COGNITIVE SCIENCE 8, 363–385 (1984)

Causal Reasoning in Medicine: Analysis of a Protocol

BENJAMIN KUIPERS
Massachusetts Institute of Technology
JEROME P. KASSIRER
Tufts University

The ability to identify and represent the knowledge that a human expert has about a particular domain is a key method in the creation of an expert computer system. The first part of this paper demonstrates a methodology for collecting and analyzing observations of experts at work, in order to find the conceptual framework used for the particular domain. The second part develops a representation for qualitative knowledge of the structure and behavior of a mechanism. The qualitative simulation, or envisionment, process is given a qualitative structural description of a mechanism and some initialization information, and produces a detailed description of the mechanism's behavior. The simulation process has been fully implemented, and its results are shown for a particular disease mechanisms in nephrology. This vertical slice of the construction of a cognitive model demonstrates an effective knowledge acquisition method for the purpose of determining the structure of the representation itself, not simply the content of the knowledge to be encoded in that representation. Most importantly, it demonstrates the interaction among constraints derived from the textbook knowledge of the domain, from observations of the human expert, and from the computational requirements of successful performance.

1. INTRODUCTION

How does an expert physician reason about the mechanisms of the body? We are exploring the hypothesis that the physician has a cognitive "causal model" of the patient: a description of the mechanisms of the human body and how they influence each other. This causal model, incorporating the ex-

This research was supported in part by NIH Grant LM 03603 from the National Library of Medicine to the first author. Correspondence and requests for reprints should be sent to Benjamin Kuipers at MIT Laboratory for Computer Science, 545 Technology Square, Cambridge, MA 02139.

pert's knowledge of anatomy and physiology, can be used to simulate the normal working of the body, its pathological behavior in a diseased state, and the idiosyncracies that characterize a particular patient. The causal model supports the expert performance of the physician by simulating the possible courses of the patient's disease and treatment, by serving as a coherency criterion on hypotheses about the patient's state, and by providing a common framework for explanations and discussion among physicians.

If intelligent computer programs are to provide genuinely expert levels of performance in medicine, they must incorporate some sort of causal model, both to support expert problem-solving and to provide an acceptable interface with physicians. Research in artificial intelligence recently has begun to address the problems of causal reasoning in diagnosis, explanation, and troubleshooting, focussing primarily on problems in electronics, in simple physics, and in medicine (de Kleer 1977, 1979; Forbus, 1981, 1982; Kuipers, 1982, in press; Patil, 1981; Pople, 1982). This work has been important in identifying computational constraints on knowledge representations for causal reasoning, but in most cases it has been only loosely constrained by empirical study of the way human experts actually solve problems. Cognitive scientists such as Chi, Feltovitch, and Glaser (1982) and Larkin, McDermott, Simon, and Simon (1980) have studied the ways that experts and novices formulate and solve word problems in physics, but without specifying the knowledge representations and implementing working computer simulations. We believe that it is important to unify these two approaches, to develop techniques for designing knowledge representations constrained by empirical observations. Our goal in this paper is to demonstrate a method we have used successfully to analyze physician behavior in detail, and derive critical properties of the knowledge representation. Taking these empirical constraints along with computational constraints on knowledge representations has allowed us to create a working program that simulates the reasoning process of the physician.

To understand in detail how a human expert reasons about causal relationships is of pragmatic benefit to the designers of expert systems for two reasons. First, if the causal model is to support clear explanations, and be an important part of the interface between expert program and expert human, then its structure should be very similar to that used by the human. Second, we are just learning how to represent causal knowledge so that programs can manipulate it effectively. We are likely to be able to extract valuable clues about the representation and manipulation of causal knowledge, at all levels of detail, by looking carefully at the behavior of expert humans.

The construction of genuinely expert knowledge-based systems requires several different methods of knowledge acquisition. Davis (1982) describes methods for supporting domain experts in providing new knowledge and debugging existing knowledge in a large rule-based system. However, his approach is limited to operating within the knowledge representation chosen

by the system designers. It is also important to develop methods for studying experts to determine the *representation* for the knowledge base, even before attempting to capture large quantities of domain knowledge. The research presented here addresses that problem: of examining the behavior of individual experts to determine the representation of their knowledge and the collection of domain concepts that should be considered fundamental.

2. DESIGN OF THE EXPERIMENT

Most existing research on clinical cognition has used experimental methods designed to gather data that could be combined across many subjects and analyzed using existing statistical techniques (e.g., deDombal, 1972; Rimoldi, 1961). These methods are appropriate to the scientific fields (e.g., biomedicine) where competing hypotheses exist to explain the existing data, and the goal of the scientist is to refute one or the other hypothesis with a reliable, repeatable experiment. In artificial intelligence, however, we typically have no detailed hypotheses adequate to explain even those facts about knowledge representations that we already know. We need a methodology of discovery, to determine constraints from human behavior that can help us develop adequate hypotheses about the structure of knowledge representations. There are two basic questions we want to answer about the behavior of an unknown knowledge representation that will aid in determining its structure:

1. What states of knowledge can be expressed?
2. What inferences can take place?

A methodology of discovery appropriate to the undoubted complexity of human knowledge requires rich data about individuals rather than easily-analyzed data about a population. Individual variation is such a striking feature of human cognition that any attempt to average data across a population is certain to mask the true structure of the knowledge. As Newell and Simon (1972) point out, only the full complexity of verbal behavior, as captured in a verbatim transcript, can do justice to the complexity of the knowledge representation. Therefore, in order to study the representation of causal knowledge in physicians, we decided to analyze verbatim transcripts of a small number of physicians solving problems using their causal knowledge.

The fidelity of the setting is another issue in studying problem-solving behavior. Experimental designs have ranged from recording the responses of subjects to data on a fixed set of cards (Rimoldi, 1961), to collecting verbatim transcripts of the responses a physician gives to a predigested case description (Kassirer & Gorry, 1978), to videotaping physician interactions

with actors trained to simulate patients (Elstein, Shulman, & Sprafka, 1978). On the one hand, it is important to allow the experimental design to reflect a richness of response sufficient to illuminate the complex structure of knowledge representations. On the other hand, the difficulty and cost of collecting and analyzing the data is an important consideration.

After analyzing the alternate methods (Kassirer, Kuipers, & Gorry, 1982), we concluded that an interview based on a detailed printed description of a patient, and resulting in a verbatim transcript, was both more cost-effective and more powerful than the simulated patient encounter to explore the knowledge representation. Note that problem-solving from predigested clinical data is a natural activity for physicians, particularly during residency but also in consultations and other conferences among physicians. While this activity is clearly distinct from the natural patient encounter, we expect that the problem-solving techniques and the nature of the medical knowledge used are very similar.

