4

Artificial Intelligence Methods and Systems for Medical Consultation

Casimir A. Kulikowski

Shortly after the preceding review article appeared, Kulikowski published the following more detailed analysis of the knowledge-engineering approach. Focusing mostly on the medical AI systems of the 1970s, he considers the major problems that arise in designing a consultation program. These problems center about choosing diagnostic interpretation and treatment-planning strategies and the knowledge representations for formalizing them. In choosing a knowledge representation, Kulikowski notes that explanation and knowledge acquisition are just as important as efficient and effective performance (Shortliffe, 1982b). Indeed, these concerns are interrelated: justifying decisions and updating the knowledge base, as the system is built incrementally or new information becomes available, place a premium on the modularity of a representation and the ease with which its reasoning procedures can be explained.

In both diagnosis and treatment decisions, schemes for quantifying the uncertainty of inferences raise difficult issues of both an empirical and a formal logical nature (see also Chapter 9). In addition, many specific practical problems of system design arise. Achieving robust performance despite uncertain relationships is a crucial requirement; an important insight resulting from the design of several systems is that robust performance can largely be achieved by a rich network of deterministic relationships that interweave the space of hypotheses.

Kulikowski also discusses several knowledge-based representational schemes that generalize the results of the early consultation programs

 From *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2: 464–476 (1980).

[EMYCIN (van Melle et al., 1981), EXPERT (Weiss and Kulikowski, 1979), AGE (Nii and Aiello, 1979)]. By providing an environment for encoding knowledge, editing the evolving knowledge base, and testing programs, these systems provide techniques and tools that promise to be very versatile in helping to design new medical expert systems.

While the earlier chapters in this volume provide motivation for applying artificial intelligence techniques to medicine, comparing the methods to those of traditional algorithmic programming and statistics, in this paper Kulikowski presents the knowledge-based perspective as a whole. This serves as a prelude to detailed discussions of particular consultation systems (Chapters 5, 6, 7, and 8) and to Szolovits and Pauker's analysis of medical reasoning in the context of these programs (Chapter 9).

4.1 Introduction

4.1.1 The Need for Computer-Based Medical Consultation

Expert medical consultation is a scarce, expensive, yet critical component of any health care system. Making the knowledge and expertise of human experts more widely available through computer consultation systems has been recognized as an important mechanism for improving the access to high-quality health care (Schoolman and Bernstein, 1978; Schwartz, 1970). The simulation of clinical cognition by the computer raises important scientific questions about the structure, consistency, completeness, and uncertainty of medical knowledge. These considerations are of particular interest to researchers in artificial intelligence (Minsky, 1968; Newell and Simon, 1972; Nilsson, 1980), cognitive psychology (Elstein, 1976), and medical science and education (Feinstein, 1967; Komaroff, 1979; Schoolman and Bernstein, 1978). These matters are also important if we are to assess the performance and understand the role of computer consultation systems in medical practice.

In a recent bibliography of automated medical decision-making methods and systems (Wagner et al., 1978) over 800 references are cited, and these do not include many of the simplest state-of-the-art applications or the most complex AI methods. If all of these are taken into account, it is likely that closer to 2,000 articles have been written describing medical decision-making and consultation systems. Yet the effect of automated decision making on medical practice after 20 years of fairly intense activity has not been very dramatic. There have been some notable successes, such as automated EKG interpretation, which is now routinely available, and a few institutions have on-line consultative decision capabilities, but on balance, remarkably few systems have gone beyond the prototype stage.

There are many reasons for the slow introduction of computer-based decision systems into medical practice. Some are social, some technological, yet ultimately there is a simple pragmatic reason: such systems have rarely been shown to fill an indispensable need in the clinical setting. This picture may be beginning to change: with the proliferation of new special-purpose biochemical tests and the accelerated specialization of medicine, the demand for easy reference to up-to-date consultative advice and medical information is beginning to be increasingly recognized. Medical data bases that pool information from national networks of collaborating researchers (Fries, 1976), record-keeping systems with capabilities for retrieving general medical information and references (Schultz and Davis, 1979), and computer-based medical instruction and testing systems have gradually grown and spread during the past decade. The National Library of Medicine has recently moved in the direction of supporting research into the structure and organization of medical knowledge bases and the methodologies by which they can be kept up to date and disseminated to practitioners (Schoolman and Bernstein, 1978). This complements the ongoing support programs of research and computing resources for artificial intelligence in medicine (AIM) by the Biotechnology Resources Program of the Division of Research Resources of the NIH (Ciesielski, 1978; Freiherr, 1979).

Another technological impetus for change can be expected to come from the increased availability of microprocessors, which will make inexpensive computing readily available to practitioners in their own offices. Many are already experimenting with methods of encoding their decision logic in the form of simple algorithms, and there has been a notable proliferation of small medical-computing groups and societies in the past few years that have served to focus these activities. The automated interpretation of laboratory instrument results, particularly in clinical pathology, is also becoming more prevalent (Bieman, 1979; Speicher, 1978; Young, 1976). It is likely to stimulate a need for more extensive clinical decision systems that will back up and integrate the results from several different instruments, ranging over various systems of the body. The scope of an AI model of internal medicine, such as that developed for INTERNIST (Pople et al., 1975), the modularity and explanatory capabilities of MYCIN (Shortliffe, 1976), and the pathophysiological reasoning and efficiency of compiled expert knowledge available in EXPERT (Weiss and Kulikowski, 1979) will all be useful for such tasks.

Not all work on consultation methods and systems needs to be ultimately justified in terms of their application in clinical practice. Contributing to help organize medical knowledge and research and supporting medical education are two other important fundamental objectives. The AI systems are particularly relevant in both these regards, since they have concentrated not only on achieving good performance, but on justifying and explaining this performance based on models of diseases and patients.

Three recent reviews of medical decision methods and systems have included the artificial intelligence approaches (see Schoolman and Bernstein, 1978; and Chapters 3 and 9). An article by Szolovits and Pauker (Chapter 9) describes the four earliest AIM systems, contrasting categorical (deterministic) with probabilistic components of their reasoning strategies. A review by Shortliffe, Buchanan, and Feigenbaum (Chapter 3) emphasizes the symbolic reasoning nature of the AI programs and highlights the importance of explanation and updating facilities, as well as good conversational capabilities for interacting with consultation programs; the authors draw mainly on their experience with the MYCIN system for illustrative examples. The present paper takes a somewhat different approach in that it suggests a set of characteristic representational, reasoning, and control features for describing consultation programs, and then uses these as the basis for its comparisons.

4.1.2 Goals and Approaches of Artificial Intelligence in Medicine

In reviewing artificial intelligence approaches to medical consultation, it is important to characterize the concerns and goals of AI research that have influenced the work in this field.

