15 April 1983, Volume 220, Number 4594 SCIENCE

Expert Systems Research

Richard O. Duda and Edward H. Shortliffe

Summary. Artificial intelligence, long a topic of basic computer science research, is now being applied to problems of scientific, technical, and commercial interest. Some consultation programs, although limited in versatility, have achieved levels of performance rivaling those of human experts. A collateral benefit of this work is the systematization of previously unformalized knowledge in areas such as medical diagnosis and geology.

Few areas of research have been as exciting, promising, or bewildering as artificial intelligence (AI). After 25 years of use, the very name—combining as it does a highly immodest ambition with a suggestion of deceit—still has the power to provoke controversy.

Research in AI has several goals. One is the development of computational models of intelligent behavior, including both its cognitive and perceptual aspects. A more engineering-oriented goal is the development of computer programs that can solve problems normally thought to require human intelligence.

These are ambitious aims, and neither has been achieved in any general sense. However, the research efforts have led to a substantial body of theory and techniques (*1*). In addition, during the past 10 years serious efforts have been made to apply AI to practical problems such as speech recognition, language understanding, image analysis, robotics, and consultation systems. Judged in strictly practical terms, the successes achieved to date have been modest. However, they hold great promise, and the application of AI methods to practical problems is attracting widespread interest.

This article concerns a class of AI computer programs intended to serve as consultants for decision-making. They are often called "expert systems" because they address problems normally thought to require human specialists for their solution (*2*). Some of these programs have reached expert levels of performance on the problems for which they were designed. We will describe these accomplishments as well as identify some difficult problems that must still be solved to realize their benefits in practice.

Historical Background

The goal of much of science has been to obtain quantitative descriptions of natural phenomena. Early in their training, most scientists encounter Lord Kelvin's characterization of scientific knowledge:

> When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of *science*.

Unfortunately, not all natural phenomena can be expressed well in numbers. In particular, symbolic rather than numerical operations seem to characterize cognitive activities, such as planning, problem-solving, and deduction. Serious work on AI began when it was realized that digital computers are not just fast adding machines, but are general-purpose processors of symbols, potentially capable of being programmed to exhibit such intelligent behavior (*3*). To support this view, AI researchers wrote programs to solve well-defined problems that had a distinctly nonnumerical character—programs that could play games, solve puzzles, perform symbolic integration, and even prove simple theorems in algebra, geometry, and symbolic logic. Among the important techniques that emerged were general methods for representing information in symbolic data structures, general methods for manipulating these structures, and heuristics for searching through them (*4*).

Although these results supported the theoretical possibility of machine intelligence, they fell far short of providing a basis for constructing programs that could solve complex practical problems. The early hope that a relatively small number of powerful general mechanisms would be sufficient to generate intelligent behavior gradually waned. When significant problems were addressed, it was often discovered that problem-independent heuristic methods alone were incapable of handling the sheer combinatorial complexity that was encountered. Similarly, general problem-solving techniques confronted with imprecisely stated "problems," uncertain "facts," and unreliable "axioms" were found to be inadequate to the task.

When it was asked how *people* were able to devise solutions to these problems, a frequent answer was that people possess knowledge of which the programs were wholly innocent. This knowledge is employed in a variety of ways—in clarifying the problem, suggesting the kinds of procedures to use, judging the reliability of facts, and deciding whether a solution is reasonable.

The growing recognition of the many kinds of knowledge required for high-

Richard O. Duda is a senior staff scientist at the Laboratory for Artificial Intelligence Research, Fairchild Advanced Research and Development, Palo Alto, California 94304. Edward H. Shortliffe is an assistant professor in the Division of General Internal Medicine, Stanford University School of Medicine, Stanford, California 94305.

performance reasoning systems changed the shape of AI research. In the words of Goldstein and Papert (*5*),

> Today there has been a shift in paradigm. The fundamental problem of understanding intelligence is not the identification of a few powerful techniques, but rather the question of how to represent large amounts of knowledge in a fashion that permits their effective use and interaction. . . . The current point of view is that the problem solver (whether man or machine) must know explicitly how to use its knowledge—with general techniques supplemented by domain-specific pragmatic know-how. Thus, we see AI as having shifted from a *power-based* strategy for achieving intelligence to a *knowledge-based* approach.

Expert Systems

The development of expert systems programs is one of the results of this shift to a knowledge-based approach (*6*). Paradoxically, it has proved much easier to emulate the problem-solving methods of some kinds of specialists than to write programs that approach a child's ability to perceive, to understand language, or to make "commonsense" deductions. Many human experts are distinguished by their possession of extensive knowledge about a narrow class of problems. It is this very limitation that makes it feasible to provide a computer program with enough of the knowledge needed to perform those tasks effectively.

The simplest and generally most successful expert systems are classification programs. Designed to be used in a well-defined context, their purpose is to weigh and balance evidence for a given case to decide how it should be categorized. Differential diagnosis is a classical medical example of such a problem. Although it is far from simple to do differential diagnosis well, the fact that much of an expert's knowledge concerns specific facts has made it easier to identify the necessary knowledge.

By contrast, it has proved much more difficult to develop expert systems for problems that have a more synthetic character, such as those that concern planning or require de novo generation of solutions. For example, despite some remarkable progress, no programs have yet been written that can rival an expert engineer at designing circuits or an expert mathematician at proving theorems. While no one doubts that mathematicians draw on their knowledge of mathematics in devising proofs, and while a remarkable amount of that knowledge has been identified (*7*), the nature of mathematical expertise remains elusive (*8*).

This article is concerned with the characteristics shared by a number of successful expert systems developed during the past decade. Before generalizing, however, we shall first describe a particular system in sufficient detail to ground our general observations in specific facts.