We designed the interview as a "thinking aloud" experiment, in which the subject is asked to report as much as possible of what he thinks about as he solves a problem. The interviewer intervenes only with non-directive reminders to keep thinking aloud. This type of experiment is particularly sensitive to the natural control structure of the subject's problem-solving method. The experimenter can usually conclude that information reported was actually in the subject's focus of attention at the time, but of course much of what the subject had in mind necessarily goes unreported. Thus, it is not possible to draw direct conclusions about the limits of the subject's knowledge.

We have complemented the "thinking aloud" experiment with a "cross examination" experiment, in which the experimenter asks probing questions about the subject's knowledge of particular topics. The "cross examination" interview is not sensitive to the natural control structure of the problem-solving method, but is much more effective for determining the limits of the knowledge represented, particularly in highly articulate subjects such as physicians. When a subject is being asked to solve only a single problem, the two methods can be combined in an interview that begins with a thinking aloud segment and ends with a cross examination.

In a recent survey (Kassirer et al., 1982), we reviewed the methodologies for investigating clinical cognition and described some of the pitfalls and promise of the analysis of verbatim transcripts of physicians solving realistic medical problems. Although the work of Elstein et al. (1978) is important and path-breaking, we criticized it for its reliance on retrospective reflections of physicians when viewing videotapes of their own behavior (Kassirer et al. 1982). In an extensive review, Nisbett and Wilson (1977) show that a subject has no privileged knowledge of the factors that influence his behavior. Ericsson and Simon (1980) develop a model of the verbaliza-

tion process and use it to clarify and refine Nisbett and Wilson's conclusion. They conclude that a subject's statement of what is currently in his focus of attention is unlikely to be in error, but that his commonsense theory of his own cognitive processes has no particular privileged status. Newell and Simon (1972) provide a clear description of their use of this distinction:

> There is much confusion in psychology about how to deal with verbal data. It is worth emphasizing that we are not treating these protocols as introspections. Actually, there are very few introspective utterances in them. An example does occur at B87:
>
> B86: *What are you thinking now?*
>
> B87: *I was just trying to think over what I was just---*
>
> We treat this utterance for the evidence it gives of the subject's knowledge or operation—in this case, essentially no evidence. The protocol is a record of the subject's ongoing behavior, and an utterance at time t is taken to indicate knowledge or operation at time t. Retrospective accounts leave much more opportunity for the subject to mix current knowledge with past knowledge, making reliable inference from the protocol difficult. Nor, in the thinking-aloud protocol, is the subject asked to theorize about his own behavior—only to report the information and intentions that are within his current sphere of conscious awareness. All theorizing about the causes and consequences of the subject's knowledge state is carried out and validated by the experimenters, not by the subject. [p. 184]

The expert physician, with many years of experience, has so "compiled" his knowledge that a long chain of inference is likely to be reduced to a single association. This feature can make it difficult for an expert to verbalize information that he actually uses in solving a problem. Faced with a difficult problem, the apprentice fails to solve it at all, the journeyman solves it after long effort, and the master sees the answer immediately. Clearly, although the master has the knowledge we want to study, the journeyman will be much easier to study by our methods. The attempts of the apprentice may also be illuminating, particularly in clarifying the relationship between textbook learning and clinical experience. Accordingly, we selected subjects at three widely spaced levels of expertise: medical school faculty members (the masters), second-year residents (the journeymen), and fourth-year medical students (the apprentices). The scope of this paper, however, only permits us to discuss results from a single subject (a journeyman). A later paper will report our comparisons across levels of expertise.

The material for the interview consisted of a slightly atypical case of a kidney disorder called the *nephrotic syndrome,* presented as a case summary on a single sheet of paper. In the nephrotic syndrome, a patient retains salt and water and suffers swelling (*edema*) of the face and legs; the swelling is an important diagnostic finding. Because of a self-induced low-salt diet,

this particular patient experienced no swelling, though all other signs and laboratory results allowed an unambiguous diagnosis to be made. The interview began with a "thinking aloud" section in which the subject made and discussed the diagnosis, and concluded with a "cross examination" section to probe for explanations of particular issues. The atypical case allowed us to compare three different causal models in the same subject: the model of salt and water handling by the healthy kidney, the pathophysiology of nephrotic syndrome, and the idiosyncracies of the particular patient.

3. THE NEPHROTIC SYNDROME

The nephrotic syndrome case was selected to investigate causal reasoning about equilibrium processes, which are central to physiological mechanisms. Two important equilibrium processes are disturbed in the nephrotic syndrome: the transfer of salt and water across capillary walls (the *Starling equilibrium*) and the transfer of salt and water from the plasma into the urine. The Starling equilibrium determines the flow of water between the plasma and the tissues (the spaces between the cells), according to the balance of competing *hydrostatic pressure* and *oncotic pressure* in the plasma and in the tissues. The second important equilibrium, also controlled by the kidney, determines the total amount of salt and water in the body. Under normal circumstances, if the body contains too much salt and water, the kidney excretes more of each into the urine; if there is too little, it cuts back on excretion.

In the nephrotic syndrome, both of these equilibria are shifted to new stable points, changing the quantity of salt and water in the body and causing problems for the patient. The basic cause of nephrotic syndrome is that the diseased kidney excretes protein that it was supposed to retain, and consequently plasma proteins (particularly albumin) are depleted. The amount of protein in the plasma determines its oncotic pressure, and hence is an important factor in the Starling equilibrium. With less protein in the blood, the Starling equilibrium shifts, moving some water from the plasma into the tissues. This movement of extra water into the tissues in itself usually causes no clinical manifestations. However, the shift of water to the tissues leaves the plasma volume low, so the kidney starts to retain water rather than allowing it to be excreted in the urine. The Starling equilibrium, of course, continues to shift much of this additional fluid into the tissues, and substantial edema develops. From the patient's point of view, this accumulation can produce as much as fifty pounds of extra water in the legs and abdomen. To understand the mechanism of edema in nephrotic syndrome requires an understanding of both equilibria and their interaction (Figure 1).

Retention of salt by the kidney is central to the mechanism whereby the kidney retains water. In response to a contraction of plasma volume, the

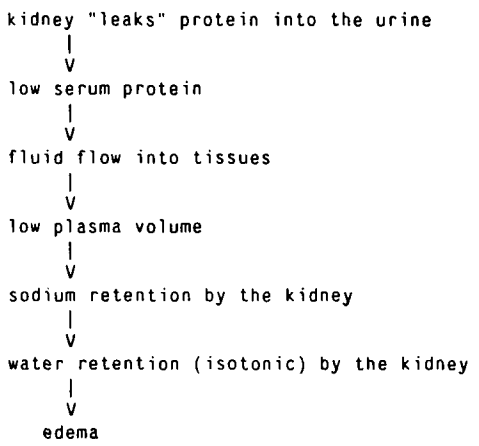

Figure 1. A diagrammatic representation of the causal relations in nephrotic syndrome.

kidney's primary response is to retain salt. Salt retention, in turn, is what causes water retention. The particular patient whose history formed the basis of the experiment had selected a low-salt diet, so the kidney was unable to retain much salt or water, and the edema was consequently much less than a physician would expect, based on the severe decrease in blood proteins. Our subjects all understood this association, but probes of the mechanism by which it works revealed limits to the subjects' knowledge.