The spectrum of research in AI can be described as ranging between two extreme approaches. The first stresses the development of theories of cognition through computer-based experimentation. Michie (1974) has given a definition of AI consonant with this view: "the development of a systematic theory of intellectual processes." In contrast, a more pragmatic concern of imitating or approximating the behavior of human problem solvers is expressed in the definition given by Minsky (1968): "The science of making machines do things that would require intelligence if done by men." The first approach shares many concerns with cognitive psychology. The major aspect of computer programs from this viewpoint is that their reasoning procedures must exhibit capabilities of understanding that imitate those used by human problem solvers. At the other extreme, the correspondence with human behavior can be viewed strictly in terms of the output performance of a computer system, regardless of whether the reasoning leading up to this performance simulates that of humans. Much of the problem solving done in robotics takes this approach (Winston, 1972). In a similar vein, research in pattern recognition, developing from engineering and the mathematical disciplines (Duda and Hart, 1973; Fukunaga, 1972), has stressed the importance of achieving accurate performance in detecting and classifying patterns, usually by mathematical and statistical techniques that do not attempt to parallel human reasoning.

Despite the contrast between the performance-oriented and understanding-oriented work in AI in the past, it can be recognized that these

represent complementary approaches that are open to researchers seeking to develop computer-based problem-solving programs. Recent work in the automated recognition of human speech (Lesser et al., 1975; Lesser and Erman, 1979) exemplifies the maturity of AI in developing systems that are oriented to both understanding and performance.

Expert medical consultation is a problem-solving process that draws on a rich, though incomplete, body of knowledge that is both empirical and conceptual in nature. Until the introduction of artificial intelligence methods, the reasoning of computer-based consultation programs relied primarily on normative knowledge (prescribed as norms or rules of reasoning) that is used directly in medical decision making. The major emphases of the AI approaches have been:

1. to clearly separate the domain-specific knowledge base of a consultation model from the reasoning and control strategies used by the consultation programs (this facilitates *modification* of the knowledge base, which is likely to require frequent changes for incorporating new results from medical research and practice);
2. to capture the expert medical knowledge about specific inferences or decisions in the form of modular rules that reference the concepts and facts of the medical domain, also organized in a modular fashion (this facilitates the *explanation* of a consultation program's reasoning processes, which is crucial to the acceptability of a computer-based system);
3. to develop logically powerful and expressive representations for describing medical concepts and facts (such as disease hierarchies and mechanisms and the corresponding courses of illness) that serve to *support* and *justify* the decision rules in terms of knowledge structures that are commonly used by physicians;
4. to experiment with a variety of reasoning and evaluation methods and to develop general strategies to control the reasoning (this introduces *flexibility* and the ability to recover from mistakes through alternative means of reasoning, hence giving a *fail-soft* capability);
5. to develop methods of facilitating *user interaction* with the programs, either by specialized natural language interpretation capabilities or by flexible command languages.

By incorporating many of the attributes described above, computer consultation programs are beginning to display some of the scope, depth, and flexibility of reasoning that characterizes expert human consultants. At the same time, the process of building these systems is uncovering new problems in the representation, application, and validation of medical knowledge.

4.2 Decision-Making Problems and Styles in Medical Consultation

4.2.1 Medical Consultation Tasks

The tasks involved in medical consultation depend on the nature of the advice that is being sought from the consultant. Whether it is feasible to capture some of the reasoning and problem-solving processes employed by a consultant within computer programs depends largely on the relative role of reasoning versus perceptual skills used by the human expert. If expertise in performing a specialized physical examination, involving the detection of subtle signs through visual, tactile, and other sensory cues, constitutes a crucial element of the expert's consultation, it is not reasonable to expect any current computer system to perform such a consultation. [Nevertheless, computer-based systems may provide valuable new modes of extracting perceptual information on the patient, such as by tomography (Kak, 1979).] If, on the contrary, the scope and definition of items in a review of systems and the elicitation of a medical history have been well determined for a given diagnostic or treatment selection problem and the major role of the human consultant is to provide a sophisticated interpretation of the findings, then it is not unreasonable to investigate such processes of interpretation and attempt to simulate their performance by computer-based systems. If, in addition, it is possible to build a knowledge base that incorporates both descriptive models of pathophysiological mechanisms as well as the normative components of expert reasoning, and if strategies of explanation can be formulated that permit the program to answer questions about its own reasoning, then it is not unreasonable to claim that such a system demonstrates certain elements of "understanding" not unlike those manifested by human problem solvers.

The major tasks of medical consultation that must be performed by a computer system can be summarized as follows:

1. the sequential elicitation of findings and the assessment of their reliability and internal consistency;
2. the interpretation of the findings in terms of a model of diagnostic classes and their relationships;
3. the extrapolation of the natural course that the illness is likely to follow (prognosis);
4. the formulation of various plans for therapeutic management and the selection of an initial treatment;
5. the explanation and justification of the above;

6. the reassessment of the patient's status on return visits and the reevaluation and possible modification of diagnostic, prognostic, and therapeutic conclusions.

At any given point in the course of a consultation, one of the tasks described above will be the main goal of the reasoning of the human or computer-based consultant. In a generalized consultation scheme these goals and their various subgoals (such as eliciting a specific finding or formulating a specific treatment for a given disease) must be explicitly represented if their sequencing is to be easily modifiable by the control strategy just as it is by the human consultant's strategy decisions. The principal types of medical facts and concepts and some of the reasoning links among them are shown in Figure 4-1.

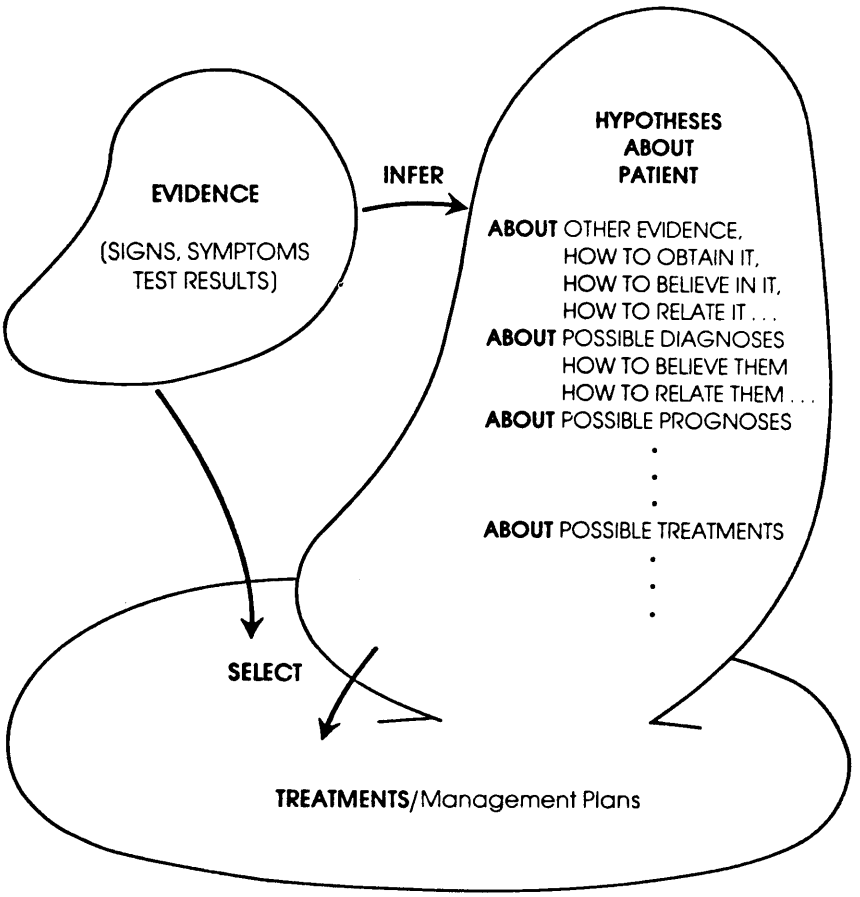

FIGURE 4-1 Problem solving in medical consultation.