The MYCIN System

MYCIN is an expert system developed at Stanford University in the mid-1970's to aid physicians in the selection of antibiotics for patients with severe infections (*9*). In several different evaluations, MYCIN has demonstrated an ability to perform at or near the level of expert physicians (*10*). As we shall emphasize later, technical performance, narrowly defined, is not the only criterion for acceptability, and current research on descendants of MYCIN is aimed at overcoming some of its deficiencies. However, MYCIN's combination of competent performance and conceptual simplicity make it particularly convenient as an illustrative example.

In MYCIN all knowledge on infectious diseases is represented in the form of rules. The current system contains about 500 rules that deal with the diagnosis and treatment of bacteremia (bacteria in the blood) and meningitis (infection in the cerebrospinal fluid). In the program these rules are expressed in a stylized form that simplifies computer interpretation and facilitates their translation into English for human examination. The following is an example of a MYCIN rule expressed in English:

> If (i) the infection is meningitis and (ii) organisms were not seen in the stain of the culture and (iii) the type of infection may be bacterial and (iv) the patient has been seriously burned, then there is suggestive evidence that *Pseudomonas aeruginosa* is one of the organisms that might be causing the infection.

To use such general knowledge about infectious diseases, MYCIN must obtain specific knowledge about a particular patient. These patient data are stored in a dynamic database in the form of "attribute-object-value" triples. For example, the database might contain the fact that the stain (attribute) of a particular organism (object) is Gram-negative (value) or that the type of a particular infection is bacterial.

```
Please enter information about the patient.

       Name       Age        Sex        Race
1) **  Pt219      42 YEARS   FEMALE     CAUCASIAN

2) Are there any positive cultures obtained from Pt219?
** NO
3) Are there any pending cultures of Pt219?
** YES

       site       date       stain        method of
                  collected  examined?    collection
                                          [urine,sputum only]
4) **  CSF        3-Jun-75   YES          N/A

       stain of smear       morph
       [type NONE if no organism was seen on the smear]
5) **  NONE
                            .
                            .
                    {more questions}
                            .
                            .

13) Has Pt219 recently had symptoms of persistent headache or other
    abnormal neurologic symptoms (dizziness, lethargy, etc.)?
** YES
14) Has Pt219 recently had objective evidence of abnormal neurologic
    signs (nuchal rigidity, coma, seizures, etc.) documented by
    physician observation or examination?
** YES

The CSF cultures will be considered to be associated with MENINGITIS.
```

Fig. 1. Excerpt from a session with the MYCIN therapy consultation program. MYCIN assumes that the user is a physician who wants to determine the most effective combination of antibiotics to treat an infection and that the organisms causing the infection may not be known. In attempting to apply rules to solve the problem, the program asks questions to obtain the needed information. For example, at entries 4 and 5 the program is told that a culture was obtained from the cerebrospinal fluid (*CSF*), but that no organisms were seen. After asking 14 questions, the program decides that the infection is probably meningitis, and turns to the problem of identifying the likely organism or organisms.

MYCIN works in two phases, diagnosis and therapy. In its diagnosis phase, the program's main goal is to apply its rules to determine the identity of all suspicious organisms. When it attempts to apply a rule, it queries its database to see whether the needed facts are available. Thus, to apply our example rule, MYCIN would begin by accessing the database to see what is known about the infection of the patient. If the infection were known not to be meningitis, the rule would be discarded at once. However, if the infection were thought to be meningitis, the program would check the other parts of the premise in turn. If all parts were satisfied, MYCIN would apply the rule, concluding that the organism's identity might be *Pseudomonas* and thereby updating the database.

A more interesting situation arises if there is no information in the database about the patient's infection, or if what is known is too uncertain to allow any conclusions. In this case the program has two options. If this is the kind of information that the user should be able to provide, MYCIN can rely on the user's knowledge of the case and ask, "What is the <attribute> of the <object>," that is, "What is the infection of the patient?" If the user knows the answer, then that information can be added to the database, and the program proceeds to the next part of the premise. However, if MYCIN has rules allowing it to infer the answer itself from other data about the case, it sets up the new goal of determining the infection. This problem is attacked by gathering together, and attempting to apply, all rules whose conclusions refer to the infection of the patient.

Thus MYCIN's strategy in rule selection is goal-oriented, and its inference method is to "reason backward" from its initial goal. It attempts to achieve any goal by applying all the directly relevant rules. The need to establish the premises of those rules sets up new subgoals that are treated in the same way. When the program eventually requests some factual information from the user, the rule that prompted the request may become applicable and, if so, is applied. The application of a rule enters a new fact into the database. This in turn is available when attempting to apply other rules. Thus the line of questioning, the rules that are applied, and the conclusions that are reached are determined by the data obtained for the particular patient.

The execution of this strategy leads to interactive consultation with the user. Portions of a sample consultation session with MYCIN are shown in Figs. 1 to 3. With the exception of a few routine initial questions, each question asked by the system is the consequence of its attempt to apply some particular rule. The responses of the user, who is presumed to be a physician, are printed in capital letters and follow a double asterisk. In most cases the user provides single-word answers, including UNKNOWN when no information is available. However, the user can also respond with one of a number of commands. Of these, the WHY command is particularly important, distinguishing MYCIN from alternative decision-tree or statistically based programs that might have produced a superficially similar line of questioning. Thus, in Fig. 3, when the user asked MYCIN why question 38 was being asked, the program could provide both the goal of the question and the relevant rule being pursued. By repeatedly invoking the WHY command the user can systematically trace back through the program's chain of reasoning.

This ability to provide understandable explanations is one of the greatest advantages of an AI approach to diagnosis. Lack of this ability is a major reason why physicians have failed to embrace statistically based diagnosis programs, even when excellent performance has been demonstrated in clinical trials (*11*). MYCIN's explanation facilities expose the program's line of reasoning in a way that a human observer can understand and critique. Furthermore, they contribute to the potential use of systems like MYCIN for computer-aided instruction (*12*).