Typical of these limits is the treatment of the physical forces, *osmotic pressure* and *oncotic pressure*. A good explanation of nephrotic syndrome must refer to both kinds of pressure, but they can be treated as "black boxes." On the other hand, the mechanisms behind these forces cannot be adequately explained using a linearized "A causes B" explanation. And, in fact, although the more expert physicians used the concepts of osmotic and oncotic pressure correctly, subjects at all levels of expertise gave very weak explanations of how they are caused.

4. ANALYSIS OF THE TRANSCRIPT

The raw data produced by the experiment is a verbatim transcript of the subject's explanation of various aspects of the nephrotic syndrome in general and of this case in particular. As it is transcribed, the transcript is broken into short lines that correspond roughly to meaningful phrases in the explanation (see Table I). How this task is accomplished is not critical, but the format eases the burden of later analysis. Out of the transcript as a whole, selections are made of excerpts in which the subject appears to be concentrating on the explanation and presenting his medical knowledge, rather than expressing an opinion about his own mental processes.

TABLE I
ANALYSIS OF REFERRING PHRASES

A second-year resident explains how loss of protein from the blood causes edema in nephrotic syndrome. The first stage in the analysis consists of identifying and classifying the phrases (italicized below) referring to *substances*. Similar analyses identify references to locations, concentrations, forces, and flow rates (cf. Table II).

L162 A: When there is a very low *albumin* in the serum,
L163 there are two forces which cause edema in my thinking---
L164 the hydrostatic and oncotic forces
L165 and we have actually opposed forces,
L166 forces [. . .break. . .] formation is secondary to
L167 the hydrostatic force of the blood going through the capillaries
L168 and causing the transudation of *fluid*
L169 as well as the osmotic force within the blood vessels,
L170 that is secondary to the *proteins* in the plasma
L171 which tend to draw *fluid*
L172 from the interstitial spaces into the blood vessels
L173 and also there is the forces in the extracellular space.
L174 There are certain *proteins* which tend to pull *water*
L175 out of the blood vessels
L176 and there is a hydrostatic force I believe also in the interstitial spaces
L177 which can counteract the force of the fluid
L178 coming out from within the vessels
L179 and if you have a very low *albumin* in the serum,
L180 there will be a decreased osmotic pressure
L181 and make it easier for the *fluid* to go out into the interstitial spaces.

Substances
protein (L162, 170, 174, 179)
fluid (L168, 171, 174, 181)

The analysis of an excerpt takes place in two stages:

1. Identify the objects and relations in the domain that the subject is referring to, as distinct from the working used to refer to them.
2. Identify the causal relationships that are described in the segment.

Table I presents an excerpt in which the subject, a second-year resident in internal medicine, is explaining (correctly) the mechanism by which the loss of protein from the blood results in edema in nephrotic syndrome. A quick reading of the excerpt shows that the physician is framing his explanation in terms of *substances* in *locations,* causing *forces* which result in *flows.* By attempting to classify each referring phrase in the extract into one of these categories, we can test whether our initial hypothesis about the framework was correct, or whether additional terms need to be added.

By classifying each of the referring phrases in the excerpt as shown in Tables I and II, we can obtain the set of domain objects and relations that constitute the framework of the explanation. The *fluid* referred to is isotonic

TABLE II
DOMAIN OBJECTS IDENTIFIED FROM ANALYSIS OF REFERRING PHRASES

Substances
protein (L162, 170, 174, 179)
fluid (L168, 171, 174, 181)
Locations
blood vessels (L162, 167, 169, 170, 172, 175, 178, 179)
interstitial spaces (L172, 173, 176, 181)
Concentrations
concentration(protein, blood) (L162, 179)
Forces
hydrostatic pressure(fluid, blood ->interstitial spaces) (L164, 167)
hydrostatic pressure(fluid, interstitial spaces ->blood) (L176-178)
serum protein oncotic pressure(fluid, interstitial spaces ->blood) (L164, 169-172, 180)
interstitial protein oncotic pressure(fluid, blood ->interstitial spaces) (L174-175)
Flow Rates
flow(fluid, blood ->interstitial spaces) (L168, 174-175)
flow(fluid, interstitial spaces ->blood) (L171-172)

sodium chloride: water with the same concentration of sodium chloride as the blood. Naturally, there will be objects and relations that are represented in the knowledge structure but were not selected for explicit mention in the explanation. We expect that computational constraints will bring these to light as we later construct a model to account for the explanation.

Once its basic terms have been formalized (Table II), the content of the explanation can be stated explicitly. Table III identifies five different statements of causal relationships in the extract, falling into two categories. Some of the key objects in the domain (concentrations, forces, and flow rates) are continuously-variable quantities, and the subject is asserting facts about those quantities. The first four statements are assertions of structural relationships that hold between certain quantities, without stating anything about the values that they may take on at particular times. The fifth statement refers to the properties that the quantities might take on under particular circumstances, and so describes the behavior of the mechanism.

Our analysis of this excerpt from the transcript, shown in Tables II and III, provides us with the following conclusions, which will serve as empirical constraints on the knowledge representation we devise for the domain knowledge.

1. The explanation refers to a relatively small set of objects and relations describing aspects of the domain.
2. Those objects that are involved in the causal assertions are symbolic descriptions of continuously-variable quantities or the values they take on at a particular time.

3. Descriptions of the structural relationships making up a mechanism are expressed separately, and therefore probably represented separately, from descriptions of the dynamic behavior of the mechanism.
4. The symbolic descriptions of quantities and values are stated in qualitative terms: *directions* of flow, *increased* and *decreased* quantities, *low* albumin, *more* perfusion, and so on. This suggests that the symbolic description of quantity and value is stated primarily in terms of ordinal relations among values.

TABLE III
ANALYSIS OF CONTENT OF STATEMENTS

The first four statements describe structural relationships that hold between continuously-variable quantities. The fifth describes the behavior of the mechanism.