The medical facts about an individual patient (findings) can be viewed as the direct *evidence* from which hypotheses about possible diagnoses, prognoses, and treatments are generated and tested. This evidence comprises the history and symptoms reported by the patient, the signs elicited by the physician during the course of an examination, and the results of specialized tests for detecting specific pathophysiological states or conditions. A data structure used for describing a finding can include details about its measurement technique, its range of values, its reliability, its timing, its cost, and its logical relation to other measurements. It will be associated by various relational links and rules of reasoning to the hypotheses.

Hypotheses usually require a very different descriptive structure. They stand for the major medical concepts used in reasoning, such as the diagnostic and prognostic categories applicable to the patient, but may also include a variety of intermediate constructs, such as syndromes, pathophysiological and clinical states, courses of illness, and clusters of clinical evidence. These intermediate concepts can be used to define the higher-level concepts. Although the major type of hypothesis is one that refers directly to the clinical condition of the patient, it is also possible that we may want to explicitly represent hypotheses that are assertions about related contexts (such as the environment of the patient, a relative of the patient, etc.). Some hypotheses may be subconcepts of others, in which case they may inherit properties of the parent concept; others may be causal antecedents, which implies that they must also occur in temporal sequence before their consequents.

A consultation system must also represent the various treatments that are potentially able to control the patient's illness. The treatments are interrelated in terms of applicability and risk/benefit factors: therapeutic effectiveness, toxicity, potential for undesirable interactions, and other constraints. To manage a patient with a complex or prolonged illness, a management plan must be formulated. The plan must consist of the various potential sequences of treatments that are available to control the alternative courses that may be followed by the illness after an initial treatment. In computer-based consultation schemes, it is important to represent a realistically large scope of alternatives and their relations to the hypotheses and findings of patients. On the basis of these relationships, rules for selecting treatments can be derived and explained.

A significant component of human consultative reasoning is often characterized as being judgmental. In designing computer-based systems, an immediate question arises as to how best to simulate such judgments, if indeed they are to be simulated at all. One school of thought holds that it would be best if they could be replaced by more objective methods, usually of a statistical decision-theory type (Grémy, 1976). But even with this approach, judgmental knowledge is needed to choose decision thresholds. Others have attempted to capture the expertise of human consultants in the form of reasoning rules that directly incorporate judgmental ele-

ments (Shortliffe and Buchanan, 1975). Regardless of the approach, the relative value of alternative reasoning outcomes (misdiagnoses, inadequate treatments, etc.) clearly enters into consultative reasoning. Thus computer-based schemes must include a representation of these values (also called *utilities*) to be used by their decision strategies. The exact manner in which such values are to be used depends on the structure of knowledge in the program, the overall strategies of reasoning, and the nature of the values involved. Values on outcomes will be very different if they are those of the patient rather than of the physician, and both will differ from any "average" or societal values for comparing outcomes. Pauker (1978) has recently discussed these problems from a decision-theory viewpoint. In addition, different experts may well disagree on how to treat a given patient, each giving a justification for his or her point of view. Such sources of variability ensure that in most situations there will be no single "correct" or "optimal" mode of treating a patient, and the role of a consultation system must be seen as one of presenting the alternatives, with a clear indication of the source for the value judgments that enter into each decision.

4.2.2 Types of Medical Consultation

The kinds of reasoning involved in medical consultation depend on the specific type of problem presented to the consultant. In the past, computer-aided methods have been used in the following consultative situations:

1. interpreting a single test and listing possible diagnoses;
2. screening the patient for a particular disease (or group of related diseases) from multiple tests and clinical findings;
3. performing some of the tasks of a primary care physician in acquiring information on the present illness of the patient, proceeding to a differential diagnosis, and making treatment recommendations if appropriate;
4. simulating the role of a specialist who is asked to provide interpretation and management suggestions for complex cases referred by primary care physicians.

The artificial intelligence consultation programs developed to date have simulated the last two types of consultation. They have been research prototypes, and it is not unreasonable to expect that if programs of this type are to become widely used in clinical practice, connections between them and the more basic types of single-test and screening programs will have to be developed.

4.2.3 Evolution of Formal Methods of Decision Making in Consultation

The applications of formal methods of decision making have concentrated on problems of diagnostic reasoning, though decision-analysis techniques have been applied to treatment-selection problems. The sequence in which different techniques have been introduced is approximately as follows:

mid-1940s: Statistical hypothesis-testing methods [mostly for screening and radiology (Yerushalmy, 1947); computations by calculator]

1954: Logical scheme for matching symptoms to diagnoses [slide rule (Nash, 1954) or hollerith cards used for sorting and matching]

1958: Statistical and logical techniques combined (Lipkin and Hardy, 1958) [computers introduced and used in most subsequent work]

1960: Bayesian and discriminant methods (Ledley and Lusted, 1959)

1968: Sequential Bayesian methods, and decision-theory approaches applied to treatment selection (McNeil et al., 1975; Schwartz et al., 1973) (also see Chapter 2)

1969: Pattern-recognition methods (Kulikowski, 1970; Patrick et al., 1977)

1970: Information-processing models for diagnosis (Wortman, 1972)

1971: Knowledge-based artificial intelligence systems (Kulikowski and Weiss, 1972; Pople et al., 1975; Shortliffe, 1976; see also Chapter 6)

Ledley and Lusted (1959) gave the first overview of the applicable methods from logic and probability, and the 1960s saw the introduction of various statistical, logical, and pattern-recognition techniques for diagnostic decision making. These methods, relying on large data bases of reliably diagnosed case histories, performed well in narrowly defined medical domains using a clearly specified (or standardized) set of patient findings. Lack of adequate statistics and problems of consistently introducing value judgments about possible misdiagnoses into the decision framework have proven to be important limitations of these methods.

A very different manner of encoding medical reasoning in a computer program has also been available: the sequence of decisions performed by a physician in reaching a diagnosis or choosing a treatment can be flowcharted and directly implemented as an algorithm. But insofar as the same conclusions may be reached by many different pathways and it is quite usual for experts to differ in their preferred sequences of tests and intermediate decisions for a given type of case, such a *flow chart algorithm approach* is usually too rigid and idiosyncratic to be widely accepted. However, characterizing the reasoning of an expert in a specialty can be useful for

teaching, for comparison with medical practice, and for guiding the decisions of physicians' assistants (Komaroff et al., 1974). Simple decision algorithms for patient self-help have been proposed recently as a technique of preventive medicine (Vickery and Fries, 1978), which may also reduce the burden on health care facilities. A mixed algorithm scheme is characteristic of one of the best-known consultation programs—Bleich's system for acid-base and electrolyte balance (Bleich, 1969). It intermingles the direct logical assessment of patient findings with calculations from mathematical formulas that describe the underlying biochemical changes.

