Economy

In parallel with problem 3, we might ask for a characterization of the power of meta-theoretical techniques in constructing proofs. For example, the formula

equation omitted in manuscript—JD

may be very difficult to prove using only ordinary rules of inference, while it is easy to prove if one admits the meta-theorem or derived rule that formulas involving only equivalences are true iff each proposition occurs an even number of times.

Can one relate the complexity of theories with particular derived rules to the complexity of theories without them? I don't quite know how to formulate the problem here, but feel there is one.

- Weyhrauch, Prolegomena to a theory of mechanized formal reasoning
- Milner et al., LCF and ML

- Bates and Constable, The definition of μPRL
- Martin-Löf, Constructive mathematics and Computer Programming

8 Summarization of Arguments

One important ability one would like in an intelligent agent is the ability to summarize the arguments or proofs it constructs, to create lemmas or derived rules for later use, thus avoiding the cost of reconstructing the argument. This ability is also important in communicating or explaining the result to other agents, and in analyzing weaknesses in its belief or inference system (to be discussed later).

Can one develop a mathematical theory of summarization? Some means is needed for stating the purpose of the desired sort of summary, and means for checking the satisfactoriness of candidate summaries. For example, one might define classes of obvious inferences and shared (or known) lemmas and reduce an argument to its residue when these inferences and facts are eliminated. Or one might define some notion of criticality of steps and facts used in arguments, and present the summary in terms of the critical elements of the arguments. The unobvious inferences and unshared axioms might be critical, or one might look to the structural complexity measures discussed above to pick out the most consequential of the axioms or steps, or the least likely to be noticed given a relevance measure. Perhaps restrictions on intelligible summaries, or motivations for wanting to summarize, can be phrased in terms of the structural complexity, in terms of the argument topology and working set.

- Davis, Obvious inferences
- Lehnert, Summarizing narratives
- Doyle, A truth maintenance system

9 Theory of Search

The structure of actions involved in searching some set space of possibilities (a tree or a graph) has long been studied as the theory of search. Several algorithms for searching are known, and can be guaranteed to find the sought node given certain sorts of information about each node and its relation to the sought node. Recently, these correctness and termination results have been supplemented with analyses of ordinary algorithmic complexity, measured in number of nodes examined, and with experimental measurements of expected efficiency.

Search theory has many connections to the developmental problems surveyed below, for developmental or learnability problems can be cast as reachability problems in a search space. However, in actual practice, few such castings of developmental problems as formal search problems have been carried through. A severe complication faced in doing so is that the most natural conception of progress in solving problems is one in which the agent is continually reformulating the problem it thinks it wants to solve. Traditional search theory explores the case of very simple sorts of problem reformulations, namely changes from “get from node N in space S to goal node G” to “get from node N' in space S to goal node G”, where N and N' are related in some way that varies with the search algorithm. Traditional search theory avoids, however, problem reformulations in which the whole search space S and goal G are traded in for new ones. One can make the analogy with Kuhn’s theory of scientific development: “normal” science is search within a fixed search space, and “revolutionary” science is search which involves replacement of the whole space. So far no detailed analyses of these more complex problems have been attempted, though various programmatic statements have been made.

- Nilsson, Principles of Artificial Intelligence
- Knuth and Moore, Analysis of α-β pruning
- Berliner, The B* tree search algorithm: a best-first proof procedure
- Gashnig, Performance measurement and analysis of certain search algorithms
- Kuhn, Structure of Scientific Revolutions
- Glymour, Theory and Evidence
- Minsky, A Framework for Representing Knowledge
- Lenat, AM: An artificial intelligence approach to discovery in mathematics as heuristic search

10 Reachability and Continuity in Development

In studies of the development of the agent’s structure or its actions, one can distinguish questions of reachability and questions of continuity. Questions of reachability ask if certain final states of agent structure or agent actions are possible given a set of inputs to the agent. These questions bear closely on the question of induction generally—what conclusions can be deduced from a given body of data. Reachability questions have been studied in several forms, discussed below as separate problems. Questions of continuity, on the other hand, ask which structures of the agent or its actions can be had as “continuous” or “incremental” changes of given structures. Questions of continuity, thus, concern the process of development. The two sorts of questions are naturally combined in theories of development to questions of reachability via continuous steps—questions on the surface very similar to the problems of graph searching mentioned earlier, where reachability questions translate into questions about the boundaries or composition of the search space (its nodes) and continuity questions translate into the arcs of the graph connecting the nodes. This separation of questions gives us a somewhat different perspective than that of search theory, in which search spaces are normally defined in terms of the nodes continuously reachable from a given node. Particular classes of continuity and continuous reachability questions are discussed as separate problems below.