L162 A: When there is a very low albumin in the serum,
L163 there are two forces which cause edema in my thinking---
L164 the hydrostatic and oncotic forces
L165 and we have actually opposed forces,
L166 forces [...break...] formation is secondary to
L167 the *hydrostatic force* of the blood going through the capillaries
L168 and *causing the transudation of fluid.*

L169 as well as the *osmotic force* within the blood vessels
L170 that is *secondary to the proteins* in the plasma
L171 which tend to *draw fluid*
L172 *from the interstitial spaces into the blood vessels.*

L173 And also there is the forces *in the extracellular space:*
L174 there are *certain proteins* which tend to *pull water*
L175 out of the blood vessels;

L176 and there is a *hydrostatic force* I believe also *in the interstitial spaces*
L177 which can counteract the force of the fluid
L178 coming out from within the vessels.

L179 And if you have a *very low albumin in the serum,*
L180 there will be a *decreased osmotic pressure,*
L181 and make it easier for the *fluid to go out into the interstitial spaces.*

Descriptions of Structure

hydrostatic pressure(fluid, blood -> interstitial spaces) (L167)
  => flow(fluid, blood -> interstitial spaces) (L168)

concentration(protein, blood) (L170)
  => serum protein oncotic pressure(fluid, interstitial spaces -> blood) (L169)
  => flow(fluid, interstitial spaces -> blood) (L171-172)

concentration(protein, interstitial spaces) (L174)
  => flow(fluid, blood -> interstitial spaces) (L174-175)

hydrostatic pressure(fluid, interstitial spaces -> blood) (L176)
  => flow(fluid, interstitial spaces -> blood) (L177-178)

Descriptions of Behavior

decreased concentration(protein, blood) (L179)
  => decreased serum protein oncotic pressure(fluid, interstitial spaces -> blood) (L180)
  => increased flow(fluid, blood -> interstitial spaces) (L181)

5. THE DOMAIN MODEL—STRUCTURAL DESCRIPTION

At this point, we have extracted the information that is directly available from the transcript. For the next step in our analysis, we must examine the phenomenon itself—in this case the Starling equilibrium—to find a way to represent the structure of its causal relationships. We need a representation for the Starling equilibrium that can support an expert level of inference about its behavior, and that is consistent with the observations we have made. The purpose of the domain model is to make explicit information that is logically necessary to answer questions correctly about the domain, but may not have been stated in the explanation.

We draw on a physiological description of the Starling equilibrium (Valtin, 1973), and express it in a way that is compatible with our observations of the human expert. Our analysis showed that the explanation was stated in terms of *substances* in *locations,* causing *forces* which result in *flows.* We also observed that the objects involved in causal relationships are symbolic descriptions of continuously-variable quantities. We begin by defining the possible substances and locations, along with quantities representing their amounts and concentrations, and the constraints among those quantities (Table IV). These constraints among quantities are what will make it possible to draw new inferences about the state of the equilibrium from a small set of hypotheses.

TABLE IV

DOMAIN MODEL: Substances, locations, amounts, and concentrations, and some of the constraints holding among the quantities.

Substances: protein, fluid	
Locations: plasma compartment (P), interstitial compartment (I)	
Amounts: amt(protein,P), amt(protein,I), amt(fluid,P), amt(fluid,I)	
Concentrations: c(protein,P), c(protein,I)	
Constraints:	amt(protein,P)=c(protein,P)*amt(fluid,P)
	amt(protein,I)=c(protein,I)*amt(fluid,I)

The Starling equilibrium is an equilibrium between four forces: the hydrostatic pressures and the oncotic pressures in the two compartments (P and I). There are several different ways to combine the effects of these forces to produce a net flow rate, each with different sets of intermediate terms. We select the combination method that provides the best match with the terms used in stating the causal relations (Table III). Thus, we combine two pressures of each type to produce net hydrostatic and net oncotic pressures, each of which causes a flow between the two compartments, which are in turn combined to produce a net rate of flow (Table V).

Other constraints, such as the way the hydrostatic pressure in the blood depends on the amount of fluid in the blood compartment, are very complex and may not even be known to the expert. The physician does,

TABLE V

DOMAIN MODEL: Pressures, rates of flow, and constraints holding between them.

Hydrostatic pressures
- HP(fluid,P->I)
- HP(fluid,I->P)
- net HP(fluid,P->I)

Oncotic pressures
- OncP(fluid,I->P)
- OncP(fluid,P->I)
- net OncP(fluid,I->P)

Flow rates
- flow(fluid,P->I)
- flow(fluid,I->P)
- net flow(fluid,P->I)

Constraints (component addition)
- net HP(fluid,P->I)=HP(fluid,P->I)-HP(fluid,I->P)
- net OncP(fluid,I->P)=OncP(fluid,I->P)-OncP(fluid,P->I)
- net flow(fluid,P->I)=flow(fluid,P->I)-flow(fluid,I->P)

however, know that the functional relationship is strictly monotonically increasing, at least for the situations now being considered. Accordingly, we define a *functional constraint* (M^+) that states that one quantity is an unknown but strictly increasing function of the other. The constraint can be modified (M_z^+) to indicate that the function passes through the origin, as well. In Table III, we see that the functional constraints correspond to statements giving the direction in which one quantity depends on another. The fact that a functional relationship is strictly monotonic provides exactly enough information to support this inference. Table VI gives the functional relationships required to model the Starling equilibrium.

Finally, the rate of flow of fluid from one compartment to another specifies the rate of change of the amount of fluid in each compartment. To capture this domain relationship we must formulate and use a *derivative constraint*. There is no specific phrase in the excerpt that we can identify with

TABLE VI

DOMAIN MODEL: Relationship between hydrostatic pressure and amount of fluid, between oncotic pressure and protein concentration, and between rate of flow and pressure.

Constraints (embedded processes)
- HP(fluid,P->I)=M^+(amt(fluid,P))
- HP(fluid,I->P)=M^+(amt(fluid,I))
- OncP(fluid,I->P)=M_z^+(c(protein,P))
- OncP(fluid,P->I)=M_z^+(c(protein,I))
- flow(fluid,P->I)=M_z^+(net HP(fluid,P->I))
- flow(fluid,I->P)=M_z^+(net OncP(fluid,I->P))

TABLE VII
DOMAIN MODEL: Rate of flow related to change in amount.

Constraints (derivative)
$\frac{d}{dt}$ amt(fluid,I) = net flow(fluid,P->I)
$\frac{d}{dt}$ amt(fluid,P) = - net flow(fluid, P->I)

the use of a derivative constraint, but such a constraint is required for computational adequacy of the model.

This system of equations (Tables IV–VII) constitutes the domain model of the structure of the mechanism of the Starling equilibrium. Figure 2 is a graphical depiction of the structural model, in which the constraint equations are drawn as linking the quantities involved. Sections 7 and 8 will discuss the qualitative simulation process whereby this structural model is used to simulate the mechanism's behavior.

Figure 2 makes it relatively easy to see that the four structural assertions identified in the explanation correspond to the four branches of the domain model.