To provide information about past experiences with prognosis and treatment, several different groups have relied on the *logical matching* of current patient profiles to prior stored cases in a large data base. The ARAMIS system in rheumatology at Stanford University (Fries, 1972; 1976) and similar ones in lung cancer at Yale University (Feinstein et al., 1972) and cardiovascular diseases at Duke University (Rosati et al., 1975) are well-known examples. The major methodological question for these systems is the form in which patient profiles are to be specified and the choice of query types that can be easily supported by the data base structure. Although they have not addressed the problems of how to incorporate their results into the broader interpretation of a patient's condition, they represent an important step in the direction of standardizing knowledge about the time course of diseases within a data base. And insofar as all interpretation is left to the physician using the system, they have been more readily accepted than many of the consultation programs.

In the late 1960s and early 1970s various *pattern-recognition methods* began to be applied to medical decision making (Kulikowski, 1970; Patrick et al., 1977). In some instances they provided the means of overcoming the limitations of small-sized statistical samples through the use of well-chosen heuristics; in others they enabled the summarization of large numbers of findings through synthetic "features" (Kulikowski, 1970), but in common with the statistical approaches, they suffered from being a "black box" approach to medical reasoning. That is, the patient's findings would be transformed mathematically into some heuristic score or weight, which would then become the sole basis for ranking diagnoses or treatment recommendations.

Figure 4-2 shows a schematic diagram of a typical pattern-recognition or statistical system for medical consultation. The sequence of operations specified by algorithm typically consists of a preprocessing, or filtering, to extract the set of patient findings relevant to the clinical problem under consideration, and the extraction of features (logical or mathematical transformations) that when selected for best discriminatory performance enable the classifier to be both simple and effective. The domain-specific knowledge base used by the algorithm is composed of various patterns of association between findings and hypotheses (for statistical methods), profiles of correctly diagnosed cases (for nonparametric sample-based methods, such as the nearest-neighbor technique), or explicit sequences of decisions (for the flowcharting methods). Most programs implementing these meth-

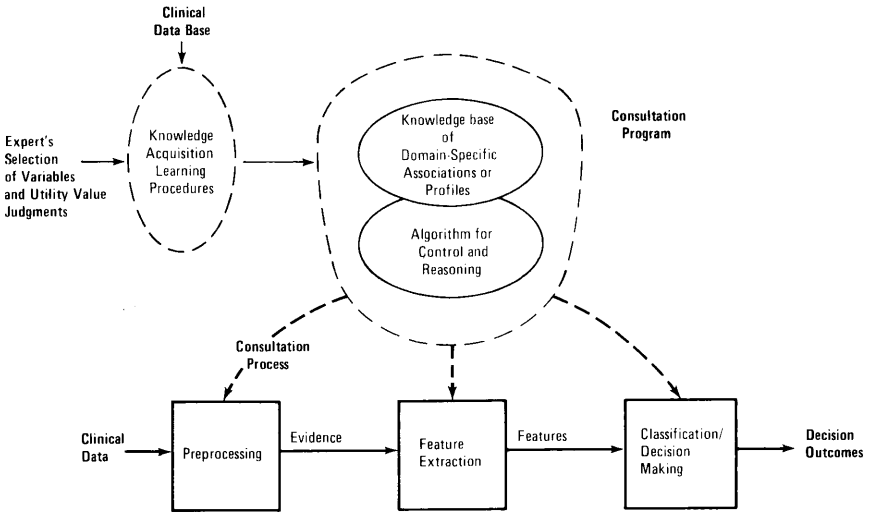

FIGURE 4-2 Statistical or pattern-recognition system for consultation.

ods intermingle elements of domain knowledge and reasoning mechanisms under algorithm control in a relatively fixed manner. The outcome, rather than the process of reasoning, is the main concern, so considerations of computational efficiency often override the possibility of introducing more flexible or general modes of reasoning that would come closer to imitating human expert behavior.

In designing such a system, the knowledge acquisition phase usually consists of analyzing the data base of clinical cases that have well-established diagnostic and treatment endpoints. The decision rules to be used by the classifier can be "learned" by various techniques (Chilanski et al., 1976; Duda and Hart, 1973; Fukunaga, 1972). The medical expert defines the scope of the problem by specifying the variables that are to be examined in the data base. If a decision-analysis method is to be used, the expert must also provide the utility or cost factors (and prior probabilities of hypotheses for subjectively estimated situations) to be used as part of the decision rule thresholds (McNeil et al., 1975).

The application of *artificial intelligence* methods sought to remedy the "black box" situation by introducing a structure of knowledge familiar to the physician into the decision-making schemes. The approach of using a computer-based model to study the decision making of clinicians was begun by researchers interested in cognitive processes. Kleinmuntz and McLean (1968) developed a program for simulating a consultation session in neurology, and Wortman (1972) developed an information-processing model for medical reasoning (including conceptual hierarchies and memory mechanisms). Initial prototype consultation programs using AI con-

cepts were developed in ophthalmology [CASNET (Weiss et al., 1978)], infectious diseases [MYCIN (Shortliffe, 1976)], internal medicine [INTERNIST (Pople et al., 1975)], and renal disease [PIP (Pauker et al., 1976)], while an article by Gorry (see Chapter 2) advocated the introduction of conceptual structures, language development, and explanation into medical decision-making systems.

All of the AI approaches use heuristic measures for scoring the weight of confidence or credibility that they assign to a hypothesis as an explanation of the patient's condition. These measures are typically computed from uncertainty weights attached by the human experts to the various reasoning rules in the consultation model. The reasoning strategies of all of the systems, however, rely as much on the structure of connectivities among concepts and between concepts and facts as on the scoring mechanisms themselves. This provides the systems with a natural way of supporting explanations, and often allows alternative and sometimes redundant lines of reasoning to be pursued, giving a measure of flexibility to their behavior.

Contemporary with the evolution of the AI approaches, several other investigators have introduced constraints and intermediate reasoning constructs into probabilistic frameworks. These include Bayesian approaches (Patrick, 1977; Warner, 1978) and a latent factor method (Woodbury and Clive, 1980). Fuzzy logic has also been applied to diagnostic problems (Wechsler, 1976).

The subsequent sections review the early AI systems and trace the evolution of the knowledge-based schemes that have been developed to the present.

4.3 Artificial Intelligence Methods in Consultation

4.3.1 A Comparative Overview of Early AI Consultation Systems

In this section we discuss the first major AIM systems—CASNET, MYCIN, INTERNIST, and PIP, each of which is described in greater detail in later chapters.

CASNET/Glaucoma Consultation System

A causal-associational network (CASNET) was developed as a means of representing the pathogenesis of a disease, in terms of which the patient's findings are interpreted. The causal relations, with associated degrees of

strength, express not only the mechanisms of a disease but also their modifications under various regimens of treatment. Different patterns over the causal network are associated with the various elements in a classification scheme of diagnostic hypotheses, which can include degrees of severity and progression of a disease. Appropriate treatment plans can be associated with the diagnostic hypotheses, and specific treatments within the plans are related to each other by constraints of how they cover for particular illnesses, how they may interact, etc. Normative knowledge is in the form of inferential rules linking patient findings to the intermediate hypotheses about pathophysiological states and preference rules linking findings to treatments. Uncertainty measures on these links range from $+1$ for full confirmation to -1 for full disconfirmation.

