AI Magazine Volume 6 Number 2 (1985) (© AAAI)

# Tenth Annual Workshop on Artificial Intelligence in Medicine: An Overview

B. Chandrasekaran

*Laboratory for AI Research, Dept. of Computer and Information Science, The Ohio State University, Columbus, Ohio 43210*

Jack W. Smith, Jr.

*Laboratory for AI Research, Dept. of Computer and Information Science and Dept. of Pathology, The Ohio State University, Columbus, Ohio 43210*

## AIM Workshops—A Tradition

The Artificial Intelligence in Medicine (AIM) Workshop has become a tradition. Meeting every year for the past nine years, it has been the forum where all the issues from basic research through applications to implementations have been discussed; it has also become a community building activity, bringing together researchers, medical practitioners, and government and industry sponsors of AIM activities. The AIM Workshop held at the Fawcett Center for Tomorrow at Ohio State University, June 30 - July 3, 1984, was no exception. It brought together more than 100 active participants in AIM. Hosted by the Ohio State University Artificial Intelligence Group, it was co-directed by Prof B. Chandrasekaran of the Computer Science Department and Prof. Jack Smith of the Department of Pathology in the College of Medicine. The Workshop was sponsored in part by the Rutgers Research Resource on Artificial Intelligence in Medicine under Grant RR2230 from the Biotechnology Resources Program of the Division of Research Resources, National Institutes of Health; by the American Association for Artificial Intelligence, who especially supported attendance by graduate students; by Ohio State University; and by contributions from numerous industrial organizations.

The OSU AI Group has been active in AIM research for a number of years with researchers from both the Computer Science Department and the College of Medicine. A number of systems for medical decision making, experimenting with new ideas for knowledge organization and problem solving, have been built there. The College of Medicine has just started a center for research in knowledge-based systems in medicine. Thus, after a number of years when the Workshop had been hosted by the AIM groups of MIT, University of Pittsburgh, Rutgers, and Stanford, sometimes in conjunction with major AI and medical computing conferences, holding the Workshop at Ohio State University was an indication of a broader base for research activities in AIM.

This report gives an overview of the Workshop discussions, without any claim of being complete or even representative—a report of this kind can only be an impressionistic account. We apologize in advance for omissions or distortions of points of view that were represented at the various Workshop panels.

## The Workshop Program

Presentations at the Workshop covered a wide spectrum and came in various forms: project reports, demonstrations of working AIM systems, and panel discussions. There were also two "after dinner talks," one each by Prof. Ed Feigenbaum and Prof. Herb Simon. Ed Feigenbaum reviewed the history of artificial intelligence in medicine and its seminal role in developing the knowledge engineering and expert systems approaches, following it with a highlight on new generation architectures being investigated at Stanford. Herb Simon gave an account of the evolution of his own thinking on the importance of research on learning for AI, following it with a description of recent advances in learning research and their significance. There were more than a dozen project reports summarizing

We thank Kaz Kulikowski and Priscilla Rasmussen of Rutgers University for their great help in organization of the Workshop One of the particularly satisfying aspects of this Workshop was the attendance by a large number of graduate students and medical fellows active in AIM research; this was made possible by a generous grant from AAAI Chris Putnam and OSU AI graduate students worked very hard in taking care of a number of details. We thank Sriram Mahalingam, David C Brown, John Josephson, John Svirbely, Jon Sticklen, Tom Bylander, Ann Kueneke, Mike Tanner, and Bill Punch for providing summaries of the various sessions for us to use in this report

progress in specific application projects. The presenters came from almost all the institutions in the U.S., as well as from some foreign countries, where AIM work is going on. Cleveland Clinic, University of Marseilles, MIT, University of Missouri, Ohio State, Rutgers, Stanford, Yale, and the University of Toronto were some of the institutions represented.