6. QUALITATIVE SIMULATION IN THE EXPLANATION

We have constructed a precise model of the structure of the mechanism of the Starling equilibrium. The structural assertions identified in the explanation specify the relevant objects, relations, and some of their connections. Examination of the scientific theory of the domain mechanism allowed us to express those connections precisely as computational constraints without sacrificing the qualitative nature of the explanation.

The next step is to augment the representation until it can carry out a qualitative simulation of the *behavior* of the mechanism, given the qualitative description of its structure. Just as we did with the structural description, we hope to use constraints from the observed explanation, from the computational requirements of the representation, and from knowledge of the domain, to specify the representation and its behavior. When this operation is completed, the portions of the explanation describing the behavior of the mechanism should correspond with a well-defined part of the qualitative simulation.

We can now make our analysis of the behavioral parts of the explanation more explicit by overlaying the described behavior onto the structural description. Figure 3 shows how the final statement of the explanation can be overlaid onto Figure 2, showing the causal pathway by which loss of plasma protein causes a shift in the Starling equilibrium, thus translocating fluid from the plasma into the interstitial space.

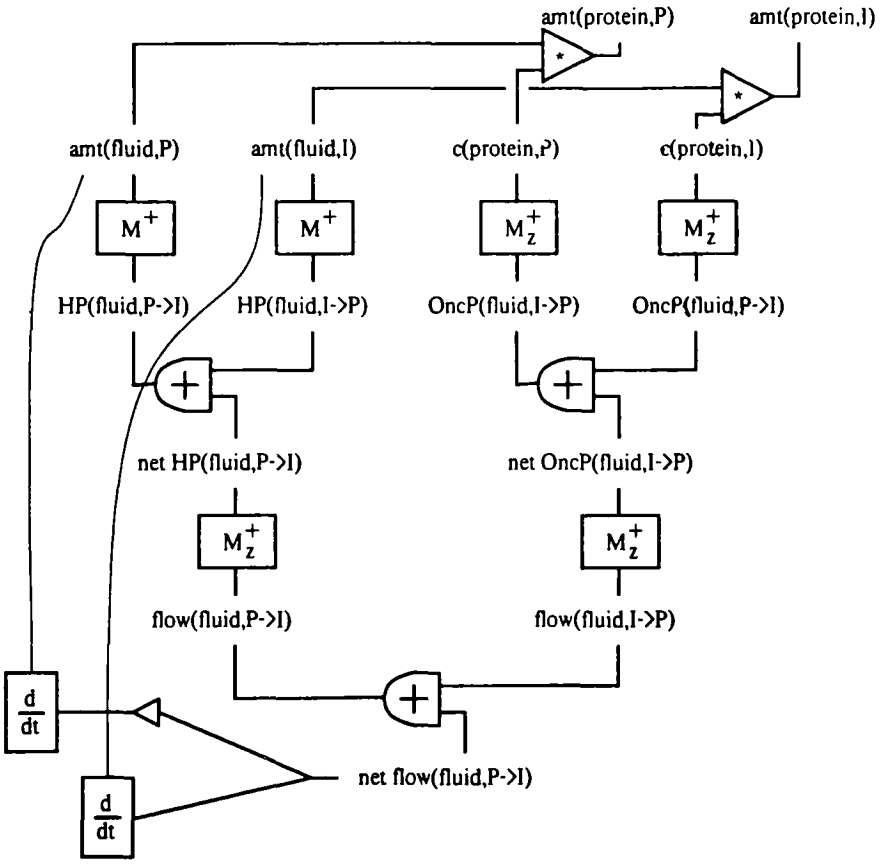

Figure 2. The domain model of the Starling equilibrium showing quantities and constraints. At any point in time, the values of the quantities must obey all of the constraints. The system as a whole changes over time while continuing to satisfy the constraints.

The effect of the change to the Starling equilibrium is primarily to reduce the plasma volume, which in turn causes the kidney to retain salt and water rather than excreting them. The Starling equilibrium continues to shift much of this additional fluid into the tissues, causing the visible swelling of the appendages. In the excerpt below, the subject is explaining this latter process, using only behavioral statements. Table VIII shows the excerpt and its analysis, and Figure 4 shows the qualitative changes overlaid onto the same domain model.

This analysis of the transcript helps specify the behavior we want from the simulation process, and gives us confidence that the terms chosen for the structural description are correct.

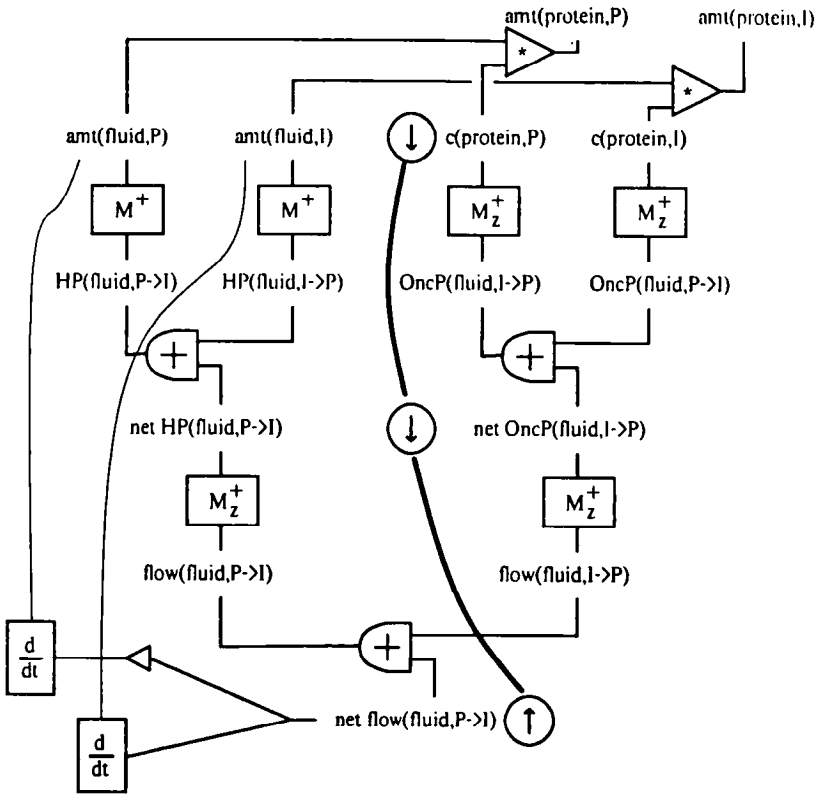

Figure 3. The portion of the explanation referring to the behavior of the mechanism can be analyzed as asserting changes to the quantities involved in the structural description (figure 2).

TABLE VIII
ANALYSIS OF A SEPARATE EXTRACT

The physician is explaining the hypothetical consequences of *increased* salt intake, which would result in increased fluid retention, and hence increased edema. The fragment shown here is only that portion of the explanation which deals with the Starling equilibrium.