Can one develop theories of abstract reachability and continuity so that one can discuss tradeoffs? That is, given a notion of reachability and a notion of continuity, can one say how refined or widened notions of continuity change the set of continuously reachable nodes?

And given a notion of reachability, can one characterize what sorts of continuity relations are possible? All of these questions must be formulated in terms of some decomposition of nodes—in terms of “atoms”, where reachable nodes are “possible molecules” and continuity relations on “molecules” are generated by continuity relations on “atoms”.

[Are there Betti numbers for theories and problems?]

Another question that arises in this framework of reachability and continuity is that of limits. Are there limit points in the learning space? Can one derive the theory of practice, with its power-laws and diminishing returns, from topological considerations?

- Kelly, General Topology
- Alexandroff, Elementary Concepts of Topology
- Whitehead, Homotopy Theory
- Hilton and Wylie, Homology Theory

11 Parameter learnability

Minsky and Papert’s study of perceptrons focused on a particularly simple sort of machine that computed with sets of numerical parameters. Outside of their negative results concerning the computational powers of such machines, they were also able to show that for certain perceptron-computable functions the parameters must be astronomical in size compared to the size of the input. One can make an analogy with Gödel numbering and view this phenomenon as the result of having to code a table of values in a single set of numbers.

For perceptrons, can one characterize the sets recognizable with bounds on the sizes of parameters, i.e., characterize the “space” hierarchy for perceptrons?

Can one generalize these results to other sorts of machines in interesting ways? For Gödel numbering problems, elementary results say that any pairing function of integers must grow as the product of the absolute values of the integers. Are all natural generalizations of parameter encodings related to space utilization in counter automata, and hence to space utilization in other sorts of machines as already studied in computational complexity?

- Minsky and Papert, Perceptrons: An introduction to computational geometry
- Garey and Johnson, Computers and Intractability: A guide to the theory of NP-complete problems
- Minsky, Computation: Finite and Infinite Machines
- Ehrenfeucht and Mycielski

12 Parameter learning

Along with their results on learnability for perceptron parameters, Minsky and Papert prove the perceptron convergence theorems about the efficacy of parameter adjustments in response to success and error. Have all the interesting questions been solved? Perhaps not. For example, they ask for a comparison of the relative efficiency of perceptron convergence algorithms and related, non-incremental algorithms for solving linear programming problems.

- Same references as for Problem 11

13 Grammar learnability and learning

Initial studies have produced several results about the learnability of grammars, given grammars of certain recursiveness classes and example sentences of certain forms. Other work has also investigated the learnability of restricted grammar classes under certain local grammar modification rules. What is still lacking is a mathematical classification of learnable languages which takes into account not just the recursiveness class of the language, but also other constraints, such as sorts of input data, continuity of learning rules, and existence of efficient parsers. The problem here is that we still lack all but the most crude characterizations of these subsidiary constraints (e.g., interpreting "existence of efficient parsers" to mean simply "context-free").

- Gold, Language identification in the limit
- Hamburger and Wexler, A mathematical theory of learning transformational grammar
- Wexler and Culicover, Formal Principles of Language Acquisition
- Berwick, Locality Principles and the acquisition of syntactic knowledge
- Berwick and Weinberg, Parsing efficiency, computational complexity, and the evaluation of grammatical theories

14 General theory of inductive inference

Solmonoff, Blum, and others have developed basic results about the induction problem in the setting of finding Turing machines of minimal size which recognize the input sequence of data. These results say something about all learnability problems, but the detailed connections may not have been pursued. For example, work is still carried out in AI on learning procedures from examples (to be discussed later in more detail), but no clear connections have been made to the general results.

One might pursue these general results along different lines. The criteria of minimal recognizer size is a powerful one, and has connections with the Chaitin-Kolmogorov theory of randomness. However, in light of our earlier problems, one might explore induction based on given space-time resource limits, rather than just space alone. One cannot ask simply for minimum time recognizers, for any finite set of data can be coded directly into the state table of the machine. But does this trivialize all restrictions on time?

- Solmonoff
- Blum
- Chaitin, A measure of program size formally identical to information theory

15 Concept learning and learnability

Much work still proceeds in AI on learning of concepts, where we may view this work as the automated formation of logical theories describing the input examples. It would be nice if this work was related to other sorts of learnability results as mentioned above, but to my knowledge it has not yet been so related.