My therapy recommendations will be designed to treat for organisms that either are very likely or, although less likely, would have a significant effect on therapy selection if they were present. It is important to cover for the following probable infection(s) and associated organism(s):

```
INFECTION-1 is MENINGITIS

  + <ITEM-1>  MYCOBACTERIUM-TB [from clinical evidence only]
  +    ...    COCCIDIOIDES     [from clinical evidence only]
  +    ...    CRYPTOCOCCUS     [from clinical evidence only]

  43) Does Pt219 have a clinically significant allergic reaction to any
      antimicrobial agent?
  ** NO
               [Considering ETHAMBUTAL-AND-INH-AND-RIFAMPIN
                for use against INFECTION-1...]
  44) Is Pt219 pregnant or breast feeding?
  ** NO
  45) Do you have reason to suspect that Pt219 may have impaired renal
      function?
  ** NO
  46) Pt219's weight in kilograms (or <number> POUNDS):
  ** 51.3

[REC-1] My preferred therapy recommendation is as follows:
        Give the following in combination:
        1) ETHAMBUTAL
        Dose: 1.28g (13.0 100mg-tablets) q24h PO for 60 days
              [calculated on basis of 25 mg/kg]
              then 770 mg (7.5 100mg-tablets) q24h PO
              [calculated on basis of 15 mg/kg]
           [The dose should be modified in renal failure.
           I can adjust the dose only if a creatinine clearance
           or a serum creatinine is obtained.]
        Comments: Periodic vision screening tests are recommended
                  for optic neuritis.

        2) INH
        Dose: 513 mg (5.0 100mg-tablets) q24h PO
              [calculated on basis of 10 mg/kg]

        3) RIFAMPIN
        Dose: 600 mg PO q24h
        Comments: Administer dose on an empty stomach

Although I suspect that fungus may be a possible pathogen, in the
absence of definitive evidence, I will not recommend antifungal
therapy at this time.  Please obtain samples for fungal, TB, and
viral cultures, cytology, VDRL (blood and CSF), coccidioides
complement-fixation (blood and CSF), cryptococcal antigen
(blood and CSF), viral titers (blood and CSF).  An infectious
disease consult may be advisable.
```

Fig. 2. MYCIN's therapy advice. After completing the diagnosis phase, MYCIN determines a combination of antibiotics to be used against the suspected organisms. The program also suggests alternative therapies and allows the user to enter a query mode to probe the reasons for these conclusions.

Accomplishments to Date

MYCIN is only one of many expert systems built during the past decade. Most of these programs have resulted from applied research. The general ideas behind their methods for knowledge representation and inference were well known within the AI community; the challenge was to show that the techniques could be effectively applied to problems of scientific or economic interest.

Expert systems have typically been developed by running the evolving program on test cases, noting problems, and refining the knowledge base or problem-solving strategy accordingly. In many cases the more mature programs have undergone a systematic evaluation to assess their performance relative to some accepted criterion, such as agreement with the decisions of human experts. These evaluations are complicated by the many dimensions of performance and by disagreement among the experts as to what constitutes correct behavior (*13*).

Some representative expert systems are listed in Table 1 (*10, 14–22*). Each program mentioned either has undergone a semiformal evaluation or is in routine use. Although the list is by no means exhaustive, it reflects the state of the art and the ways in which validation experiments must often be adapted to the problem domain for which the expert system was developed. To illustrate these issues, we will briefly describe some of the better known systems and their validation.

DENDRAL. One of the earliest expert systems, DENDRAL analyzes mass spectral patterns to suggest the chemical structure of unknown compounds (*14*). Its use had led to approximately 50 publications in the chemistry literature, and it has been validated by running analyses on several families of compounds: aliphatic structures (ketones, ethers, alcohols, and amines), generalized aliphatic monofunctional compounds, cyclic ketones, estrogenic steroids, and prostaglandins. The validation methodology has involved developing analytical rules on a small set of compounds from the group of interest, testing and refining on several more members, and then running the system on a total of 25 to 40 compounds in the same class. The system's contributions to refereed journal publications, coupled with its acceptance and routine use by chemists, have been viewed as an effective validation of DENDRAL's performance.

MYCIN. The meningitis evaluation for MYCIN involved a more formal study design in which the expert evaluators did not know whether they were assessing the advice of an infectious disease consultant, a medical resident, a medical student, or the program. In that study, MYCIN's recommendations were uniformly judged preferable, or equivalent, to those of five infectious disease experts who recommended therapy for the same patients (*10*). MYCIN tends to do least well when serious infections are present at sites about which the program has no knowledge. One reason the system has not been implemented clinically, therefore, is the incompleteness of its knowledge on infectious diseases.

INTERNIST-1. INTERNIST-1 is an ambitious program designed to undertake diagnosis for all problems in internal medicine (*15*). The knowledge base was developed and refined by running the program on difficult cases taken from medical journals. In a recent evaluation, new cases representing 43 different diagnostic problems were selected from the *New England Journal of Medicine*. INTERNIST-1 made the correct diagnosis 25 times, compared with 28 times for the physicians caring for the patients and 35 times for the expert clinicians who discussed the cases in the journal. Although the program did miss several diagnoses, no previous diagnostic system could have adequately coped with such a large number of complex cases drawn from the broad field of internal medicine. An analysis of those cases on which INTER-

```
34) Is the meningitis a hospital-acquired infection?
** NO
35) Is Pt219's clinical history consistent with EPIGLOTTITIS?
** NO
36) Is Pt219's clinical history consistent with OTITIS-MEDIA?
** NO
37) Has the patient ever been to an area that is endemic for
    coccidiomycoses?
** UNKNOWN
38) Is Pt219 a burn patient?
** WHY

[i.e. WHY is it important to determine whether PATIENT-219 has been
seriously burned?]

[4.0] This will aid in determining the organisms (other than those
      seen on cultures or smears) which might be causing the infection.
  It has already been established that
     [4.1] the infection which requires therapy is meningitis, and
     [4.2] organisms were not seen on the stain of the pending csf
           culture (CULTURE-1), and
     [4.3] the type of the infection may be bacterial
  Therefore, if
     [4.4] Pt-219 has been seriously burned
  then
     there is suggestive evidence that Pseudomonas-aeruginosa is one
     of the organisms (other than those seen on cultures or smears)
     which might be causing the infection
  Author: Yu
  [RULE578]

   [back to question 38...]
   ** HOW 4.3

[i.e. HOW was it established that the type of the infection may be
bacterial?]

The following rules concluded about: the likelihood that the type
of the infection may be bacterial

                 Rule          Yes         No

     [4.5]     RULE500       (.4)
     [4.6]     RULE501                   (.17)
     [4.7]     RULE502       (.1)
     [4.8]     RULE524       (.42)
     [4.9]     RULE504       (.28)

Thus, it has been established that the type of the infection may be
bacterial.
```