L215 The hydrostatic pressure now will increase.
L216 The tissues will be perfused more,
L217 and because of the increased osm... hydrostatic pressure within the vessels,
L218 and the decreased osmotic pressure,
L219 that is the decreased albumin also within the vessels,
L220 we'll get a transudation of fluid, that is, salt water,
L221 from the vessels into the interstitium.

Descriptions of Behavior

increased hydrostatic pressure(fluid, blood->interstitial spaces)(L215)
=>increased flow(fluid, blood->interstitial spaces)(L216)

increased hydrostatic pressure(fluid,blood->interstitial spaces)(L217)
=>increased flow(fluid, blood->interstitial spaces)(L220-221)

decreased amount(protein, blood)(L219)
=>decreased oncotic pressure(fluid, interstitial spaces->blood)(L218)
=>increased flow(fluid, blood->interstitial spaces)(L220-221)

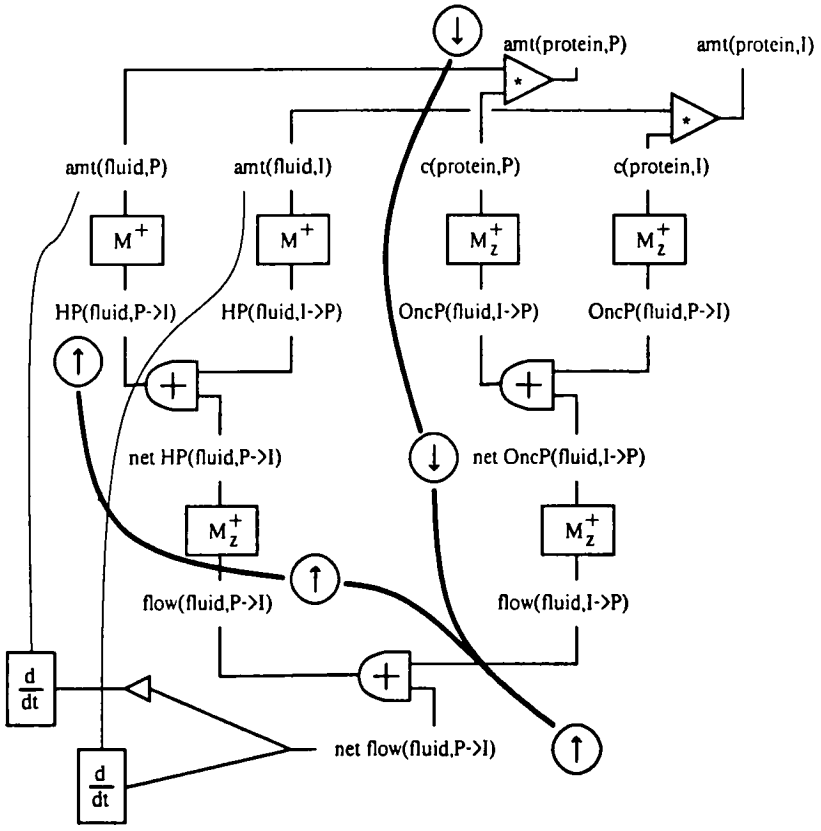

Figure 4. The trace of the behavior described in Table 8 is overlaid onto the domain model.

7. THE DOMAIN MODEL—QUALITATIVE DESCRIPTION OF STATE

The fifth statement in the explanation describes the behavior of the mechanism. By examining the relations described in the transcript, and attempting to maintain logical adequacy, we can propose a representation for the dynamic state of the qualitative simulation, and for the inference rules that drive it.

One conspicuous characteristic of the transcript is the qualitative vocabulary used to describe quantities: *directions* of flow, *increased* and *decreased* quantities, *low* albumin, *more* perfusion, and so on. This suggests that the simulation works primarily with *ordinal relations* among the values of the quantities in the structural domain model: e.g., a quantity is *increased* if its current value is greater than its previous (or its normal) value. The

numerical values of particular quantities (e.g., plasma oncotic pressure) at different times are unspecified and sometimes unknown to the physician. Thus, the knowledge representation must function with *descriptions* of values, not with the numerical values themselves. Since all that is mentioned about those values are their ordinal relationships, we might conclude that the description of a value consists of exactly its ordinal relationships with other values.

Logical adequacy, however, requires us to distinguish between two closely related concepts:

1. The *ordinal relation* between two values: greater-than, equal, less-than
2. The *direction of change* of a single value over time: increasing, steady, decreasing.

A patient's current blood pressure, for example, could be in any one of the nine states combining these two attributes, with different clinical significance in each case. Therefore, the qualitative description of a value must contain both its ordinal relations with other values and its direction of change. The logical necessity of this distinction forces us to include it in any representation for expert causal reasoning, even though the two concepts are difficult to distinguish in the transcript.

This qualitative description in terms of ordinal relations provides a powerful representation for partial knowledge of a collection of related quantities. The representation is rich in states of partial knowledge: where little is known, it is possible to express precisely what is known without having to make additional assumptions or discard useful information (Kuipers, 1979). On the other hand, if there are many "landmark" values of a quantity, then ordinal relationships can specify where the current value lies with respect to the landmarks and provide arbitrarily high resolution.

The constraint types defined before for the structural description interact almost perfectly with these qualitative descriptions of value. Essentially, each constraint acts as a local theorem-prover operating in an unquantified relational calculus, having access to its own axioms and the information known about the associated quantities, and communicating with its neighbors through shared quantities. For example, the constraint $X + Y = Z$ makes inferences of the form:

if $X_1 > 0$ and $Y_1 = 0$ then $Z_1 > 0$,
if $X_1 > X_2$ and $Z_1 = Z_2$ then $Y_1 < Y_2$,
if decreasing (X_1) and steady (Z_1) then increasing (Y_1).

Kuipers (in press) defines this representation in detail, based on a design by Steele (1980) that operates on integer values.

This propagation of information through constraints does not correspond to a sequence of events taking place over time. Rather, we start with a small amount of information about the current state of the mechanism and deduce a much more complete description of the state of the mechanism at the same point in time. The actual simulation process analyzes the configuration of changing values to predict the next state after the passage of time (Kuipers, in press). These two processes correspond to two different senses of "causality." In the first, one assertion is logically subsequent to the other, but temporally simultaneous. In the second case, the second assertion both logically and temporally follows the first.

8. THE DOMAIN MODEL—QUALITATIVE SIMULATION

The propagation of information across the constraints provides an increasingly complete description of the state of the mechanism at a particular point in time, deriving new information about its intermediate variables. Once a sufficiently well-specified description of the current state exists, the simulation process examines the configuration of changing values to determine what can be asserted about the next state whose qualitative description is distinct from the current one. The propagation process then begins again for this new time-point, until yet another state can be determined. DeKleer (1977) introduced the term *envisionment* for this process. The qualitative simulation system has been implemented, and is described in detail in (Kuipers, in press).