Winston made one of the earliest studies of concept learning from examples and non-examples, and his apparent intent was to construct minimal conceptual complexity descriptions of the examples. Later work has followed along these lines. Fahlman makes suggestions about learning techniques which mix both conceptual complexity and inferential complexity measures, but as far as I know, no detailed work has been done on minimal inferential complexity learning algorithms or on combinations of the two. There may be a better chance of obtaining an interesting theory here than in the Turing machine case discussed before because databases of concepts and axioms need not be fully associative. This means one may not be able to simply add the input example to the database and have it be rapidly retrievable; instead, a full search of the database may be necessary. Thus theories of concept induction involving inferential complexity must be built on particular notions of database structure and retrieval procedures. Perhaps this sort of consideration can be reflected back into the Turing machine case by assigning costs to state-table accesses, but given the restricted nature of Turing machine operations, the prospects seem bleak for doing this in such a way that minimal time complexity does not imply minimal machine complexity.

- Winston, Learning Structural Descriptions from Examples
- Mitchell
- Fahlman, NETL: A system for Representing and Using Real-World Knowledge
- Lenat and Harris, Cognitive Economy
- Langley

16 Circumscription and concept learning

A very interesting alternative to characterizing minimal conceptual complexity concept learning (such as Winston's) is in terms of circumscription. All the induction procedures explored make the tentative assumption that the concept aimed for involves only what the presented examples do, and not anything else. This, in fact, is just the sort of inference that can be stated formally in terms of circumscription. For example, Reiter has pursued the use of circumscription in answering database queries, in which completions of the database are induced from the actual entries contained by means of circumscription.

The mathematics of circumscription is still full of problems. One of the most serious is that of choosing instantiations for the inductive schema one develops by applying circumscription to data. The general problem of automated deduction from schema is intractable, but can one get most of what is needed with simpler techniques? Reiter explores a simple-minded instance of the schema which can be constructed in a straightforward manner and which permits drawing the desired results in many cases. Can this be improved on? Also, can one develop a classification of the difficulty of the circumscription problem as a function of the input theory? Is there ever a "maximal" schema instance which alone generates the entire set of circumscriptive theorems? If there is no "maximal" instance, can one say how many there are? What their syntactic size is? Can one precisely relate minimal Turing machine formulations of induction to the minimal-model formulation underlying circumscription?

- McCarthy, Circumscription: a form of non-monotonic reasoning
- Reiter, Logical Theory of Databases
- Davis, The Mathematics of Non-Montonic Reasoning

17 Procedure learning and learnability

The learning of procedures is closely connected to the learning of concepts in some respects, in that one can take the concept being learned to be a description of a procedure—the instances of the concept corresponding to the instances (input-output pairs) of the procedure. This connection works both ways. We get all the results about concept and Turing machine learning in the context of learning of procedures, and we can feed back results about procedures to results about concepts.

One promising area of investigation draws on the connection between computation and deduction, and applies proof-theoretic techniques to both extract procedures from proofs, and to refine procedures by manipulating proofs. For example, given a notion of the procedural interpretation of logical formulae, as in PROLOG, one can attempt to prove theorems from an initial set of axioms which represent a more efficient procedure. Alternatively, one can take the proof and manipulate or refine it to produce a more efficient program. Aside from the exploration of these ideas in themselves, can one use them to formulate interesting notions of inferential complexity of concepts for use in concept learning theories?

The problem of learning of procedures is more general than that of concept learning, however, since procedures may have side-effects, that is, change the state of the world. The procedures corresponding to concepts are not of this character. For procedures of actions, as I will call them, the problem is not so much learning from examples as learning from errors, where errors show up as failed executions or unexpected effects. Learning of these sorts of procedures is still in its infancy because analysis of failures entails some notion of the intentions of the procedure. This connection with the mental attitudes of the agent means this problem is much more complex, and not likely to be purely mathematical, at least without a mathematical psychology and theory of action to ground it in. The theory of procedure correction also depends on the decomposability of domains, of the ability to localize causes. Mathematical models of the world which allow such characterization of the complexity of causal relations are largely yet to be developed. (See the problems on revision of attitudes below, however.) It might be possible, however, to characterize the complexity or possibility of procedure learning by procedure correction given specified sorts of causality. If errors can always be traced to single causes, the problem should be simple, and should admit rapid learning procedures. If errors can have multiple causes, things should be less nice.