As mentioned earlier, working programs were demonstrated The Stanford Group presented two major demonstrations. One was on NEOMYCIN and GUIDON, projects directed by Bill Clancey at Stanford; these programs investigate new ideas for knowledge structuring and explanation of decisions for AIM systems. The other demonstration from Stanford was on the system called ONCOCIN, a project conducted by Ted Shortliffe, Larry Fagan, and associates. ONCOCIN is an AIM system for cancer therapy consultation and is being designed for clinical deployment in the short term, so it was interesting to note the number of human factors issues that they needed to address. The Ohio State Group had three demonstrations available at the workshop One was on a system called RED, which was a blood-typing classification system built using CSRL, the group's own diagnostic system building language. This system incorporates a number of new ideas in the abductive aspects of diagnostic reasoning. The OSU Group also had a demonstration of a system which could compile diagnostic knowledge from a deeper functional representation of device-like systems. The third program from the Ohio State Group was one called MDX/MYCIN, which was an attempt to compare the group's MDX approach to design of diagnostic systems with the MYCIN approach. The methodology adopted was to use the MYCIN knowledge base but to design an MDX-like diagnostic system for the same knowledge base so that both the system building and performance advantages could be comparatively evaluated. The expert consultant program in rheumatology, AI/RHEUM, was presented by the Missouri (Lindberg, Kingsland) and Rutgers (Kulikowski, Weiss) groups. It uses a knowledge representation in the form of criteria tables, developed at Missouri, and is implemented within Rutgers' EXPERT rule-based consultation system. The diagnostic knowledge base comprises 26 rheumatic diseases, with over 800 findings, 400 hypotheses, and 1000 rules. It is being tested and has been transfered to a microprocessor environment (the WICAT). The management component, still under development, serves as a testbed for experimenting with the representation of the effectiveness, contraindications, and temporal relations found in sequences of treatments. Perry Miller from Yale demonstrated a system called ATTENDING, which critiqued user plans for therapy. The idea was that, from a human factors perspective, a system which intervenes with comments only if there is a significant reason to question a physician's plan but which otherwise remains silent has a greater likelihood of physician acceptance. The Rutgers group demonstrated a system for the management of herpes simplex of the eye, which had won an award earlier at a conference on ophthalmology. The system was particularly interesting because it was an IBM PC-based system and as such has tremendous potential for widespread use in the immediate future.

The remainder of our report summarizes a number of the panel discussions. Most of the work in the Workshop took place as part of these panel discussions, where not only the speakers, but also the audience made significant contributions.

### Panel 1: Strategies for Diagnostic Reasoning

*Panelists: Bill Clancey of Stanford, Bill Long of the MIT Laboratory of Computer Science, Jon Sticklen of OSU, and Kaz Kulikowski of Rutgers.*

The first generation of work in diagnostic reasoning, *e.g.,* MYCIN, showed that certain kinds of diagnostic reasoning can be performed impressively, but the rule representation, by providing a gloss of uniformity, hid a number of important distinctions. Bill Clancey of Stanford Heuristic Programming Project, in his work on GUIDON, showed how identifying these distinct kinds of knowledge and strategies can be useful in generating explanations In NEOMYCIN, he has similarly been concerned with useful distinctions that help structure diagnostic knowledge and help search that knowledge and transform the diagnostic hypothesis space. He identified a number of diagnostic operators, such as group, focus, and explain, that are based on diagnostic strategies exhibited by physicians. Thus, Clancey belongs to an emerging group of researchers who wish to identify higher level (or knowledge-level) distinctions that need to be made. Note that in GUIDON, the problem solving itself is done by the inference engine that ran on a rule base; it used the labelings of knowledge that was added to the pieces of knowledge in the base to produce explanations, i.e., the explanations used the labels to couch the explanations. In NEOMYCIN, on the other hand, these diagnostic operators and the structuring of the knowledge base are not added on after the fact for explanatory purposes, but are part of the knowledge representation and of the problem-solving activity.

Jon Sticklen of Ohio State discussed control issues in diagnostic reasoning –in particular, classificatory aspects of diagnosis. The issues arise from the MDX framework, which proposes a topdown, "establish-refine" control strategy for classification, where control passes from consideration of more abstract (or general) hypotheses to more specific successors of the higher level hypotheses that have been established. When multiple independent diseases are present, this approach presents no particular problem. When multiple diseases of the form "A secondary to B" are present, a blackboard can be useful in controlling problem-solving, since information about the status of hypothesis

B can be made available through the blackboard when hypothesis "A secondary to B" is being considered. When symptoms are more quantitative and different parts of the quantity can be accounted for by different disease hypotheses, complex control issues result, requiring a more global perspective, *i.e* , how much of a given symptom ought to be accounted for by a given hypothesis cannot often be done locally by the hypothesis itself. Sticklen discussed the need for, and aspects of, the global critic needed in classificatory diagnosis.