The reasoning control strategy of CASNET can be characterized as mainly event-driven: the incoming clinical data trigger the inference rules that assign weights to the pathophysiological states. A thresholding evaluation mechanism then yields a logical status of "confirmed," "disconfirmed," or "undetermined" to each causal state. The subgraph of confirmed and undetermined states forms a *patient-specific interpretation model* at every stage of the consultation. The system uses the causal model to constrain the search for possible hypotheses by guiding the requests for further patient data. This is carried out by first propagating direct and inverse causal weights throughout the net every time a data item is entered. Such a global assessment is made efficient by the partially ordered and precompiled nature of the causal net. Once the weights are computed, the choice of next question is hypothesis-driven: a criterion of maximal diagnostic information for a given cost range guides the selection that will add to the weight of evidence of the most likely intermediate hypothesis (state). This strategy may be superseded by domain-specific strategies for data acquisition, which can encode prespecified protocols given by experts; this was the case in the specialized CASNET/Glaucoma system. When all the data having a bearing on the consultation have been accumulated, the system carries out a final evaluation over the entire causal net, producing a weighting of the root nodes (primary causes). These trigger the higher-level diagnostic, prognostic, and treatment categories in a purely deterministic fashion. The choice of specific treatment, including the dosage, mode of administration, and time course, is then carried out by evaluation over the preference rules. These contain the various restrictions on the applicability of treatments, such as allergies, past history of treatment effectiveness, drug interactions, and so on.

A knowledge-base acquisition program for building CASNET-type models was developed at Rutgers University (see Chapter 20), and an in-depth model for consultation in the glaucomas was built incorporating the knowledge of clinical experts from five major ophthalmology research centers. The consultation model was tested with many cases of disease (from the U.S. and Japan) and participated in a national symposium on glaucoma, performing at an expert level (Lichter and Anderson, 1977).

MYCIN/Infectious Disease Therapy Consultant

A system of production rules with associated uncertainty weights serves to capture most of the expert knowledge in MYCIN (Shortliffe et al., 1973; Shortliffe, 1976). Rules are of the following form: IF premise assertions are true, THEN consequent assertions are true with confidence weight X. The assertions can be Boolean combinations of clauses, each of which consists of a predicate statement about an <attribute, object, value> triple. The triples represent medical facts and hypotheses about the patient and related objects or contexts, such as infections, cultures, and organisms. For example, <GRAMSTAIN, E.COLI, GRAMNEG> stands for "the gram stain of the *E. coli* organism is gram-negative." Goals and subgoals of the consultation process, such as "select therapies to cover for all diagnosed infections," can also be explicitly represented by the predicate structure of an assertion.

The uniformity of representation for both domain-specific inferences and reasoning goals makes it possible for MYCIN to use a very general and simple control strategy: a goal-directed backward chaining of rules. In this approach, the first rule to be evaluated is one containing the highest-level goal—to select treatments for all the infections of the patient. This requires that the infections be known. But since they are usually unknown, the system must then try to satisfy subgoals that will allow the infections to be inferred. Discovering the results of cultures or other clinical parameters of the patient would be the most direct subgoals. These in turn may be deduced from other rules, but eventually the attempt to satisfy rule premises will end with assertions that can only be confirmed by directly questioning the user for the appropriate information. Once this happens, the system can begin to reason deductively by successively satisfying subgoals that it had previously unwound. A hierarchical tree of contexts (patient-infections-cultures-organisms) anchors and constrains the order in which the rules are invoked. This, together with a network of links among clinical parameter values and the templates for the parameters, constitutes the descriptive component of the MYCIN knowledge base.

The reasoning evaluation mechanisms include a fuzzy logic function for combining the effect of uncertain assertions within a rule (a minimum for conjunctive and a maximum for disjunctive combinations) and a heuristic cumulative function to add the confidence weights from rules with different sources of evidence in their premises (Shortliffe and Buchanan, 1975). The confidence weights (or factors) are expressed on a scale from -1 for complete disbelief in an assertion to $+1$ for complete belief. Separate measures of belief and disbelief are used in updating hypothesis weights, because of the need to avoid the probabilistic constraint that an assignment of probability P to a hypothesis implies a probability of $1-P$ for its negation. Shortliffe developed his scheme of confidence factors to provide physicians with a means of expressing their belief or disbelief in a hypothesis independently of one another. Although the MYCIN reason-

ing strategy is almost entirely based on the rule evaluation procedures, the final selection of therapy is carried out by a specialized algorithm, which uses the deduced knowledge of the patient's infections, the causative organisms, and the ranking of drugs by sensitivities and preference categories (of effectiveness).

The MYCIN system places special emphasis on the modular nature of its knowledge and on the ease that this modularity entails for generating explanations. A question-answering program interacts with the performance program to find out about the reasoning sequences leading to a given conclusion and the reasons behind the latter's requests for patient data. The user interface of the system has been developed with careful attention to its "friendliness" and the capability to express its rules in English. The system is able to understand a domain-specific vocabulary of commands and descriptions of patient-related facts, using a keyword-recognition scheme. Various Interlisp facilities are used to advantage in giving the system a good "conversational style." There have been formal evaluations of the MYCIN system by a number of independent consultants that demonstrated that the program performed at a level comparable to that of experts (Yu et al., 1979a; 1979b).

INTERNIST/Diagnostic Consultant in Internal Medicine

One of the principal aims of INTERNIST system development has been to explore the manner in which expert clinicians reason about diagnosis when the space of possibilities is large and hierarchically structured, as in internal medicine (Pople, 1975; 1977; Pople et al., 1975). The program uses a knowledge base in the form of a hierarchy of diseases, from the general (liver disease, heart disease, etc.) to the specific (hepatocellular infection, aortic stenosis, etc.), with the typical findings linked to the most specific form of each disease group. Other links include finding-to-disease evocation and disease-to-disease causal connections. A cost-related specification (history questions, signs, or the more expensive tests) and global weights of import are attached to the findings. There are uncertainty weights associated with most of the links, expressed on a scale that ranges from 1 (for least confirmation) to 5 (for maximum confirmation). The weights are subjectively estimated by the medical expert.

The reasoning strategy of INTERNIST begins in an event-driven fashion: the initial data presented to the system evoke a set of related disease hypotheses. For each of the evoked diseases, the system builds a patient-specific model, consisting of four lists: observed findings consistent with the disease, those unexplained by the disease, findings as yet unobserved that would be consistent with the disease, and those that ought to be observed if the disease is the correct diagnosis. Each disease model is scored positively for explained findings and negatively for the unexplained ones, with the individual findings weighted according to their importance.

Bonuses are added to hypotheses that are linked causally to other confirmed diseases. A partitioning heuristic then splits the space of hypotheses into those that compete and those that complement the most highly ranked one. For example, if thyroid carcinoma is found to be the most likely disease from the first evaluation, diseases like a thyroid cyst would be competing hypotheses, whereas a heart disease would be complementary in that it accounts for other findings largely unrelated to the thyroid problem.