A satisfactory analysis of causality and types of errors should allow formulation of a generative theory of test-cases for procedures, so that learning might progress by testing procedures at their weakest points.

- Goad
- Clark and Sickel
- Sussman, A Computer Model of Skill Acquisition
- Bates and Constable, The definition of μPRL
- Goldstein and Miller
- de Kleer

18 Logics of Assumptions and Inference

The default formulation of concepts requires a careful determination of the meaning of defaults, since defaults might conflict, and in any event appear to have unusual properties. The task of characterizing interpretations of sets of axioms and defaults has been studied as default logic and as non-monotonic logic. I have recently developed a unified extension to these approaches, which I am writing up. There are many interesting mathematical problems yet to be solved, but I leave them for that exposition.

- Reiter, A logic for default reasoning
- McDermott and Doyle, Non-monotonic logic I
- Doyle, Logics of Assumptions and Inference, in preparation

19 Revision of Attitudes

If logics of assumption and inference characterize the “reachable” sets of attitudes, i.e., the coherent interpretations of axioms and defaults, the recent work on revision of attitudes addresses the related continuity questions. The attitude revision procedures are mappings from sets of attitudes and additions to new sets of attitudes which both preserve coherence (as formulated by the logics) and which incorporate notions similar to Quine’s “minimum mutilation” principle. Unfortunately, these notions of distance for attitude sets have to date been left implicit in the programs, so one problem is their precise characterization given the language of the logics of assumption and inference.

One unexplored question is that of what sets of attitudes are reachable from a given set. McDermott and Doyle have raised this problem for the pure reachability problem, but we can also ask it for the continuous reachability problem.

One might also formulate other sorts of notions of continuity than minimum mutilation. Touretzky (*Hands off this one, folks*) is developing a thesis in which the continuous changes also preserve determinateness (non-ambiguity) as well as coherence of assumptions.

Finally, the existing algorithms for attitude revision need lots of exploration, analysis, and improvement, especially in light of the subsequent work on their logics. What is their worst case and average complexity? This should be asked for several notions of complexity—running time, number of statements examined, number of statements changed, etc. Some of the answers are bound to be disappointing—e.g., in the current algorithms one might have to examine an arbitrarily large set of conclusions even if one changes none of them. Can other algorithms be devised which avoid this? And what are the tradeoffs between algorithms which allow easy access to several (e.g., chronological) states of attitudes vs. those which do not?

- McDermott and Doyle, Non-monotonic logic I
- Doyle, Logics of Assumptions and Inference
- Touretzky, in preparation
- McAllester
- McDermott

Postscript—August 15, 1994

Complaining of an acute attack of *la vida breve*, and inspired by Minsky's example in [3], I wrote up these notes and delivered them in a Carnegie Mellon University Computer Science Department seminar. I typed up the first half, but never completed the task of typesetting and revising it—delayed indefinitely, I can guess, by the rapid progress I was making in my thinking at the time. In lieu of a properly typeset copy, I gave out photocopies of the manuscript to various people over the years.

For this reprinting, I have limited my editing to correction of misspellings and minor typos, and the addition of the partial and partially-specified bibliography I prepared at the time, which I have incorporated into the text by putting the references for each problem at the end of the problem section. Rereading the text tempts me to add many more historical notes concerning the state of knowledge at the time, the evolution of these ideas in my then future work, and the ways that AI eventually addressed some of them. But I defer this task to a later postscript, as checking and restoring the accuracy of my memories of these matters requires rereading my notebooks and writings of the time.

In spite of this disclaimer, I will clarify one reference in the text. The "work in progress" mentioned in the introduction and again in problem 18 refers to my unfinished and unpublished monograph *Logics of Assumptions and Inference*, which as my mathematical thinking progressed was rewritten first into the monograph *A Mathematical Basis for Psychology*, and then transformed into the substantially different widely-circulated monograph *Some Theories of Reasoned Assumptions* [1]. Some of the material in the text above was contained in versions of these earlier works, and I took up statement the mathematical view again in [2].

References

[1] Jon Doyle. Some theories of reasoned assumptions: An essay in rational psychology. Technical Report 83-125, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1983.

[2] Jon Doyle. What is rational psychology? toward a modern mental philosophy. *AI Magazine*, 4(3):50–53, 1983.

[3] M. Minsky. Problems of formulation for artificial intelligence. In *Proceedings of a Symposium on Mathematical Problems in Biology*, pages 35–46. American Mathematical Society, Providence, 1962.

Contents