The role of causal relations between diseases and pathophysiological states in diagnostic reasoning was discussed by Bill Long of MIT and Gordon Banks of Pittsburgh. Long also discussed the integration of diagnostic and therapeutic reasoning, a contrast to normal models, where therapy is viewed as a separate process that follows diagnosis. In many domains (including that of heart failure, with which Bill Long has been concerned), diagnostic and therapeutic reasoning are closely related and mutually influence each other. This produces a need for an underlying representation that integrates diagnostic and therapeutic states. The representation problem can be further complicated by complex temporal relations.

Gordon Banks discussed his experience with INTERNIST and CADUCEUS. Both of these systems use the initial complaint to form a context for constraining search. Following this, both of them attempt to match the manifestations with the diseases, so that the best, most likely disease hypothesis is produced. The major difference between CADUCEUS and INTERNIST in the search is that the former uses causal pathways at the patho-physiological level to modify its confidence in a disease hypothesis, while INTERNIST generally uses a fairly straightforward numerical matching to produce these confidences. Thus, the role of more detailed causal knowledge in CADUCEUS is to change the method by which disease hypotheses acquire credibility. This issue of how to take into account deeper causal knowledge has been a recurrent issue in AIM work in recent years A later panel on deep models further considered some aspects of this issue.

Kulikowski noted that one aspect of uncertainty that AI systems have ignored is that of handling unreliable data from the patient. For example, some patients might fear medical treatment to such an extent that they might lie or not disclose all the facts about their problems to the physician This leads to difficult issues, such as determining patient strategies, belief, and assumptions. Although this was considered an interesting problem (and generated a good deal of discussion), most people felt that medical systems should perhaps not be too deeply involved in this issue yet, as there are still many other difficult problems to tackle.

### Panel 2: Cognitive Psychology and AIM

*Panelists. Paul Johnson of Minnesota, Herb Simon of CMU, Alan Lesgold of Pittsburgh, Ben Kuipers of Tufts, and Bill Clancey of Stanford.*

One of the explicit aims of the workshop was to examine the role and possible common ground between cognitive psychological research and AIM work Simon remarked on the points of contact between AI and cognitive psychology and how these fields can progress by borrowing ideas from each other. A number of psychologists have studied the diagnostic process at various times before the advent of AIM. Recently, psychological analysis of the diagnostic process, inspired by the AIM paradigm, has also taken place. Panelists engaged in a wide-ranging discussion on aspects of the relationship between cognitive psychology and AIM

In the area of studying human performance in diagnosis, skill levels have been of interest to cognitive psychology. Johnson and Lesgold described experiments regarding skill levels in diagnosis. The following skill levels have generally been identified:

- novices (typically first and second year medical fellows)
- trainees or intermediate level experts (typically third and fourth year medical fellows)
- acknowledged experts

In general, the performance of intermediate-level experts was judged to be the worst. Lesgold remarked that a basic phenomenon of psychology, the monotonicity of the learning curve, was being violated. The learning curve had points of non-monotonicity, and these points corresponded closely to the stages where the residents were getting more responsibilities and were on their way to becoming experts.

Kuipers and Clancey discussed how the experts and the trainees tended to do recognition-driven reasoning The experts and trainees had a rich vocabulary of prototypes, and their descriptions of problem solutions were in terms of those prototypes To force them into causal reasoning, the stimuli contained some hard cases known to be troublesome to the trainees.

The acknowledged experts reached the correct diagnosis in a fairly short time. The trainees experienced a great deal of difficulty before reaching a correct decision, if at all. For a number of reasons, the diagnostic situation is often full of garden path cues. The intermediate-level physicians are the ones most misled by this situation, and misclassify early in the problem-solving process. Having made a commitment, they experience difficulty in overcoming that choice and moving to considering the correct alternative. The novices may also reach the correct conclusions, but will usually be much less aware of the potential red herrings and can miss them altogether.