Once the partitioning is completed, a number of different strategies may be pursued by the system, depending on the size of the competing hypothesis set. If there are more than four competitors, the system will try to rule them out by asking questions about the findings that are expected to be present in the disease. If the number of alternatives ranges from two to four, a discriminatory strategy is followed that consists of seeking results that are strongly indicated by one disease but only weakly indicated by the other. Finally, if there are no competitors, the strategy will ask for data that will strongly confirm the highest-ranking hypothesis. When this process has been completed by the confirmation of the first major disease (or one of its competitors), the program repeats the cycle with the next most highly ranked hypothesis in order to account for findings that remain unexplained. This process continues until all findings have been accounted for. The reasoning of INTERNIST is therefore strongly focused around the highly ranked hypotheses once the initial phase of data entry is completed.

The INTERNIST system has been reported to cover a large proportion of the field of internal medicine and is routinely tested with complex cases from clinical-pathological conference case reports in the major medical journals (Pople, 1977). Once its knowledge base has been expanded sufficiently, it is expected to be tested outside the University of Pittsburgh in a formal manner. The system is also being used for educational purposes, and it is expected to be linked to other diagnostic systems (Freiherr, 1979).

PIP/ Present Illness Program

To develop an understanding of the problem-solving methods used by physicians for patients who present with a varied and potentially large set of complaints was the underlying motivation of the project in clinical cognition (see Chapter 6). The system, developed at M.I.T. and Tufts–New England Medical Center, evolved from Gorry's proposal to introduce conceptual structure to guide and support reasoning in diagnosis (see Chapter 2). The representation chosen for the system was the frame scheme developed by Minsky (1975). A *frame* is a prototypical description, which in PIP is centered around disease categories. Each frame is a structure with a name and a number of slots, which can be filled by various properties, logical and semantic relations, and associated inference rules. The disease

frames in PIP contain slots for descriptive relations (causal, complementary, complicational, etc.), logical conditions (necessary and sufficient findings), and reasoning rules of various types (suggestive, discriminatory, or conclusive rules). The most important slots are those containing a listing of evocative or triggering findings, and a listing of expected findings. Like CASNET and INTERNIST, PIP initiates its reasoning in an event-driven fashion: the initial data trigger a number of hypotheses, which are then considered to be "activated." PIP maintains a three-level status for its hypotheses during a consultation. All start out in long-term memory, with inactive status. Once a hypothesis is activated, it brings along all hypotheses that are directly complementary to it into "semiactive" status. A semiactive hypothesis is eligible to become active if any one of its typical findings is found to be true, whereas "inactive" hypotheses can only be activated by their triggering findings.

Once the reasoning process begins by triggering, the system attempts to "fill in the frame" by asking questions that will tend to confirm it or rule it out. This may be done categorically by matching findings that are logically sufficient or necessary (MUST-HAVE or MUST-NOT-HAVE relations) or probabilistically by thresholding a local score evaluated for the hypothesis. This score is computed from the uncertainty rules associated with the frames and has two components: a measure of the fit of observed-to-expected findings for the hypothesis and a ratio of the number of findings explained by the hypothesis to the total number of observed findings. PIP also propagates scores so that the effect of findings that are explained by lower-level hypotheses—the clinical or pathophysiological states, such as "nephrotic syndrome"—can be taken into account in the likelihood computations of hierarchically or causally related hypotheses (such as "glomerulonephritis"). The sequential questioning of the system is therefore hypothesis-directed in that the filling of a frame results in asking about its expected findings or those that will discriminate it from other hypotheses. Focus is shifted to other frames once the truth value of the original one has been established with a sufficiently high level of certainty. The process continues until all reported findings have been accounted for.

PIP was an experimental system, and it was tested with a knowledge base of about 70 hypothesis frames in renal disease and related disorders (see Chapter 9). Problems were uncovered in maintaining a sufficiently focused and clinically acceptable line of reasoning, and this contributed to a shift in emphasis toward more tightly structured and physiologically determined domains (acid-base balance and digitalis therapy) on the part of its developers (Gorry et al., 1978; Patil, 1979; Silverman, 1975). It has been suggested that one major reason for the difficulty of generating lines of reasoning that parallel those of clinical experts lies in the use of generalized scoring functions and in termination criteria that lead to exhaustive explanations of the observed findings (Szolovits and Pauker, 1979) (also see Chapter 9). When several top-ranking hypotheses have scores that are close in value, reflecting a very ambiguous case, the interpretation of additional

data may often result in rapid changes in the focus of the reasoning, as one piece of evidence pushes the score of one hypothesis above that of its competitors, and then another finding elevates the score of an alternative hypothesis above that of the first. To avoid an overdependence on scoring functions, all AIM systems have tried to incorporate into their knowledge bases as many categorical reasoning links as possible.

4.3.2 Characteristic Elements of AIM Consultation Systems

The four initial AIM systems and their successors all share certain characteristic properties. Figure 4-3 illustrates some of the principal components of the systems and the resulting consultation process.

In contrast to the pattern-recognition and statistical approaches, there is a deliberate separation of the domain-specific knowledge base, the general mechanisms of evaluation, and the control strategies of the system. The reasoning evaluation and control components are sometimes called the *inference engine* (Davis, 1979; Feigenbaum, 1977). The knowledge base is often also clearly divided between a *descriptive component* of data structures linked by domain-specific relations (hierarchical categorizations, subcomponent membership, causal precedence or antecedence, etc.) and a *normative component* of prescriptive reasoning rules that operates over the descriptive component using the evaluation mechanisms in a manner specified by the control strategies. This organization can be viewed as a specialized variant of the structure used in generalized production systems in AI (Newell and Simon, 1972; Nilsson, 1980).

In CASNET, INTERNIST, and PIP the reasoning process is centered around an explicit, structural descriptive component. The causal nets and hierarchical taxonomies can be viewed as special cases of *semantic networks* (Quillian, 1968), which were the first and most widely used means of representing knowledge for natural language interpretation. The *frame* (or unit) schemes offer a very natural alternative way of representing knowledge, which emphasizes the "chunking" or partitioning used by human experts to separate different topics, concepts, or hypotheses. The normative or reasoning knowledge in these systems is expressed as decision rules or procedures attached to the nodes of the semantic net, or as logical constraint conditions contained in the frames.