Fig. 3. Explanation features. This excerpt from the MYCIN run illustrates some of the explanation facilities provided by the program. At entry 38, MYCIN asks whether the patient has been seriously burned. Instead of answering, the user asks why this particular question is being asked. In response, the program states both the goal of the question and the rule it was attempting to apply. This explanation leads the user to wonder how MYCIN established that the infection might be bacterial; in response to the HOW command, the program lists the five rules that it applied, four of which supported this conclusion. This ability to inspect the program's reasoning methods provides the user with a firmer basis for understanding the final conclusions.

Table 1. Representative expert systems.

Expert system	Domain	Type of evaluation	Routine use	Reference
DENDRAL	Mass spectroscopy interpretation	Case studies	Yes	*(14)*
MYCIN	Antimicrobial therapy	Randomized trials	No	*(10)*
INTERNIST-1	Internal medicine diagnosis	Case studies	No	*(15)*
CASNET	Glaucoma assessment and therapy	Case studies	No	*(16)*
PROSPECTOR	Geological exploration	Case studies	No	*(17)*
R1	Computer layout and configuration	Case studies	Yes	*(18)*
Digitalis Advisor	Digitalis dosing advice	Randomized trials	No	*(19)*
PUFF	Pulmonary function test interpretation	Randomized trials	Yes	*(20)*
Microprocessor EXPERT	Protein electrophoresis interpretation	Case studies	Yes	*(21)*
HASP and SIAP	Ocean surveillance (signal processing)	Case studies	No	*(22)*

NIST-1 failed to perform as well as the discussants has helped identify specific deficiencies to be overcome in subsequent versions of the system: the program's inability to reason by using anatomic knowledge or knowledge of the time course of disease, its occasional attribution of clinical findings to improper causes, and its inability to explain the basis for its decisions.

PROSPECTOR. PROSPECTOR is a mineral exploration consultation system designed for problems in regional resource evaluation, ore deposit identification, and drilling site selection *(17)*. Its knowledge base is organized around models of different types of ore deposits, including kuroko-type massive sulfide, Mississippi Valley lead and zinc, komatiitic nickel sulfide, Yerington porphyry copper, Butte porphyry copper, island-arc porphyry copper, hood porphyry molybdenum, zoned vertical-cylinder porphyry molybdenum, roll-front sandstone uranium, and Grants sandstone uranium. As with MYCIN, PROSPECTOR's coverage is incomplete, and much work remains to include all deposit types of economic significance.

In PROSPECTOR's domain, the usual problems of validation are further complicated by the relatively small number of well-known ore deposits of any given type and by the long time between the initial discovery of a deposit and its final characterization. In formal tests with data on known deposits, PROSPECTOR's assessments have repeatedly agreed closely with those of the geological consultants who provided the models. In addition, in the one test involving a prospect undergoing exploration, the program accurately identified the location and extent of ore-grade mineralization for a previously unknown portion of a porphyry molybdenum deposit. While this was not a formal statistical study, its success justifies extending PROSPECTOR's knowledge base.

R1. Also known as XCON, R1 is a rule-based expert system that configures VAX computers, determining the physical layout and interconnection of their many components *(18)*. The program both adds support components missing from the order and saves engineering time, providing technicians who assemble the systems with information that is much more detailed than the traditional hand-generated specifications. Developed at Carnegie Mellon University in the late 1970's, R1 is now used by the Digital Equipment Corporation to configure every VAX that is sold. Of the more than 3000 orders that were processed in one 3-month period, over 85 percent of the configurations were flawless, and most of the rest were usable with minor corrections. Many of the errors occurred merely because R1 lacked information on recently introduced products, and most of the rest were due to known, correctable problems with the rules. Although a formal validation procedure was performed before a decision was made to put R1 into production operation, the acceptance of R1 in practice is the most convincing demonstration of its usefulness.

The programs mentioned here, and the others listed in Table 1, serve to illustrate the current status of expert systems research. Without exception, the successes that have been obtained were due to extensive effort devoted to formalizing and organizing a large amount of knowledge. This knowledge is neither a large database of unstructured facts nor a small set of formal axioms for a general theory. Rather, it is typically a substantial collection of semiorganized, perhaps incomplete, and often subjective, information. Encoding this kind of subjective information in a computer program serves to make it, if not objective, at least explicit and public. If the knowledge is valuable and is faithfully represented, the resulting program can make it more widely available and permit it to be more uniformly applied as an aid to decision-making. Indeed, one of the most important results of this enterprise may be the development of ways to express formally, and to record systematically, knowledge that is usually unexpressed and unrecorded.

Research Issues

Despite the impressive decision-making performance that has been achieved, only four of the systems listed in Table 1 are in routine use, and two of them (PUFF and the electrophoresis analyzer) are rather small and simple by AI standards. Thus, it is important to ask what is impeding the greater use of expert systems.