The virtuosos appear to be aware that they are in a garden path environment, and they allocate their attentional resources in a highly strategic way. The lesser experts do not know how to monitor their problem-solving

behavior to that extent and hence respond with the automated structures at the wrong points, thereby misinterpreting or overlooking the cues Thus the acquisition of highly data-specific heuristics to interpret cues in garden path cases and a greater control over an otherwise automated process seem to be what distinguish a virtuoso from a lesser expert.

### Panel 3: Deep Models for AIM Systems, including Anatomical and Physiological Models

*Panelists: B Chandrasekaran of OSU, Ben Kuipers of Tufts, Bill Long of* MIT, *Ramesh Patil of* MIT, *Gordon Banks of Pittsurgh, Robert Blum of Stanford, William Swartout of the University of Southern California Information Sciences Institute, and John Kunz of Intellicorp*

There has recently been increasing interest in AIM work about endowing AIM systems with knowledge about "how the body works" or, less ambitiously, about models of disease causation at the pathophysiological level. One class of research in this area can be thought of as "the levels of detail" school, which views the problem as one of changing levels of detail with respect to causal relations; *i e* , in this model, depth relates to fineness of detail in the causal pathways Thus, "A causes B" at the top level can be refined as "A causes S1 causes S2 causes B," where S1 and S2 are more detailed, typically patho-physiological states. Patil and Long discussed a number of issues in this class of models. Patil has worked out a number of operators for refining into greater levels of detail and abstracting into less detailed top levels, *i e.*, shifting levels of detail. (CADUCEUS and CASNET share this broad view of changing levels of detail in causal network as the basic issue in this form of reasoning. There is considerable disagreement, of course, about representation and manipulation of such causal nets at different levels of detail)

Another class of approaches to deep models envisions one or more separate representations, which correspond in some sense to how a device or a subsystem works. The idea is that these representations can be processed to yield various causal relations. Such a processing is, broadly speaking, a kind of qualitative simulation of a subsystem. The work of Ben Kuipers of Tufts and Chandrasekaran and Sembugamoorthy of OSU belong to this class. John Kunz's work in physiological and anatomical models shares some aspects of this, as well as the levels of detail approach. A number of different representations are currently under consideration. Two examples were described at the Workshop. Kuipers described a structural description from which qualitative behaviors can be obtained by a process called "envisioning;" Chandrasekaran and his colleague described an approach that processes a "functional model" of a system to yield a diagnostic knowledge structure.

Jack Smith described the representation and organization issues in the PATHEX (PATHology EXpert) project, where a number of organ parts and organ part abnormalities need to be represented, *e.g* , for histologic analysis of tissue samples Since the type of reasoning involved in this task is distinct from that for the diagnostic task, the MDX metaphor of problem solving by distinct specialist communities is appropriate here Since this organization also reflects the way the medical community is organized, a study of the communication between internists and pathologists was used to obtain a number of constraints on the knowledge and message structures for the PATHEX system.

Gordon Banks discussed how to model and represent spatial relations in anatomical structures, such as neurological structures. The issues here are not cognitive in the sense of representing knowledge of spatial relations, but of organization of the representation so as to enable localization rapidly. He has been experimenting with a representation that uses nested three-dimensional cubes

Robert Blum of Stanford has also been concerned with "causal reasoning" in medicine. His focus is completely different from using causal knowledge to do diagnostic reasoning; he is more concerned with extracting plausible causal hypotheses from medical databases. Thus, the emphasis is more on "discovery" of causal connections than in using known causal connections to generate and evaluate diagnostic hypotheses.

### Panel 4: Explanation in AIM Systems

*Panelists: Bill Clancey of Stanford, Chandrasekaran of OSU, and William Swartout of USC/ISI.*

The major thrust in the discussion was on making a number of distinctions in knowledge and problem-solving structures as well as to arrive at a possible taxonomy of types of explanations that will be needed in AIM systems.

Chandrasekaran distinguished three components that need to be explained (or justified):

- How runtime problem-solving steps were made, *i.e* , how they depend on the data of the case.
- How pieces of knowledge in the knowledge base that were used in solving a given problem can be justified Three types of justification here are:
  (a) by appeal to authority;
  (b) by statistical generalization;
  (c) by derivation from deeper knowledge structures
- How the control strategy can be justified (especially in terms of problem-solving goals).

