# HOW TO DISCOVER A KNOWLEDGE REPRESENTATION FOR CAUSAL REASONING BY STUDYING AN EXPERT PHYSICIAN.[1]

Benjamin Kuipers
Jerome P. Kassirer

Department of Mathematics, Tufts University, Medford, MA 02155.
and
Department of Medicine, New England Medical Center,
171 Harrison Avenue, Boston, MA 02111.

## Abstract

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 expert computer system. The first part of this paper demonstrates a methodology for collecting and analysing 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. This "vertical" slice of the construction of a cognitive model demonstrator, 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 way the body works? We are exploring the hypothesis that the physician has a cognitive "causal model" of the patient that 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. 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 create a working program that simulates the reasoning processes of the physician.

Research in artificial intelligence has recently begun to address the problems of causal reasoning in diagnosis, explanation, and trouble-shooting, focussing primarily on problems in electronics, in simple physics, and in medicine [2,3,6,7.9,10,1'1,15]. 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, et al [1] and Larkin, et al [11] 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 methods are designed for determining the knowledge representation of the knowledge base, even before attempting to capture large quantities of domain knowledge.

1. This research was supported in part by NIH Grant LM 03603 from the National Library of Medicine to the first author. We also gratefully acknowledge the intellectual environment of the Clinical Decision Making Group at the MIT Laboratory for Computer Science, supported in part by NIH Grant LM 03374 from the National Library of Medicine.

## 2. Design of the Experiment

Our methodology must determine constraints from human behavioi that can help us develop adequate hypotheses about the structure of knowledge representations [8]. 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?
(?) What inferences can take place?

A methodology of discovery appropriate to the undoubted complexity of human knowledge requires richly-structured data about individuals rather than easily analyzed data about a population. As Newell and Simon [12] 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. Our study included 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 tc discuss results from a single subject (a journeyman).

The interview is designed 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. This type of experiment is particularly sensitive to the natural control structure of the subject's problem solving method, but cannot support direct conclusions about the limits of the subject's knowledge. The

"thinking aloud" experiment is complemented by 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.

In a recent survey, Kassirer, Kuipers, and Gorry [0] review the methodologies for investigating clinical cognition and describe 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 (4| is important and path-breaking, Kassirer, et al [8] criticize it for its reliance on retrospective reflections of physicians when viewing videotapes of their own behavior. In an extensive review, Nisbett and Wilson (131 show that a subject has no privileged knowledge of the factors that influence his behavior Ericsson and Simon [5] develop a model of the verbalization 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. It is the subject's commonsense theory of his own cognitive processes that has no particular privileged status.

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. 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 atypical case allowed us to compare throe 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 standing *equilibrium)* and the transfer of salt and water from the plasma into the urine. The Standing 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. 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, keeping the body in balance but causing problems for the patient. The basic cause of nephrotic syndrome is that the kidney excretes protein that it was supposed to retain, and consequently plasma proteins 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. 1 he Starting equilibrium, of course, continues to shift much of this additional fluid into the tissues, and substantial *edema* (swelling, particularly in the legs) 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.

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 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.

## 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, it is broken into short lines that correspond roughly to meaningful phrases in the explanation (see Table 1). Excerpts are selected 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. 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 wording used to refer to them,

(2) identify the causal relationships that are described in the segment.

Table 1 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 1 and 2, we can obtain the set of domain objects and relations that constitute the framework of the explanation. Naturally, there will be objects and relations that are represented in the knowledge structure but were not selected for explicit mention in the explanation. Computational constraints will bring these to light as we later construct a model to account for the explanation.

L162 A: When there is a very low albumin in the scrum,
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
LI 71 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)

Table 1. 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 in the excerpt referring to substances. Similar analyses identify references to locations, concentrations, forces, and flow rates (cf. Table 2).

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)

Table 2. The complete set of objects and relations identified in the excerpt in Table 1.

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)

Table 3. The first four statements describe structural relationships that hold between continuously-variable quantities. The fifth describes the behavior of the mechanism.

Once its basic terms have been formalized (Table 2), the content of the explanation can be stated explicitly. Table 3 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 values 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 2 and 3, provides us with the following conclusions, which will serve as empirical constraints on the knowledge representation we devise for the domain.

(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.

## 5. The Domain Model - Structural Description

For the next step in our analysis, we must examine the Starling equilibrium itself to find a way to represent the structure of its causal relationships 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. Based on our observations of the transcript, we begin by defining the possible substances and locations, along with quantities representing their amounts and concentrations, and the constraints among those quantities (Table 4).

**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)

**Table 4**. Domain model: substances, locations, amounts, and concentrations, and some of the constraints holding among the quantities.

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 the explanation. 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 5).

**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)

**Table 5**. Domain model: pressures, rates of flow, and constraints holding between them.

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, 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. Table 6 gives the functional relationships required to model the Starling equilibrium.