There are several answers to this question. One is that the work is relatively new. Most of the expert system projects have been aimed at demonstrating the possibility of applying AI methods to significant problems. Such a demonstration is merely the first step toward practical operation, after which other considerations, such as cost, speed, reliability, versatility, convenience, and user acceptance, become dominant.

However, there are also some more fundamental problems that are not only holding back the exploitation of these particular programs, but are also limiting the potential for future applications. In this section we consider some of these more basic research issues.

Knowledge acquisition. One common characteristic of MYCIN, INTERNIST, and PROSPECTOR is that their knowledge bases are incomplete. The identification and encoding of knowledge is one of the most complex and arduous tasks encountered in the construction of an expert system. The very attempt to build a knowledge base often discloses gaps in our understanding of the subject domain and weaknesses in available representation techniques. Even when an adequate knowledge representation formalism has been developed, experts often have difficulty expressing their knowledge in that form. Thus the process of building a knowledge base has usually required a time-consuming collaboration between a domain expert and an AI researcher. While an experienced team can put together a small prototype system in 1 or 2 man-months, the effort required to produce a system that is ready for serious evaluation (well before contemplation of actual use) is more often measured in man-years.

It has frequently been suggested that some kind of learning process might solve this problem. A related idea is to provide the expert with an appropriate way to "teach" the system directly (*23*). While both of these ideas are plausible, programs that can learn or be taught seem to need a significant amount of initial knowledge, together with mechanisms for assimilating that knowledge properly. Although this is an excellent area for future research, learning techniques currently cannot solve the problems facing the builder of expert systems.

Knowledge representation. MYCIN's knowledge about bacterial infections and R1's knowledge about computer configuration are represented by rules. The advantages and limitations of such simple, uniform approach are well appreciated by AI researchers, who have developed a variety of alternative formalisms for knowledge representation (*24*). The methods currently in use in expert systems rarely capture subtleties and sometimes fail to reflect major aspects of an expert's knowledge. For example, while MYCIN and INTERNIST-1 have effective mechanisms for representing empirical associations, neither has appropriate ways to express physiological mechanisms or temporal trends in the evolution of disease processes.

Ideally, a knowledge representation formalism should (i) represent the concepts and intentions of the expert faithfully, (ii) be able to be interpreted by the program correctly and effectively, (iii) support explanations that convey a line of reasoning that the human observer can understand and critique, (iv) facilitate the process of finding gaps and errors in the knowledge base, and (v) allow separation of domain knowledge from the interpretation program so that the knowledge base can be enlarged or corrected without the need for reprogramming the interpreter. These criteria place conflicting demands on the system designer. The first two (fidelity and effectiveness) lead toward complex representations specific to each situation, whereas the other three favor a single, uniform formalism that is simple to interpret.

While the choice of uniform representations has allowed the construction of large systems, current research is showing a trend toward more complex and heterogeneous approaches. It has frequently been noted that humans seem to exploit several different representations of the same phenomena. In particular, experts seem to employ rule-like associations to solve routine problems quickly, but can shift to using more reasoned arguments based on first principles when the need arises (*25*).

Consider, for example, the rule in MYCIN that tells the program to avoid giving tetracycline to patients under 8 years of age. This rule effectively prevents the system from recommending therapy with tetracycline in this pediatric age group, but the underlying reason for this rule (that the drug can produce dental staining during tooth development and may depress bone growth) is not represented. This is a level of detail that the developers of MYCIN chose not to include. As a result, the system cannot deal with exceptions (cases in which the severity of disease requires administration of tetracycline despite the potential cosmetic side effects) and cannot explain the basis for the rule. As this example illustrates, augmenting empirical knowledge with causal or mechanistic links that represent functional behavior promises a significant increase in capabilities (*26*).

Inference and uncertainty. MYCIN is a goal-driven inference system that reasons backward from goals to data. Other inference strategies have been used in other domains. For example, R1 uses a so-called data-driven strategy in which the user initially enters all of the information about the problem into the dynamic database and the rules are then applied to "reason forward" from the data to the conclusions. More generally, one would like to use knowledge about the problem to decide the best strategy to pursue.

When inference steps are less than certain, a new level of complexity is introduced. Most expert systems that can tolerate uncertainty employ some kind of probability-like measure to weigh and balance conflicting evidence. PROSPECTOR assigns probabilities to conclusions, using an approximate form of Bayes's rule to update these probabilities as information is obtained; however, this leads to problems with assumptions about statistical independence and prior probabilities. MYCIN avoids these problems by employing a novel calculus of certainty values, but the operational meaning of the computed numbers is not always clear. Possibility theory (*27*) and the Dempster/Shafer theory of evidence (*28*) have been advocated as formal methods for dealing with the problems of vagueness and ignorance. However, questions about how a program should reason in the presence of ignorance, or how it can even recognize the limits of its knowledge, are largely unanswered.

Explanation. One of the most important features of MYCIN is its ability to provide explanations of the program's behavior. MYCIN's explanations are given in terms of its goals and its rules, and can be very illuminating. However, when asked the same question, the expert who provided those rules might give a very different explanation in terms of physiological mechanisms or disease processes. The desire to provide such causal explanations is another motivation for employing multiple levels of representation.

Another characteristic of effective human consultants is that their explanations are adjusted to satisfy the perceived needs of questioners. For a program to respond similarly it must maintain a model of the user, an assessment of what the user does and does not know and what he is trying to accomplish (*29*). However, models of users will have to be more sophisticated before they solve more problems than they cause.

Ultimate limitations. Expert systems are frequently presented as surrogate consultants, programs one can turn to for advice when the need arises, just as one would turn to a human consultant. While this is a reasonable metaphor, if taken too literally it leads to the conclusion that success will not be achieved until all the problems of AI have been solved—that the program must not only be able to reason at the level of expert humans but must also be able to converse in idiomatic natural language, perceive evidence directly, and possess that breadth of knowledge that is called common sense.