For generating explanations of type (2c), he proposed that his group's work on functional representations as deep knowledge structures can be useful, since there exist compilation processes which can process these representations to generate diagnostic knowledge in the knowledge base.

Thus, pieces of diagnostic knowledge used during reasoning can be justified by reference to how the subsystem concerned works. For (3) above, he proposed that different kinds of generic problem solving have different generic goals. An explanation of the control strategy that was used by an AIM system can be given in terms of the generic goals of the different types of problem-solving into which the task can be decomposed.

Clancey advanced some key distinctions to keep in mind when analyzing explanation facilities for expert systems:

- The distinction between the knowledge level and the implementation level; this has similarities to the point made by Chandrasekaran about the different types of problem-solving that can be seen at the information processing level as underlying a complex task
- The distinction arising from the different perspectives of psychology and epistemology
- The distinctions between conceptual knowledge (which relates to inference), process and heuristics (which are related to the strategic behavior of the problem solver), and structural knowledge (which relates the two)

He noted that there were justifications underlying every relation, whether it be an inferential association or process ordering. That is, every act of the problem solver, whether it is to make an inference about the problem at hand (make a domain conclusion) or to make strategic decisions about how to make the inference has an underlying justification structure.

Bill Swartout first gave an overview of his group's view of explanation:

Explanation requires knowledge of program development as well as knowledge of the program. In his architecture an automatic programmer is used to create the expert system from abstract domain knowledge. Swartout distinguished various kinds of knowledge that a system with an explanation capability will need to have, *e.g.*, descriptive domain knowledge; problem-solving knowledge, which he calls domain principles; knowledge about tradeoffs, preferences, terminology, definitions and, preconditions for various default assumptions. A very useful part of his work is a compilation of the types of questions a user might ask an expert system. Some examples are:

1. How does the system do ⟨ action⟩ ?
2. How was ⟨ parameter⟩ used?
3. Why is this value being requested? (What are you thinking of?)
4. Why is ⟨ plan-a—conclusion-a⟩ preferred ⟨ over b⟩ ? Why not ⟨ action—conclusion⟩ ? What would it take to get the system to ⟨ action—conclusion⟩ ?
5. What would be the effect of ⟨ x⟩ on the recommendations? On the patient?
6. Capability questions: What can the system do? What does the system know about x?

### Panel 5: Medical Problems Ripe for AI Application

*Panelists: Robert Galen of Cleveland Clinic, Carl Speicher of OSU, Larry Fagan of Stanford, and Ted Shortliffe of Stanford*

An often expressed thought in this panel was that a number of problems in medicine could be successfully handled by the current crop of AIM techniques, but cultural problems in AI and medicine may prevent appropriate rewards for attacking them; those problems may be "trivial" from an AI point of view. They do not present new research problems in AI, or medical departments may not see new medical knowledge coming from these attempts. This is a problem in all new, pioneering, interdisciplinary activities, and at this point it seems to be particularly acute in AIM.

Among the problems that can use current AIM techniques and have done so successfully is the problem of test results, which is repeatedly mentioned (Galen, Speicher). The Coronory Care Unit (CCU) was cited (Fagan, Speicher) as an application area with significant potential utilizing AI techniques; success here would also make AIM highly visible to other divisions of medicine. This is an area where the depth of problem-solving is limited, but where processing large amounts of data in limited time is essential.

Several points are becoming increasingly clear for successful AIM applications and were explicitly stated by Fagan during the panel discussion: the essential role for well-engineered human interaction; the need for a balance between do-able and useful; and the need to choose domains that can be reasonably viewed as "closed word" systems.

Galen pointed out that changing technologies and distributed laboratories are actually making it possible to introduce AIM more successfully in a number of limited, well-circumscribed domains. There was uniform agreement that the role of the physician as the ultimate decision maker must be maintained, and mechanisms for the physician to override the AIM systems' advice need to be provided (Shortliffe).

General consensus indicated a need for a family of higher level languages in which to extract and represent knowledge, *i.e.*, software has become the limiting factor. This was viewed as one of the major bottlenecks preventing wider and wider diffusion of AIM systems in the clinical world. Shortliffe suggested that it was important that a segment of the AIM community concentrate on getting a number of applications out soon by choosing problems in which codified knowledge is available.