**Constraints (functional)**

$HP(fluid,P,I) = M^+(amt(fluid,P))$

$HP(fluid,I,P) = M^+(amt(fluid,I))$

$OncP(fluid,I,P) = M_Z^+(c(pr,P))$

$OncP(fluid,P,I) = M_Z^+(c(pr,I))$

$flow(fluid,P,I) = M_Z^+(net\ HP(fluid,P,I))$

$flow(fluid,I,P) = M_Z^+(net\ OncP(fluid,I,P))$

**Table 6**. Domain model: relationship between hydrostatic pressure and amount of fluid, between oncotic pressure and protein concentration, and between rate of flow and pressure.

Finally, the rate of flow of fluid from one compartment to another specifies the change in 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 the use of a derivative constraint, but such a constraint is required for computational adequacy of the model.

**Constraints (derivative)**

$$\frac{d}{dt}\,\text{amt(fluid,I)} = \text{net flow(fluid,P,I)}$$

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

Table 7. Domain model: rate of flow related to change in amount.

This system of equations (Tables 4 - 7) constitutes the domain model of the structure of the mechanism of the Starling equilibrium. Figure 1 is a graphical depiction of the structural model, in which the constraint equations are drawn as linking the quantities involved.

## 6. Qualitative Simulation in the Explanation

The structural assertions we have identified in the explanation specify the relevant objects, relations, and their connections. The next step is to augment the representation until it can carry out a qualitative simulation of the *behavior* of the mechanism, given this qualitative description of its structure. Just as we did with the structural description, we will 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.

We can illustrate our analysis of the behavioral parts of the explanation by overlaying the described behavior onto the structural description. Figure 2 illustrates the final statement of the explanation, 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.

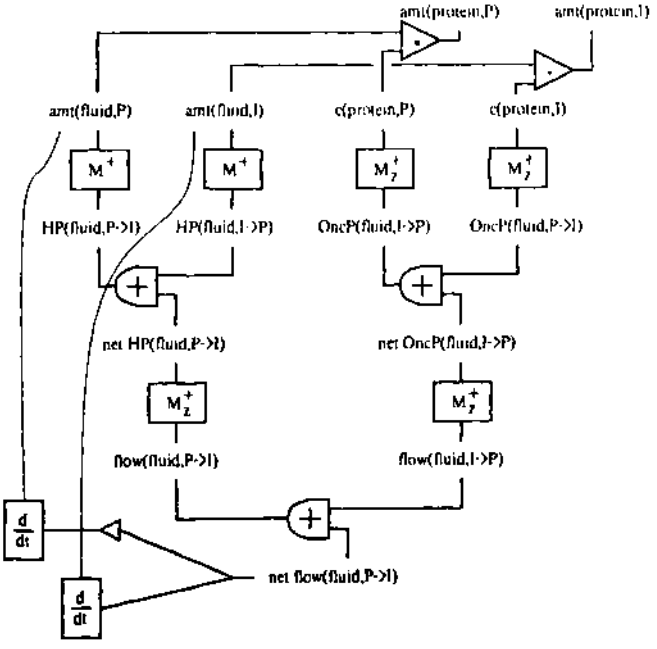


Figure 1. A diagrammatic representation of the domain model of the Starling equilibrium (Tables 4 - 7) showing quantities related by arithmetic, functional, and derivative constraints. (The sign of net flow(fluid,P,I) is inverted before reaching one of the derivative constraints.) At any point in time, 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.

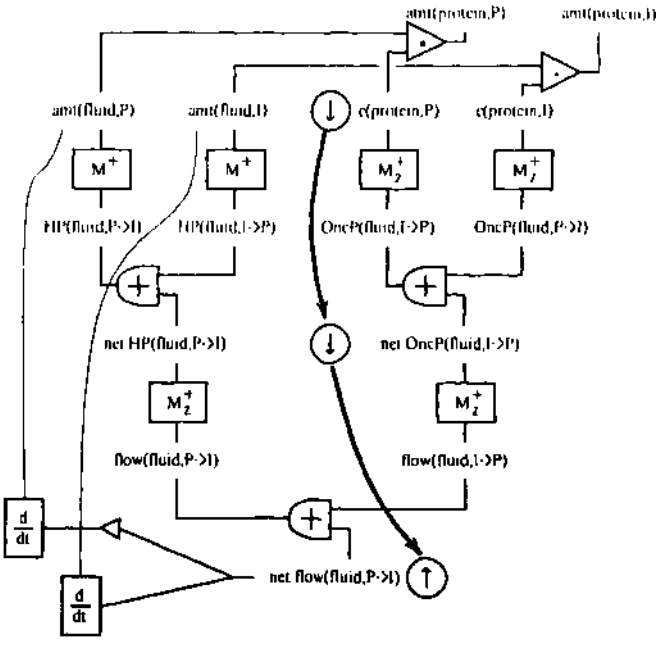


Figure 2. The portion of the explanation referring to the behavior of the mechanism can be analyzed as asserting changes to the quantities involved in the structural description (figure 1).

## 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 patients 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.

The constraint types defined above 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 [10] defines this representation in detail, based on a design by Steele [16] 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 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.

## 8. The Domain Model - Qualitative Simulation

The propagation of information across the constraints provides a more complete description of the state of the mechanism at a particular point in time, deriving new information about the states of 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 [2] introduced the term *envisionment* for this cyclic process.