This is a needlessly pessimistic conclusion. The goal of expert systems research is to provide tools that exploit new ways to encode and use knowledge to solve problems, not to duplicate intelligent human behavior in all its aspects. The challenge at this stage of expert systems development is therefore to constrain the problems addressed in realistic ways to allow useful solutions to real-world problems.

Commentary

The field of expert systems is one of the most active and exciting areas of applied research in AI. As we have outlined, the work of the past decade has shown that programs that can operate at or near the level of human experts are feasible; several have been demonstrated to be capable of such performance in carefully selected, well-specified domains. As a result, the field is beginning to undergo the transition from basic research to application.

Current technology seems best suited to diagnosis or classification problems whose solutions depend primarily on the possession of a large amount of specialized factual and empirical knowledge. However, progress has also been made on synthetic problems such as planning and design. Successes in these areas not only point to the potential of the field but also help define the most important topics for ongoing basic research. The limitations of current expert systems have exposed unsolved problems in such basic areas as knowledge representation, inference, perception, and learning. Progress in solving these fundamental problems will lead to significant advances in the capabilities of expert systems.

The success of expert systems research is disclosing additional new problems, many of which are sociological. In domains such as medicine, important legal and ethical questions remain to be resolved. Commercial and industrial interest, stimulated by perhaps unrealistic expectations about the power of currently understood AI techniques, has created a shortage of appropriately trained and motivated professionals. The standard problems of transforming a concept into a commerical product are further complicated by the lack of any tradition in producing applications of AI research.

The greatest contributions of expert systems research may well go beyond the development of high-performance programs. Equally as important is the field's impact on the systematization and codification of knowledge previously thought unsuited for formal organization. Improved approaches to formalizing and managing knowledge are certain to be of importance to a variety of scientific and economic endeavors.