The rules for determining the next qualitatively-distinct state are elaborations on the following two types of qualitative changes, which depend on the ordinal relationship between the current value of a quantity and nearby "landmarks" or distinguished values.

> **Move From Distinguished Value:** If the current value of a changing quantity is equal to a distinguished value, then let the next value be an undistinguished value perturbed in the direction of change, closer to the starting point than any other distinguished value.
>
> **Move To Limit:** If the current value of a changing quantity is not equal to a distinguished value, and there is a distinguished value in the direction of change, let the value of that quantity in the next time-point be equal to the next distinguished value.

The subject's goal in his explanation is to show how the Starling equilibrium contributes to edema in the nephrotic syndrome(Table I, L162-163). Our hypothesis is that the explanation is derived from the qualitative simulation of the Starling equilibrium mechanism, based on its structural description. The result we want the explanation to justify is:

amt(protein,P)< normal =>amt(fluid,I)>normal.

Table IX shows the result of envisioning the Starling equilibrium. We assume that the reasoning system has, from its previous knowledge of nephrology, a description of the normal state of the Starling mechanism in equilibrium. (State (N) in Table IX represents that normal state; the term "norm" in each line refers to the normal value of *that* quantity, to simplify the notation. State (1) is created by asserting the initial conditions defining the nephrotic syndrome:

amt(protein,P) < normal and held constant,
amt(protein,I) = normal and held constant,
amt(fluid,P) = normal,
amt(fluid,I) = normal.

Thereafter, the propagation process completes the description of state (1). The simulation process asserts new ordinal relations in state (2) for each changing quantity in state (1), and propagation adds the directions of change to complete the description of state (2). The simulation process must diagnose which of several qualitative changes take place after state (2). It concludes that the first qualitative change is the one that makes *net flow(fluid,P -> I) = 0,* but leaves all other changing quantities different from their previous normal values. The propagation process fills in the directions of change (all *steady*) to show that state (3) is an equilibrium.

TABLE IX
ENVISIONMENT OF THE STARLING MECHANISM
Use of the envisionment to show that
amt(protein,P) < normal = > amt(fluid,I) > normal.

Quantity	*(N)*	*(1)*	*(2)*	*(3)*
amt(protein,P)	=norm (std)	<norm const	<norm const	<norm const
amt(protein,I)	=norm (std)	=norm const	=norm const	=norm const
amt(fluid,P)	=norm (std)	=norm (dec)	<norm (dec)	<norm (std)
amt(fluid,I)	=norm (std)	=norm (inc)	>norm (inc)	>norm (std)
c(protein,P)	=norm (std)	<norm (inc)	<norm (inc)	<norm (std)
c(protein,I)	=norm (std)	=norm (dec)	<norm (dec)	<norm (std)
HP(fluid,I->P)	=norm (std)	=norm (inc)	>norm (inc)	>norm (std)
HP(fluid,P->I)	=norm (std)	=norm (dec)	<norm (dec)	<norm (std)
net HP(fluid,P->I)	=norm (std)	=norm (dec)	<norm (dec)	<norm (std)
OncP(fluid,I->P)	=norm (std)	<norm (inc)	<norm (inc)	<norm (std)
OncP(fluid,P->I)	=norm (std)	=norm (dec)	<norm (dec)	<norm (std)
net OncP(fluid,I->P)	=norm (std)	<norm (inc)	<norm (inc)	<norm (std)
flow(fluid,I->P)	=norm (std)	<norm (inc)	<norm (inc)	=f* <norm (std)
flow(fluid,P->I)	=norm (std)	=norm (dec)	<norm (dec)	=f* <norm (std)
net flow(fluid,P->I)	=0 (std)	>0 (dec)	>0 (dec)	=0 (std)

- "norm" refers to the normal value of *that* quantity.
- initial inequalities propagate to provide ordinal relations.
- derivative constraints provide directions of change, which then propagate.
- (1) = > (2) as many values move from distinguished values.
- (2) = > (3) as flow(fluid,I->P) = flow(fluid,P->I) precedes any other qualitative change.
- f* is the new distinguished value for flow(fluid,I->P) and flow(fluid, P->I).

Examining the qualitative values in Table IX, we see that the original goal was achieved, of explaining the link:

amt(protein,P)<normal = >amt(fluid,I)<normal,

since the antecedent of this causal link was asserted as an initial condition, and the consequent holds true in the final equilibrium state. An additional important feature of this simulation process is the fact that many other facts are derived and stored about the states of the other variables in the mechanism. These other variables are critical as the interfaces to other physiological mechanisms. In this case, the value of *amt(fluid,P)* in state (3) acts as the interface with the total body fluid equilibrium.

The requirement of computational adequacy tells us that the reasoning process must carry out this simulation in order for the reasoner to predict the behavior of the mechanism. It must produce a wealth of detail in order to interface correctly with the many other mechanisms in human physiology. On the other hand, a careful examination of the behavioral assertion in Table III and its overlay representation in Figure 3 shows that the content of the subject's explanation is derived solely from the propagation of information through the network to complete state (1). A possible explanation for this is that the qualitative simulation is both complicated to express, and capable of running to conclusion on its own, so the most effective explanation omits the simulation trace.

9. CONCLUSION

We have followed the derivation of a working computer simulation of an aspect of causal reasoning from end to end. The first part of the paper demonstrates a methodology for collecting and analyzing observations of experts at work, in order to find the conceptual framework used for the particular domain. The second part of the paper develops a representation for qualitative knowledge of the structure and behavior of a mechanism. The qualitative simulation, or envisionment, process is given a qualitative structural description of a mechanism along with initialization information, and produces a detailed description of the mechanism's behavior.

By following the construction of a knowledge representation from the identification of the problem to the running computer simulation, this paper provides a vertical slice of the construction of a cognitive model. It demonstrates an effective knowledge acquisition method for the purpose of determining the structure of the representation itself, not simply the content of the knowledge to be encoded in that representation. Most importantly, it demonstrates the interaction among constraints derived from textbook knowledge of the domain, from observations of the human expert, and from the computational requirements of successful performance.

The knowledge representation for causal reasoning is presented in greater detail in (Kuipers, in press), along with several examples in nonmedical domains that reveal more of its interesting properties. Since the objects of the representation are descriptions of continuously-variable quantities, and their relationships are expressed as arithmetic, derivative, or functional relations, the resulting models look very similar to physiological models in the style of Guyton, Jones, & Coleman, (1973) or systems dynamics models in the style of Forrester (1969). One might ask how the models differ, and whether we could avoid the analysis of transcripts and create the models directly from the scientific literature in physiology.