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 1, L162 163). Our hypothesis is that the explanation is derived from the qualitative simulation of the Starting 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 8 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 8 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 makes net flow(fluid.P.h = 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.

| quantity | (Normal) | (State 1) | (State 2) | (State 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) |

Move From Distinguished Values — Collision

**Table 8.** Use of the envisionment to show that amt(protein,P) < normal => amt(fluid,I) > normal.

- "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.
- State 2 is caused by State 1 as quantities move from distinguished values.
- The first qualitative change from State 2 occurs as two changing quantities, flow(fluid,I,P) and flow(fluid,P,I), become equal (i.e. collide) at the new distinguished value f*.

Examining the qualitative values in Table 8. 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. Patil [14] addresses the problem of maintaining the correspondence between this detailed causal model description and the clinical level of description. The many facts derived about the states of other variables in the mechanism serve as the interface 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 description in Table 3 and its illustration in Figure 2 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 developed a representation for qualitative knowledge of the structure and behavior of a mechanism. The qualitative simulation, or envisionment, process is given a structural description of a mechanism and some initialization information, and produces a detailed description of the mechanism's behavior. The knowledge representation for causal reasoning is presented in greater detail in [10], along with several examples in nonmedical domains that reveal more of its interesting properties.

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 the textbook knowledge of the domain, observations of the human expert, and the computational requirements of successful performance.

This representation for the structure and behavior of a mechanism is similar to differential equations, but expresses descriptions that are strictly weaker, in the sense that several different differential equations would be consistent with a single causal model.

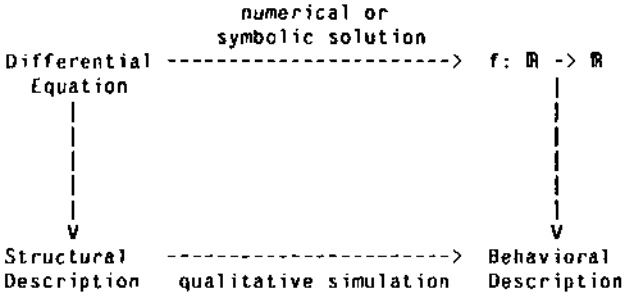


Figure 3. The qualitative structural description is capable of capturing more partial states of knowledge than differential equations, and produces a partial description of the mechanism's behavior. Because the qualitative simulation occasionally uses heuristics, the two paths through the above diagram do not necessarily yield the same result.

The fact that the causal model is strictly weaker than the corresponding differential equation model may have important implications for the construction and validation of a truly large medical knowledge base. 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 weaker descriptive language allows the system to reason effectively with the type of mixed qualitative and quantitative information that is typically available to physicians. The systematic relationship with formal scientific models of physiology suggests a possible alternative to the current slow and unverifiable methods for constructing large knowledge bases.

## 10. References

[1] M. T. H. Chi, P. J. Feltovich, and R. Glaser. 1982. Categorization and representation of physics problems by experts and novices. Cognitive Science 5: 121-152.

[2] J. de Kleer. 1977. Multiple representations of knowledge in a mechanics problem-solver. Proceedings of the Fifth International Joint Conference on Artificial Intelligence. Cambridge, Mass.

[3] J. de Kleer. 1979. The origin and resolution of ambiguities in causal arguments. Proceedings of the Sixth International Joint Conference on Artificial Intelligence. Tokyo, Japan.

[4] A. S. Elstein, L. S. Shulman, and S. A. Sprafka. 1978. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press.

[5] K. A. Ericsson and H. A. Simon. 1980. Verbal reports as data. Psychological Review, 87, 215-251.

[6] K. D. Forbus. 1981. Qualitative reasoning about physical processes. Proceedings of the Seventh International Joint Conference on Artificial Intelligence, Vancouver, BC.

[7] K. D. Forbus. 1982. Qualitative Process Theory. Cambridge, MA: MIT Artificial Intelligence Laboratory Memo 664.

[8] J. P. Kassirer, B. J. Kuipers and G. A. Gorry, Toward a theory of clinical expertise. The American Journal of Medicine 73: 251-259,1982.

[9] B. J. Kuipers. 1982. Getting the envisionment right. In Proceedings of the National Conference on Artificial Intelligence (AAAI 82). Pittsburgh, Pennsylvania, August 1982.

[10] B. J. Kuipers. 1982. Commonsense reasoning about causality: Deriving behavior from structure. Tufts University Working Papers in Cognitive Science, No. 18.

[11] J. Larkin, J. McDermott, D. P. Simon, H. A. Simon. 1980. Expert and novice performance in solving physics problems. Science 208: 1335-1342.

[12] A. Newell and H. A. Simon. 1972. Human Problem Solving. Englewood Cliffs, NJ: Prentice-Hall.

[13] R. E. Nisbett and T. D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. Psychological Review 84: 231-259.

[14] R. S. Patil. 1979. Design of a program for expert diagnosis of acid base and electrolyte disturbances. Cambridge, MA: MIT Laboratory for Computer Science TM-132.

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

[16] G. L. Steele, Jr. 1980. The definition and implementation of a computer programming language based on constraints. MIT Artificial Intelligence Laboratory TR-595.