References and Notes

1. The theoretical foundations of AI are clearly presented by N. J. Nilsson [*Principles of Artificial Intelligence* (Tioga, Palo Alto, Calif., 1980)]. A good introduction to the practice of AI as a programming activity is presented by P. H. Winston [*Artificial Intelligence* (Addison-Wesley, Reading, Mass., 1977)]; more advanced techniques are described by E. Charniak, C. K. Riesbeck, and D. V. McDermott [*Artificial Intelligence Programming* (Erlbaum, Hillsdale, N.J., 1980)]. Other important AI references are by A. Newell and H. A. Simon [*Human Problem Solving* (Prentice-Hall, Englewood-Cliffs, N.J., 1972)] and A. Barr, P. R. Cohen, and E. A. Feigenbaum, Eds. [*Handbook of Artificial Intelligence* (Kaufmann, Los Altos, Calif., 1981 and 1982), vols. 1 to 3]. The last three chapters of the book by M. Boden [*Artificial Intelligence and Natural Man* (Basic Books, New York, 1977)] discuss some of the psychological, philosophical, and social issues raised by AI research, including the concerns voiced by J. Weizenbaum [*Computer Power and Human Reason* (Freeman, San Francisco, 1976)].
2. There is as yet no systematic textbook on expert systems, although the general principles are described in F. Hayes-Roth, D. Waterman, and D. Lenat, Eds. [*Building Expert Systems* (Addison-Wesley, New York, 1983)]. Two useful collections of papers are in D. Michie, Ed. [*Expert Systems in the Microelectronic Age* (Edinburgh Univ. Press, Edinburgh, 1979)] and B. L. Webber and N. J. Nilsson, Eds. [*Readings in Artificial Intelligence* (Tioga, Palo Alto, Calif., 1981)]. Several of the best-known expert systems for medical decision-making are described in P. Szolovits, Ed. [*Artificial Intelligence in Medicine* (Westview, Boulder, Colo., 1982)]; for a general survey of the field, see the papers by B. G. Buchanan [in *Machine Intelligence*, J. E. Hayes, D. Michie, Y.-H. Pao, Eds. (Wiley, New York, 1982), vol. 10, pp. 269–300] and D. S. Nau [*Computer* **16**, 63 (1983)].
3. E. A. Feigenbaum and J. Feldman, *Computers and Thought* (McGraw-Hill, New York, 1963). The view of intelligence as a symbol-processing activity is cogently articulated by A. Newell and H. A. Simon [*Commun. ACM* **19**, 113 (1975)].
4. N. J. Nilsson, *Problem-Solving Methods in Artificial Intelligence* (McGraw-Hill, New York, 1971).
5. I. Goldstein and S. Papert, *Cogn. Sci.* **1**, 84 (1977).
6. The phrase "knowledge-based systems" is often preferred to "expert systems," since there are no uniquely qualified human experts for a large number of AI applications; however, both phrases are sufficiently vague that the latter can be applied to almost any program that works well and the former to almost any program at all. While usage is far from uniform, we shall define a knowledge-based system as an AI program whose performance depends more on the explicit presence of a large body of knowledge than on the possession of ingenious computational procedures; by expert system we mean a knowledge-based system whose performance is intended to rival that of human experts.
7. The role of knowledge in theorem proving was described by W. W. Bledsoe [*Artif. Intell.* **9**, 1 (1977)]. Bledsoe observed, "The word 'knowledge' is a key to much of this modern theorem proving. Somehow we want to see the knowledge accumulated by humans over the last few thousand years, to help direct the *search* for proofs."
8. Observers of AI research frequently note that the hardest part of many problems is converting them from a vague initial statement to a form that is sufficiently precise to allow formal problem-solving to begin [H. A. Simon, *Artif. Intell.* **4**, 181 (1973); H. E. Pople, Jr., in *Artificial Intelligence in Medicine*, P. Szolovits, Ed. (Westview, Boulder, Colo., 1982), pp. 119–190; M. Stefik and L. Conway, *AI Mag.* **3**, 4 (1982)].
9. E. H. Shortliffe, *Computer-Based Medical Consultations: MYCIN* (Elsevier/North-Holland, New York, 1976).
10. MYCIN has been favorably evaluated for its ability to handle isolated bacteremias [V. L. Yu, B. G. Buchanan, E. H. Shortliffe, S. M. Wraith, R. Davis, A. C. Scott, F. M. Cohen, *Comput. Programs Biomed.* **9**, 95 (1979)] and meningitis [V. L. Yu *et al.*, *J. Am. Med. Assoc.* **242**, 1279 (1979)].
11. E. H. Shortliffe, B. G. Buchanan, E. A. Feigenbaum, *Proc. IEEE* **67**, 1207 (1979). In a recent survey, physicians overwhelmingly cited explanation capabilities as a principal requirement for clinically acceptable computer-based consultation systems [R. L. Teach and E. H. Shortliffe, *Comput. Biomed. Res.* **14**, 542 (1981)].
12. MYCIN's explanation facilities were refined and expanded in the TEIRESIAS program [R. Davis and D. B. Lenat, *Knowledge Systems in Artificial Intelligence* (McGraw-Hill, New York, 1982)] and further adapted for the education of medical students in GUIDON [W. J. Clancey, *Int. J. Man-Mach. Stud.* **11**, 25 (1979)].
13. J. Gaschnig, P. Klahr, H. E. Pople, Jr., A. Terry, E. H. Shortliffe, in *Building Expert Systems*, F. Hayes-Roth, D. Waterman, D. Lenat, Eds. (Addison-Wesley, New York, in press).
14. R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project* (McGraw-Hill, New York, 1980).
15. The design and evaluation of INTERNIST-1 were recently described by R. Miller, H. Pople, Jr., and J. Myers [*N. Engl. J. Med.* **307**, 468 (1982)], and the limitations of the system have been outlined to motivate the design of INTERNIST's successor, CADUCEUS [H. Pople, Jr., in (*8*)].
16. CASNET is an expert system that uses a causal model to provide consultations for patients with glaucoma [S. M. Weiss, C. A. Kulikowski, S. Amarel, A. Safir, *Artif. Intell.* **11**, 145 (1978)]. It was successfully evaluated at a symposium on glaucoma sponsored by the National Society for the Prevention of Blindness, where its performance was judged comparable to that of a distinguished panel of glaucoma experts [P. R. Lichter and D. R. Anderson, *Discussions on Glaucoma* (Grune & Stratton, New York, 1977)].
17. PROSPECTOR is based on a network reasoning structure that incorporates the knowledge of the geologists who collaborated on the project [R. Duda, J. Gaschnig, P. Hart, in *Expert Systems in the Microelectronic Age*, D. Michie, Ed. (Edinburgh Univ. Press, Edinburgh, 1979), pp. 153–167]. The results of a validation study are described by J. Gaschnig [in *Machine Intelligence*, J. E. Hayes, D. Michie, Y.-H. Pao, Eds. (Wiley, New York, 1982), vol. 10, pp. 301–323)], and the successful prediction of a deposit is reported by A. N. Campbell, V. F. Hollister, R. O. Duda, and P. E. Hart [*Science* **217**, 927 (1982)].
18. J. McDermott, *Artif. Intell.* **19**, 29 (1982). The evaluation data cited in the text are from J. McDermott (personal communication).
19. The Digitalis Advisor is a consultation system that uses a mathematical pharmacokinetic model of digitalis distribution and excretion, coupled with a patient-specific model that uses AI representation techniques [G. A. Gorry, H. Silverman, S. G. Pauker, *Am. J. Med.* **64**, 452 (1978)]. In an evaluation of the system's performance in recommending digitalis dosing for a group of 50 patients at a large hospital, the computer gave advice that was rated of comparable quality to that of the attending physician (W. Long, P. Szolovits, S. G. Pauker, personal communication).
20. PUFF differs from the other expert systems discussed in that it was originally developed and tested by using EMYCIN, a system-building tool based on the routines originally developed for MYCIN [W. van Melle, "A domain-independent system that aids in constructing knowledge-based consultation programs" (Rep. HPP-80-11, Computer Science Department, Stanford University, Stanford, Calif., June 1980)]. Once its EMYCIN-based performance was stable, PUFF was rewritten in BASIC to run on a minicomputer. The resulting system is used routinely to analyze pulmonary function tests and to print clinical reports at Pacific Medical Center in San Francisco. PUFF agrees with clinical experts interpreting the same tests in over 90 percent of the cases [J. S. Aikins, J. C. Kunz, E. H. Shortliffe, R. J. Fallat, "PUFF: An expert system for interpretation of pulmonary function data" (Rep. HPP-82-13, Computer Science Department, Stanford University, Stanford, Calif., August 1982)]
21. Computer scientists, working in collaboration with a pathologist and a commercial clinical instrumentation firm, developed a small expert system for interpreting serum protein electrophoresis tracings [S. M. Weiss, C. A., Kulikowski, R. S. Galen, in *Proceedings of the Seventh International Joint Conference on Artificial Intelligence* (University of British Columbia, Vancouver, August 1981), p. 835]. The resulting system, after validation on large numbers of test cases, was then implemented on a microprocessor chip and is incorporated in protein electrophoresis instruments sold by the commercial firm.
22. HASP (and its successor SIAP) are knowledge-based systems that use information about vessels and the sea and expertise about signal interpretation to analyze signals from ocean sensors. The system performed well in an independent evaluation by the Mitre Corporation [H. P. Nii, E. A. Feigenbaum, J. J. Anton, A. J. Rockmore, *AI Mag.* **3**, 23 (1982)].
23. Learning has long fascinated AI researchers, but the problem of computer induction is a difficult one and progress has been slow. The uniformity of rule-based programs simplifies the development of programs that can help the user build their knowledge bases. Such programs include TEIRESIAS for EMYCIN systems and SEEK for a system-building program called EXPERT [P. Politakis and S. M. Weiss, paper presented at the 15th Hawaii International Conference on Systems Sciences, Honolulu (1982)].
24. Knowledge representation is a core topic for AI