The detailed analysis of physician behavior suggested the *level of description* for the causal models: the set of qualitative relationships and their inference rules that can express incomplete knowledge while remaining able to draw useful conclusions about behavior. Once we have determined an appropriate qualitative representation, it is possible that existing techniques for acquiring knowledge in expert systems (Davis, 1982) will be adequate to specify the content of the models using input from human experts and from the scientific literature.

The representation presented here differs from the Guyton and Forrester models in its ability to express a larger, more flexible set of states of partial knowledge. In particular, the functional constraints M^+ and M^- express functional relationships known to be monotonic in a specific direction but otherwise unknown. Furthermore, the simulation based on this structural description is not limited to precisely specified numerical values, but can operate on symbolic descriptions that constrain the possible numerical values a quantity could take on at a particular time.

Another important difference is how the use of the model influences its size and its scope. When the laboratory scientist formulates a Guyton-style model to account for a phenomenon, he attempts to include every possible factor and relationship that influences the mechanism, so the models tend to become very large. An expert physician reasoning about a case uses only those factors he considers particularly relevant, and thus is able to restrict his attention to a much smaller model. To make up for the lack of detail, the expert must then have many different small models, each with its own assumptions and thus expressing different "points of view." The causal model representation is intended to express this highly modularized knowledge structure, so its models will typically be relatively small. Indeed, it appears that there is a match between the limited working memory and processing capacity of the human and the inability of the causal model representation to handle very large models.

This representation for the structure and behavior of a mechanism is intended to express descriptions that are strictly weaker than the corresponding differential equation, in the sense that several different differential

equations would be consistent with a single causal model. Figure 5 shows the ideal relationship between the two descriptive systems.

Having found the causal model representation by detailed study of the behavior of human experts, we can return to the suggestion that we concentrate on the medical facts of the domain as captured in the medical literature. The fact that the causal model is strictly weaker than the corresponding differential equation model may allow us to construct and validate truly large medical knowledge bases. It suggests the possibility that causal models might be constructed by systematically transforming precise models from the scientific literature into the weaker causal model representation. The resulting causal models would then constitute the knowledge base. Weakening the descriptive language allows the system to reason effectively with the type of mixed qualitative and quantitative information that is typically available to physicians. Much more work is needed before this method can be tested and realized, but it is an attractive alternative to the current slow and unverifiable methods for constructing large knowledge bases.

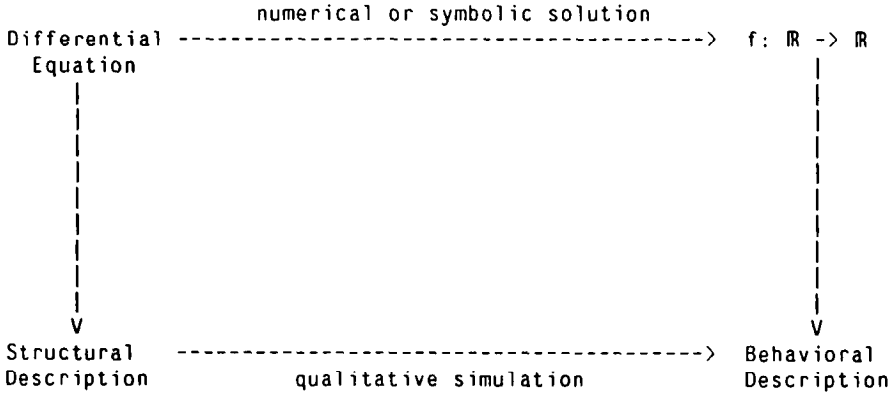

Figure 5. The qualitative structural description is capable of capturing more partial states of knowledge than differential equations, and produces a partial description of the mechanism's behavior. Because the qualitative simulation occasionally uses heuristics, the two paths through the above diagram do not necessarily yield the same result.

REFERENCES

Chi, M.T.H., Feltovich, P. J., & Glaser, R. (1982). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Davis, R., Lenat, D. B. (1982). *Knowledge-based systems in artificial intelligence.* New York: McGraw-Hill.

deDombal, F. T., Leaper, D. J., Staniland, J. R., McCann, A. P., & Horrocks, J. C. (1972). Computer-aided diagnosis of abdominal pain. *British Medical Journal, 2,* 9-13.

de Kleer, J. (1977, August). Multiple representations of knowledge in a mechanics problem-solver. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence.* Cambridge, MA.

de Kleer, J. (1979, August). The origin and resolution of ambiguities in causal arguments. *Proceedings of the Sixth International Joint Conference on Artificial Intelligence.* Tokyo, Japan.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215-251.

Forbus, K. D. (1981, August). Qualitative reasoning about physical processes. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence,* Vancouver, BC.

Forbus, K. D. (1982). Qualitative Process Theory. Cambridge, MA: MIT Artificial Intelligence Laboratory Memo 664.

Forrester, J. (1969). *Urban dynamics.* Cambridge, MA: MIT Press.

Guyton, A. C., Jones, C. E., & Coleman, T. G. (1973). *Circulatory physiology: Cardiac output and its regulation* (2nd ed.). Philadelphia: W. B. Saunders.

Kassirer, J. P., & Gorry, G. A. (1978). Clinical problem solving: a behavioral analysis. *Annals of Internal Medicine, 89,* 245-255.

Kassirer, J. P., Kuipers, B. J., & Gorry, G. A. (1982). Toward a theory of clinical expertise. *The American Journal of Medicine, 73,* 251-259.

Kuipers, B. J. (1979). On representing commonsense knowledge. In N. V. Findler (Ed.), *Associative networks: The representation and use of knowledge by computers.* New York: Academic Press.

Kuipers, B. J. (1982, August). Getting the envisionment right. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-82).* Pittsburgh, PA.

Kuipers, B. J. (1984). Commonsense reasoning about causality: Deriving behavior from structure. *Artificial Intelligence, 24*(1).

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208,* 1335-1342.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

Patil, R. S. (1981). Causal representation of patient illness for electrolyte and acid-base diagnosis. (Tech. Rep. No. 267). Cambridge, MA: MIT Laboratory for Computer Science.

Pople, Jr., H. E. (1982). Heuristic methods for imposing structure on ill structured problems: The structuring of medical diagnostics. In P. Szolovits (Ed.), *Artificial Intelligence in Medicine.* Washington, DC: AAAS/Westview Press.

Rimoldi, H. J. A. (1961). The test of diagnostic skills. *Journal of Medical Education, 36:* 73.

Steele, Jr., G. L. (1980). The definition and implementation of a computer programming language based on constraints. (Tech. Rep. No. 595). Cambridge, MA: MIT Artificial Intelligence Laboratory.

Valtin, H. (1973). *Renal function: Mechanisms preserving fluid and solute balance in health.* Boston: Little, Brown.