In contrast, MYCIN centers its knowledge around the normative component: the *production rules*. Its descriptive component is deemphasized, although the context tree and network for updating values of clinical parameters are crucial to the effective invocation of rules. This approach may facilitate the acquisition of the strictly inferential knowledge, but leaves open the question of how to relate the specific productions to prototypical concepts in the medical domain. The context tree does this, but in a very specific and understated manner. It has been suggested that the operation

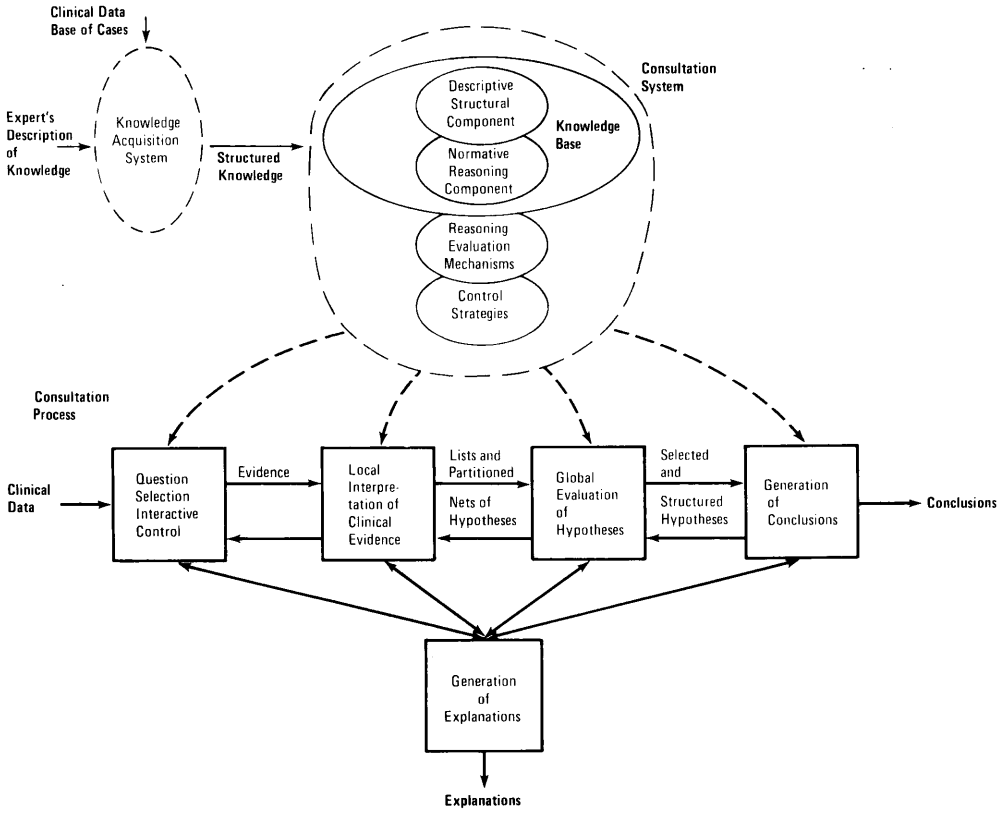

FIGURE 4-3 Artificial intelligence system for medical consultation.

of MYCIN could be turned "inside out" (see Chapter 9), with the contexts represented by frames, which will be filled up as the production rules that are attached to them are evaluated. The recent implementation of a mixed frame-and-production-rule representation [the CENTAUR system (Aikins, 1979; 1983)] has shown this to be a feasible approach.

Methods for quantifying uncertainty vary from system to system but share certain common properties: they treat confirmation and disconfirmation of hypotheses as independent processes (although combining functions are needed to produce measures of overall confidence for guiding the course of reasoning); the number of distinct uncertainty levels subjectively estimated by the experts is usually five or six; and they use fuzzy logic combining functions for evaluating the uncertainty of a Boolean combination of assertions.

Depending on the complexity of the consultation task, reasoning mechanisms may include: focus-of-attention heuristics to concentrate on a subspace of the space of possible hypotheses; pattern-matching mechanisms to actively scan incoming data for patterns that will trigger a hypothesis; goal generators to specify how sequences of subgoals ought to be pursued; global evaluation heuristics to piece together the results of several partial interpretations; and explanation mechanisms for tracing the reasoning. The control strategies specify how the different reasoning mechanisms are to be invoked, either automatically or in response to interactive commands given by the user.

The characteristic flow of information illustrated in Figure 4-3 shows that after an initial set of clinical data has been presented to the program, the control strategies can lead it to generate local interpretations (such as deciding on the normality, abnormality, or consistency of findings, or their interpretation in terms of directly related hypotheses), request more data as suggested by the initial interpretation, proceed to a global interpretation over the entire knowledge base (evaluating and comparing the partial interpretations, and selecting the most likely and coherently structured groupings of hypotheses), generate conclusions (integrating the various hypotheses into a final statement), and produce explanations for any of the preceding stages. The ability to recycle through previous stages of reasoning, allowing the user to request explanations and possibly changing the focus of reasoning by selectively introducing new data, introduces a significant degree of flexibility and generality that is characteristic of the AI approaches. It is interesting, however, that those consultation systems that give advice on treatment have done so without resorting to general methods of planning (Sacerdoti, 1977). This may reflect the fact that many treatment plans in medicine are short in length and center around the control of a limited number of clinical or physiological variables, making it possible to use relatively simple strategies of selection over prespecified alternative plans.

In building an AI consultation system, we rely more heavily on the knowledge of medical experts than in building probabilistic or pattern-recognition systems. The variety of structures employed by experts results in a much more complex knowledge acquisition process than must be faced by designers of the traditional systems, and a considerable effort has been devoted to these problems by subsequent AI system developers.

4.4 Evolution of AIM Systems and Knowledge Engineering

While the initial AIM systems were still evolving, several other systems were designed, taking advantage of the experiences and results obtained in the first cycle of development. The Digitalis Therapy Advisor (Gorry et al., 1978; Silverman, 1975) combined a single-compartment mathematical model for the effects of digitalis treatment with symbolic reasoning methods for the interpretation of patient-specific findings. After arriving at an initial determination of digitalis dosage based on the mathematical model, the system uses feedback information about the patient's clinical response to the dose (including both quantitative aspects, such as serum digoxin level, and qualitative cardiac signs and symptoms) to modify its recommendations for subsequent digitalis levels. The system was subjected to careful formal evaluation (Gorry et al., 1978), which demonstrated that its recommendations were comparable in effect to those of the clinical experts, suggesting that the system might be useful in health care situations where expert cardiac consultation is unavailable.

A generalization of the CASNET representational structures was included in the IRIS system, which used a semantic net to represent the descriptive knowledge of disease processes, reasoning primitives, and control states (Trigoboff and Kulikowski, 1977). IRIS was designed as a tool for experimenting with different reasoning and control strategies, rather than as a complete consultation system. It provided the user with a general mechanism for instantiating domain-specific facts and hypotheses and a mechanism for propagating inferences between them based on production rules. Specific control strategies could be written in Interlisp making use of the knowledge-base structure and reasoning elements of IRIS. Parts of the control strategies of MYCIN, INTERNIST, PIP, and CASNET were easily emulated in this manner. The MEDICO system, also applied in ophthalmology, used semantic and inference networks for knowledge acquisition (Walser and McCormick, 1976) and the design of a consultation system.

The PROSPECTOR system similarly combined the modularity of a rule-based scheme [using subjective Bayesian inferencing (Duda et al.,

1976) rather than the confidence-weight method of MYCIN] with a semantic network representation (Hart and Duda, 1977). Although this is a mineral exploration consultant rather than a medical consultant, PROSPECTOR is important in that it introduced the concept of a *partitioned semantic net* (Hendrix, 1975) to facilitate the attachment of rules to the appropriate set of semantic categories.