research. Several general approaches have been developed, including logic, rules, semantic networks, and frames. The flexibility and precision of mathematical logic make it both a useful method and a standard of comparison for alternative representation schemes. Rules provide a modular and uniform mechanism that has proved to be popular for both expert systems and psychological modeling. Semantic network representations simplify certain deductions (such as inferences through taxonomic relations) by reflecting them directly in a network structure. Frames generalize this notion, providing structures or frameworks for organizing knowledge. A good overview of these topics and further references are given in A. Barr and E. A. Feigenbaum, Eds. [*Handbook of Artificial Intelligence* (Kaufman, Los Altos, Calif., 1981), vol. 1].
25. Systematic comparisons of the behavior of experts and novices are reported by M. T. H. Chi, P. J. Feltovich, and R. Glaser [*Cogn. Sci.* **5**, 121 (1982)] and by J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon [*Science* **208**, 1335 (1980)].
26. P. E. Hart, *SIGART Newslett. No. 79* (January 1982), p. 11. The need for mechanistic or support knowledge when teaching students the MYCIN knowledge base is discussed by W. J. Clancey (*Artif. Intell.*, in press). ABEL, an experimental system that suggests how multiple levels of representation may be used, has been developed by R. S. Patil, P. Szolovits, and W. B. Schwartz [in *Proceedings of the Seventh International Joint Conference on Artificial Intelligence* (University of British Columbia, Vancouver, 1981), p. 893].
27. L. A. Zadeh, *Fuzzy Sets Syst.* **1**, 3 (1978).
28. G. Shafer, *A Mathematical Theory of Evidence* (Princeton Univ. Press, Princeton, N.J., 1976).
29. User models are particularly important for instructional use of expert systems. Examples of user modeling include the MACSYMA Advisor [M. R. Genesereth, paper presented at the Sixth International Joint Conference on Artificial Intelligence, Tokyo (1979)] and GUIDON [see W. J. Clancey in (*12*)].
30. We thank S. Amarel, D. R. Barstow, R. J. Brachman, B. G. Buchanan, W. J. Clancey, R. Davis, J. Early, O. Firschein, M. A. Fischler, P. E. Hart, C. Kulikowski, R. Maffly, J. McDermott, N. J. Nilsson, M. Stefik, and P. Szolovits for their helpful suggestions. E.H.S. is supported by the Office of Naval Research under contract NR 049-479 and by National Library of Medicine research career development award LM-00048.

Neural Crest and the Origin of Vertebrates: A New Head

Carl Gans and R. Glenn Northcutt

Summary. Most of the morphological and functional differences between vertebrates and other chordates occur in the head and are derived embryologically from muscularized hypomere, neural crest, and epidermal (neurogenic) placodes. In the head, the neural crest functions as mesoderm and forms connective, skeletal, and muscular tissue. Both the neural crest and the epidermal placodes form special sense organs and other neural structures. These structures may be homologous to portions of the epidermal nerve plexus of protochordates. The transition to vertebrates apparently was associated with a shift from a passive to an active mode of predation, so that many of the features occurring only in vertebrates became concentrated in the head.

In the extensive discussions on the origin of vertebrates (*1–3*), there has been an emphasis on characters shared between early vertebrates (and their descendants), and other deuterostomes—particularly the protochordates (Table 1). This emphasis on similarities, and thus on characters that appear to be primitive for chordates, has masked some major differences between vertebrates and all other deuterostomes. Consideration of these disparities allows analyses of the functional shifts that seem to have occurred with the origin of vertebrates.

Our analysis of new data from developmental biology, neurobiology, functional morphology, and systematics leads us to propose the hypothesis that vertebrates have evolved from protochordate-like ancestors primarily by elaboration and differentiation of their epidermal nerve plexus and by muscularization of their hypomere. This hypothesis is supported by the observation that many of the sensory, integrative, and motor systems of vertebrates, as well as their supportive skeletal structures, are derived embryologically from neural crest and epidermal (neurogenic) placodes. In the process, these embryonic tissues form the anterior part of the head, most of which represents a new vertebrate unit.

The structural differences between protochordates and vertebrates are presented in Table 2, along with notes on the embryonic origins of the vertebrate structures. Consideration of the functions of these systems and of their phylogenetic development leads to a new interpretation of the phylogeny of the vertebrate head.

The authors are professors of zoology in the Division of Biological Sciences, University of Michigan, Ann Arbor 48109.

Comparison of Vertebrates and Protochordates

At some stage of their life history, all chordates show such apparently derived characters as a dorsal hollow nerve cord, a notochord, segmented muscles (in an unsegmented trunk), and a perforated pharynx (*1–3*). However, certain assumed correlates of these characters prove to be only superficially similar in different chordate types (Table 3). For instance, although protochordates and vertebrates both have a pharynx, that of vertebrates differs (i) in having a cartilaginous rather than a collagenous skeleton (*4, 5, 8*), (ii) in pumping water with the branchiomeric muscles rather than cilia (*1*), and (iii) in having gills and internal, muscular aortic arches, both of which are lacking in protochordates (*6–8*). Similarly, although the myotomes of the axial musculature are staggered in cephalochordates, those of vertebrates lie in symmetrical pairs. In addition, the trunk muscle cells of cephalochordates extend to the nerve cord, where the motor endplates lie. In contrast, the motor endplates of vertebrates lie at the termination of peripherally passing spinal nerves (*9*). The differences between protochordates and vertebrates even affect superficially similar structures.

Vertebrates differ from other chordates because they are mobile predators, the predatory activities of which, whether or not utilizing jaws, inevitably involve the modified skeletal elements and muscles of the pharynx. This active predation is directed by an elaborate array of special sense organs and their integrating circuitry. Predation is supported by an advanced metabolic mechanism with specializations for exchange and