### Panel 6: Validation of AIM Systems

*Panelists: Ted Shortliffe of Stanford, Shalom Weiss of Rutgers, Don Lindberg of National Library of Medicine, Perry Miller of Yale, and Larry Fagan of Stanford*

Shortliffe discussed a number of points that were unique to the validation of AIM systems.

- Lack of a "gold standard": external validation of knowledge is often difficult; thus, expert consensus is the only alternative.
- Need to take into account geographical variation in terms of disease prevalence, local customs, etc , so that AIM systems may be "customized;" otherwise validation would be very unreliable.

Weiss suggested a methodology of starting with the validated prototype (validated with respect to correctness of knowledge), and adding further knowledge incrementally, ensuring that previously correct performance is not adversely affected. The idea is that designing and validating the full-blown system will produce intractable validation problems. A similar point was made by Lindberg, who suggested that gradual introduction of AIM systems in less friendly atmospheres is preferable to unbiased, rigorous testing that does not allow for a repair of systems to try to bring it up to specifications. Lindberg also mentioned the problems that arise when AIM systems move from laboratory settings in field sites including complications not encountered at the field setting.

Perry Miller discussed validation of a particular class of AIM systems, critiquing systems, which, instead of providing a decision or complication, accept a physician's plan and critique it as needed. He proposed the notion of "tutorial evaluation," in which the tutorial system is first used under the watchful eye of the local experts. Thus the system is flexed under realistic conditions without endangering any patients.

These presentations were followed by animated discussion, with contributions both from the floor and the panelists. Some of the themes that ran through the discussions were:

- What are the responsibilities of the AIM system designer/implementer after the product is in the marketplace?
- Will physicians be required to use such systems for such systems once such expertise becomes available?
- Will government regulation of the industry become necessary or be forced upon the AIM commercializers?
- Recent criticism of a poor product with alleged AIM content to it[1] raised the question of need for criteria for criticism, as opposed to simple opinion.

[1]R Miller, "INTERNIST: medical diagnosis/information retrieval program," *Annals of Internal Medicine,* Vol 100, No. 4, April, 1984, pp 622-623

### Panel 7: Computing Environments for AIM Research and End User

*Panelists: Thomas Rindfleisch of Stanford, Ed Patterman of Stanford, Chris Lane of Stanford, Shalom Weiss of Rutgers, Tom Bylander of OSU, Ramesh Patil of* MIT, *Casey Quayle of Pittsburgh, and Bob Neches of USC/ISI*

The user community represented above covered a wide spectrum of hardware. In addition to the mainframes (DEC20, Vax), the experience spanned Wicats, XEROX Dandelions, PERQ's, Chipmunks (HP9836), Symbolics, and Apollo workstations. The move from a central computing resource to single-user workstations is pronounced, but the transition is not complete. Central Resources can still be useful for large disk files and large databases. The emerging heterogeneous machines have brought with it significant problems of integration and software transportability. Some comments and specific considerations/concerns follow.

- complexity of transformation from system to system
- bugs in systems and implementing new features for systems
- active networks (and improvements)
- extended data types—good or bad? Do they make for "unmodifiable" code?
- move to implementation of specific languages for specific tasks (build a complex system in a "simple way")
- "home" computing—a thing of the past?
- "Seat sharing" for effective use of time
- workstations lack cohesiveness of mainframe—no fast access to disks
- Symbolics machines: no real sense of history in machine (since none inherent, user left to own resources to implement); large number of good ideas but consistency lacking; poor documentation; powerful, if willing to delve into system

The panel concluded with the conjecture that single-user workstations need a "critical mass," *i.e.,* a cluster of workstations in one location seems to be very effective for a number of reasons.

### Concluding Remarks

The Workshop provided evidence that medicine continues to be one of the most important domains for AI: it continues to provide a rich collection of basic research problems of general interest to AI, and it also remains an application area attracting the attention and interest of some of its best practitioners. It was also clear that here was an example of a *community* in action; the several years of community-building through SUMEX and the Workshop series have borne fruit in this regard.

The 11th Workshop will be held in the summer of 1985 in Washington, D.C., with Don Lindberg and Larry Kingsland of National Library of Medicine and Larry Fagan of Stanford leading the organizational effort. The AIM community is looking forward to reporting on and hearing about another year of understanding and progress.

## Index to Advertisers