The facilitation of knowledge acquisition from experts and the updating of MYCIN-type models were the goals of the TEIRESIAS system (Davis, 1979). The system works mainly by analyzing mistakes of the consultation program, displaying the facts of the specific consultation case, the rules used by MYCIN, and its trace of reasoning. It then engages in a high-level dialogue (in a restricted set of natural language) with the expert builder of the knowledge base to try to discover the procedures by which the errors can be avoided. This knowledge is interpreted by TEIRESIAS so as to suggest possible changes in the rules of the consultation program. Taken together with the consultation model, TEIRESIAS represents an important example of a system that "knows what it knows," at least in the sense that one part of the representation can be used to represent properties and reasoning about another part. A different application of MYCIN techniques led to a consultant to help in the analysis of cases in the data base, which was implemented for use with ARAMIS (see Blum and Wiederhold, 1978; and Chapter 17).

The need to emulate the sequence of expert reasoning more accurately led to a new formulation of INTERNIST. The main concern was to develop a representation that would support strategies for handling multiple or composite hypotheses and would yield performance that converged more rapidly to the correct conclusions. Some of the elements introduced in the INTERNIST-II (Pople, 1977) system were *constrictor relationships* for describing very specific associations between findings and higher-level hypotheses, a multiproblem hypothesis generator with a modified scoring heuristic for taking advantage of the constrictor links, and control strategies for evaluating complexes of hypotheses rather than the individual hypothesis structures of the original system. Most recently, a knowledge-acquisition front end for INTERNIST has been adapted from the ZOG system, permitting the domain experts to enter their knowledge in a more natural manner (Freiherr, 1979). The problem of representing groups of related hypotheses in such a manner that they are "aggregated" in a natural way during inferencing has been a topic of concern for all of the researchers who deal with large hypothesis spaces. This question is a major consideration in the design of a new program for acid-base balance diagnosis and treatment (Patil, 1979).

A major problem that has not been adequately dealt with in the current consultation schemes is that of reasoning over temporal sequences of events and hypotheses. One approach to this problem, based on a real-time rule reevaluation within a MYCIN-like scheme, has been applied in the VM

system (Fagan, 1979) for ventilator management. In this application, the goal-directed strategy of MYCIN was not used, since the system must respond in an event-driven way to the changes in physiological status of the patient on the respirator. The inference of changes in hypotheses over the long-term course of chronic disease states was modeled in the CASNET/Glaucoma system by specialized time-dependent functions, and feedback of physiological parameter values is used in the reasoning of the Digitalis Therapy Advisor (Gorry et al., 1978). These examples represent specialized applications, and a general scheme for reasoning over time is still needed.

The *explanation* of reasoning has been a major concern of AIM systems, which has been extended recently to include tutorial advice in the GUIDON system (Clancey, 1979a; 1979c) for MYCIN-like consultants. An explanation scheme that is based on physiological and frame-based models has been developed for the Digitalis Therapy Advisor (Swartout, 1981).

A perennial problem for the designers of knowledge-based consultation programs has been to balance the mixture of declarative and procedural knowledge forms in their representations. In general, this has been alleviated by combining frames or semantic nets with production rules, as in IRIS (Trigoboff and Kulikowski, 1977), PROSPECTOR (Hart and Duda, 1977), CENTAUR (Aikins, 1979), NEUREX (Reggia, 1978), and NEUROLOGIST (Catanzarite and Greenburg, 1979), and in the knowledge-based schemata of EXPERT (Weiss and Kulikowski, 1979) and AGE (Nii and Aiello, 1979). Related to this are questions of modifying the control strategies so that the right kind of knowledge is applied to each problem-solving task, which have not as yet been explored in depth. A first attempt in this direction is the MDX system (Chandrasekaran et al., 1979), which develops a hierarchy of different "procedural experts" within a consultation system, with strict transfer of control protocols between them. The structure of experts in MDX directly parallels the links among the subspecialties of medicine. More research is needed to study not only this but other more flexible ways in which the control of concurrently operating experts can be coordinated.

As the number of examples of consultation programs and schemes increases, some common sets of techniques are beginning to emerge, which has led to the building of general tools for the construction of knowledge-based *expert systems*. This work has been characterized recently as *knowledge engineering* (Feigenbaum, 1978). Some of the general schemes for helping to build knowledge-based systems are EMYCIN, EXPERT, and AGE. The EMYCIN (van Melle, 1979) scheme is an outgrowth of MYCIN and permits the creator of a knowledge base to organize it so that it can be run with the MYCIN consultation control structure. Consultation programs in psychopharmacology (Brooks and Heiser, 1979) and structural analysis (Bennett and Englemore, 1979) illustrate the range of applications modeled with this representational scheme.

The EXPERT system (Kulikowski and Weiss, 1982; Weiss and Kulikowski, 1979) draws primarily on the CASNET experience and also provides a generalized consultation program that can be fitted with a knowledge base in any chosen medical specialty. Its representational scheme includes a hierarchical-causal network for hypotheses and treatments, a structured scheme for findings, and a set of production rules that permit the specification of contexts in terms of these elements. Models in rheumatology, neuro-ophthalmology, and endocrinology are being developed using this scheme (Freiherr, 1979). The system is designed so that physicians with some computer experience can construct models by writing them onto a file (with any system editor) using a simple descriptive language. The file is then compiled by a special program that checks for syntactical errors and produces a compiled model that can be efficiently run by the consultation program. Data-base updating and knowledge acquisition are also available to help in the process of debugging the model as it is tested against cases with reliable conclusions (Weiss and Kulikowski, 1979). A version of the system has been implemented on a minicomputer, thereby facilitating its dissemination to clinical environments.

The AGE (attempt to generalize) system (Nii and Aiello, 1979) provides a general set of technical tools for modeling consultative situations using the "blackboard" model (Lesser et al., 1975; Lesser and Erman, 1979), which was developed for handling the representation and processing of information from multiple sources of knowledge in speech understanding. Building a consultation model in AGE requires knowledge of Interlisp facilities, so this system is designed primarily for use by computer scientists working with medical specialists. Since the development of models that perform at an expert level has been shown to call for intensive interdisciplinary collaborations, such an approach is likely to continue as the main mode of research, at least until there are more experts who combine advanced training in both fields. Thus the current stage of development of knowledge engineering for medical consultation is one of constructive expansion in a number of varied applications. The next few years are likely to see many efforts at validation and application of these systems in realistic clinical environments.

The practical advances in developing knowledge-engineering tools continue to uncover new problems of a formal nature concerning representation, inference, and control in consultative problem solving. There is no lack of candidates for the title of "most difficult problem" when we attempt to study or emulate aspects of expert human reasoning on the computer. If a single set of problems qualifies for major attention, it might be those centered around the properties of concept abstraction and self-referencing that we associate strongly with "knowing what we know." Issues of concurrency in reasoning and related questions of whether and how to maintain logical and semantic consistency of the knowledge bases also present crucial open questions. These and other problems will continue to

offer sufficient challenges of an epistemological and formal nature and are likely to encourage active research that will parallel the engineering efforts for many years to come.

ACKNOWLEDGMENT

Part of this work was supported by a grant (RR 643) from the Biotechnology Resource Program, Division of Research Resources, National Institutes of Health.